MODELLING OF HUMAN AGEING, COMPOUND EMOTIONS, AND INTENSITY FOR AUTOMATIC FACIAL EXPRESSION RECOGNITION

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCES AND ENGINEERING

2019

By
Nora Algaraawi
School of Computer Science

Table of Contents

Final word count : **39**, **186**

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

AAM	Active Appearance Model
AFER	Automatic Facial Expression Recognition
ASM	Active Shape Model
BE	Basic Expression
BOW	Bag of Words
BRIEF	Binary Robust Independent Elementary Features
CLM	Constrained Local Models
DBN	Dynamic Bayesian network
DL	Deep Learning
FACS	Facial Action Coding System
FEL	Facial Expression Localization
FER	Facial Expression Recognition
FFPD	Facial Features Points Detector
GPU	Graphics Processing Unit
HMMs	Hidden Markov Models
HOG	Histogram of Gradient
LBP	Local Binary Pattern
LPQ	Local Phase Quantization
MC-SVM	Multi Class Support Vector Machine

NN	Neural Networks
PCA	Principal Components Analysis
QLZM	Quantized Local Zernike Moment
RF	Random Forest
RFRV-CLM	Random Forest Regression Voting Constrain Local Models
RFRV	Random Forest Regression Voting
RLMS	Regularized Landmark Mean Shift
RMSE	Root Mean Square Error
RR	Recognition Rate
SSM	Statistical Shape Model
SVM	Support Vector Machine
SVR	Support Vector Regressor

Abstract

After decades of research, automatic facial expression recognition (AFER) has been shown to work well when restricted to subjects with a limited range of ages, expressions, and intensities of expression. Recognition of the expressions of subjects across a large range of ages (including older people), expressions (compound emotions), and intensities (ranging from a neutral expression to the apex of the target expression) is harder and, to date, has not been studied in any particular depth.

This thesis focuses on studying the influence of these problems on the accuracy of AFER. The main concern is to investigate the possibilities that can be used for modelling facial expression recognition against the impact of the problems under study in order to ensure the solution is more generalized and effective. Since the face image is a collection of texture and shape parameters, the study starts by using texture measurement methods to understand the influence of those problems on face texture features and hence on texture-based AFER. Our first contribution shows that by using binary robust independent elementary features (BRIEF) (Calonder et al., 2012), we can develop a new face descriptor model that is able to describe face images and can generalize to new data sets. The BRIEF descriptor is able to generate the discriminative features globally from the image with an explicit shape. However, when BRIEF is used to generate feature from an image with no explicit shape such as the face image, BRIEF is unable to generate discriminative feature. We thus propose to use BRIEF locally to ensure that each pixel in the image is evaluated locally to capture the local shape surrounding around it. Empirical and comprehensive evaluation using three facial expression datasets demonstrates that this model gives satisfactory performance compared to other local face descriptors techniques evaluated on the same datasets. The study also shows that the patterns of the problems under consideration have a significant effect on the face texture features and on the accuracy of texture-based AFER.

The study is then extended by using shape measurement methods to investigate the influence of those problems on the face shape features and hence on shape-based AFER. Our second contribution shows that by using random forest regression voting in a constrained local model (RFRV-CLM) framework (Cootes et al., 2012; Lindner et al., 2015), we can develop a fully automated facial expression localization (FEL) system that is able to detect the facial key points in a multiple-stage (coarse-to-fine) scheme and can generalize

accurately to new data sets with a wide range of variations of facial appearances. Empirical and comprehensive evaluation using five different facial expression datasets demonstrates that this model gives excellent agreement with ground truth data and outperforms the results of alternative methods evaluated on the same datasets. The study also shows that the patterns of the problems under study have a significant effect on the performance of FEL, and that the FEL based on RFRV-CLM achieved good performance against that effect. It also demonstrates that appearance-based AFER (combining shape with texture) gives better results than texture-based AFER.

Our final contribution builds on the second and it is the development of an age-based AFER system that explicitly estimates age group and expression in a single framework. In this system, we show that by using the age information, in particular apparent age since some people might look younger or older than their real age, as prior knowledge to the expression recognition through using a weighted combination rule of a set of age group classifier and age-specific expression classifiers, we can significantly eliminate the influence of age features on the expression classification accuracy. Tested on three age-expression datasets, we show that the results of our novel system were encouraging in comparison to the state-of-art systems which ignore age and alternative models recently applied to the problem.

In summary, the results of the BRIEF-based face descriptor, RFRV-CLM-based FEL, and age-based AFER are encouraging and could be basic building blocks for many face applications in computer vision such face detection, face recognition, ..etc.

Declaration

I hereby declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

By
Nora Algaraawi
School of Computer Science
2019

Copyright

Dedication

This work is dedicated to:
MY LADY (**FATIMAH AL ZAHRA** (A.S))

MY PARENTS ESPECIALLY TO THE SOUL OF MY MOTHER (RAJA AL-KAFAJY)
TO ALL THOSE WHO SUPPORT ME IN MY JOURNEY

Acknowledgements

All praise belongs to Allah. I would like to offer my amazing thanks to a number of people during the 4 years of the PhD.

First and foremost, I would like to offer my gratitude to my supervisor, **Dr Tim Morris**, the person who has given me the opportunity to study the PhD at the University of Manchester, U.K. He has taught me how good research is done. I am thankful to him for his guidance, support, motivation, and encouragement all the way through my PhD. I have been extremely lucky to work with him, and I hope to cooperate with him in the future.

I am thankful to my second supervisor, **Dr. Aphrodite Galata** who inspired me to understand what a PhD involves and she provided me with a level of critical thinking. I am looking forward to working with her in the future.

I must sincerely thank my second reader, **Prof. Tim Cootes** for his endless support and guidance, for patiently sharing his knowledge, attention, suggestions and for all the opportunities he presented to me during my PhD studies which enabled me to achieve a lot, and I am glad to have another opportunity to cooperate with him in the future.

Thanks to **Dr. Evangelos Sariyanidi** for providing Quantized Local Zernike Moment(QLZM) software that I used in my thesis.

Special thanks to all members of the ADVANCE-INTERFACE-GROUP in the Computer Science Department of the Manchester University who turned out to be great buddies especially Qinhao.

Special thanks also to all members of the CV-READING-GROUP in Imaging Sciences Centre of the University of Manchester especially Raja and Luca for their moral support and technical discussion.

I acknowledge the financial support from the Higher Committee for Education Development in Iraq (HCED) for funding and Scientific Research, the Iraqi Ministry of Higher Education and Scientific Research, and the University of Kufa in Iraq. I owe to them for the scholarship and thank you for the trust which is given to me to complete my PhD.

My deepest thanks to my father, my husband and my wonderful children - Mohamedrdha, Fatimah, and Ali who their support and understanding have been a strong driving force throughout my work. Thank you all.

Chapter 1

Introduction

1.1 Introduction

The work on facial expressions was started by Charles Darwin (1872) in his book "The expression of the emotions in man and animals", in which he claimed that there are six Universal Facial expressions: happy, sad, angry, fear, disgust, and surprise plus the neutral expression. In (Ekman, 2002), the authors described the face's muscles movements that are used to generate those expressions in a framework called the Facial Action Coding System (FACS). FACS is one of the most common systems in behavioural sciences. It was developed to describe the deformation of the musculature responsible for the appearance of facial expressions. Ekman (2002) found that there was an encoding of 44 distinct action units (AUs) that are anatomically related to contraction of specific facial muscles, each of which is intrinsically related to a small set of localised muscular activations. Figure 1.1 illustrates the universal facial expressions and their AUs (highlighted in grey).

Fig. 1.1 Examples of universal facial expressions and their associated AUs. Those images are from Cohn–Kanade (CK+) dataset (Lucey et al., 2010).

1.2 Motivations

The universal facial expressions are considered an essential part of a non-verbal message and a universal language of human life; they help us to recognise and understand emotions, mood, and the mental states of others. Furthermore, Mehrabian (2008) and Kaulard et al. (2012) have proved that non-verbal signals such as facial expression convey two-thirds of human communication, as compared to the verbal message which conveys one-third. For that reason, facial expression analysis has been an interesting topic of ongoing research, primarily associated with the fields of neurology, and psychology. However, since the computer has begun to be used to automate significant parts of our lives, to the extent that our lives sometimes depend on it, automatic facial expression recognition (AFER) by a machine such as a computer has become a crucial area of research in the fields of computer science and machine learning.

AFER, which gives a computer the ability to interpret the user's emotions, has a significant role in our daily activities, and it can enable various technologies and applications in several domains including human behavioural science, human-computer interaction, security, interactive games, computer-based learning, entertainment, telecommunication, and psychiatry (Picard et al., 1995). For instance, in pharmacology, the effect of new antidepressant drugs can be evaluated more accurately based on the information conveyed via the patients' facial expressions than by asking the patients to fill out a questionnaire, as is currently done by Cohn et al. (2009). Moreover, Whitehill et al. (2008) described how automatic facial expression recognition can be effectively used to estimate the difficulty level of a delivered lecture. Consequently, facial expression recognition might enable a new generation of teaching systems to adapt to the expressions of their students' emotion in the way that good teachers and instructors will themselves attempt to do. Furthermore, in (Vural et al., 2007) expression recognition was used to assess the lassitude of drivers and pilots, which might be a useful means by which to decrease the numbers of driving and flying incidents. The state and intention of humans and their response would be assessed by robot assistants that work through automatic recognition of facial expressions.

1.3 Research Problems

Although the computer itself has no ability to have emotion it can nevertheless recognize a human's expression of emotions in a similar way to the ability of one person to recognize the feelings of another (Picard et al., 1995), the nature or characteristics of facial expressions and the methods used for modelling facial expression are not easy, and indeed are the most important challenges that need to be considered when recognizing expressions. In other words, one can ask how we convert facial images with their various proprieties and facial deformations associated with expression into a numerical form that gathers faces

with similar expressions together into groups. Historically, the majority of studies have focussed on the analysis of properties and variations of the six basic expressions including happiness, anger, disgust, fear, sadness, and surprise. However, there are several problems (face deformations) that might impact on the characteristics of those expressions. This thesis focus on the following problems:

- **Human ageing:** Research in psychology has shown that increasing age represents one of the most significant problems in recognising facial expression (Ebner and Johnson, 2009, 2010; Guo et al., 2013; Hess et al., 2012; Houstis and Kiliaridis, 2009). Ageing is a complicated process due to the biological changes in facial musculature and skin elasticity with significant variations among individuals. These changes might distort the appearance of an expression, instead making it appear similar to the appearance of age. Age-related structural changes can overlap with expression-induced changes. For example, the fold between the cheek and upper lip can appear in the happy expression of young people and the neutral expression of old people, or sagging eyelids in old people can appear in the sad expression of young people (see Figure 1.2). Since both age and expression appearances start from the same face image, the performance of AFER might be affected by the overlapping between age and expression feature and vice versa.

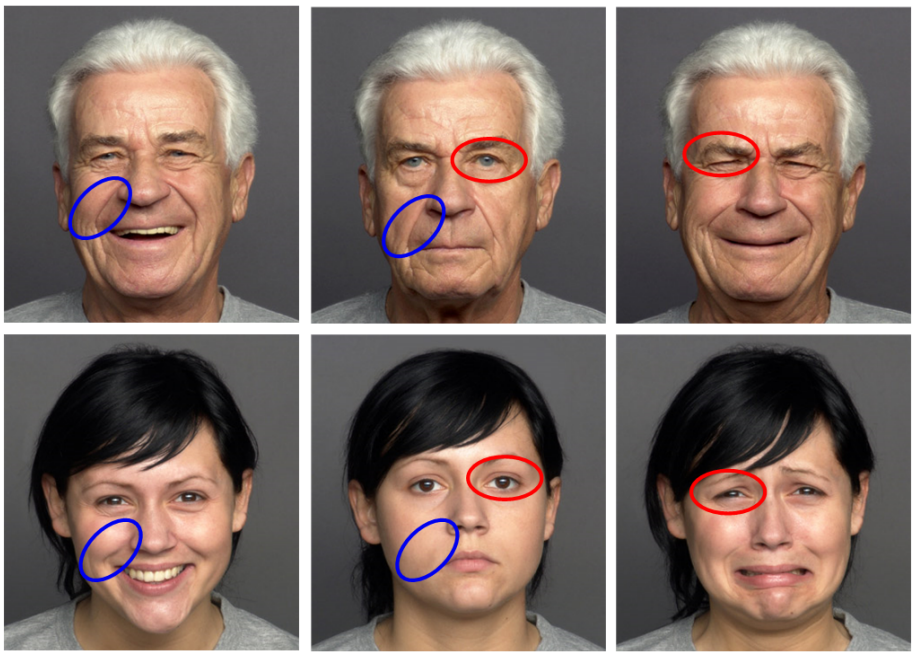


Fig. 1.2 Illustration of the similarities between ageing and expression appearances of happy, neutral, and sad expressions of old people (top row), and young people (bottom row). Those images are from FACES dataset (Ebner et al., 2010).

- **Compound emotions:** Compound emotions also represent one of the difficulties in AFER modelling (Du et al., 2014). Compound emotion is a very complex

process in which several emotions are exhibited at once. Those expressions are constructed by combining two or more basic expressions in order to create the new non-basic expression as shown in Figure 1.3. The idea here is each one of the basic emotions consists of some of the AUs, and if two of those emotions are combined a new emotion will be created, and some of those AUs will be kept and some may disappear in the resulting emotion. As a result, the compound emotion is partially similar to the basic emotions which are used to generate it. These huge similarities between the basic and non-basic expressions might lead to huge confusions and hence to a poor performance in the AFER.

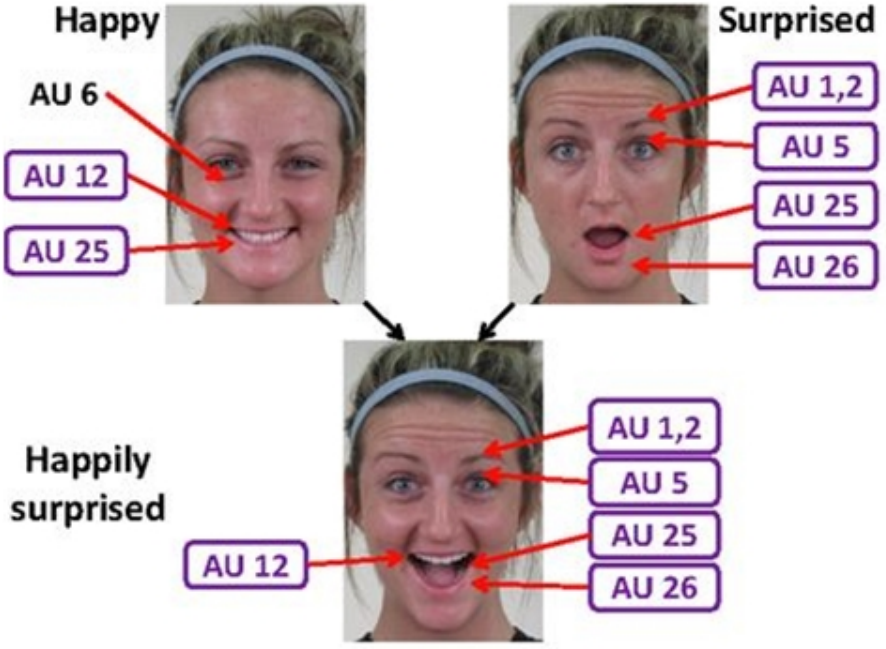


Fig. 1.3 Illustration of the similarities between the AUs of basic emotions (top images) and compound emotion (bottom image). The AUs of the basic expressions that take part to produce the compound emotion are labelled with a box in the compound emotions dataset. This figure is adapted from Du et al. (2014)

- **Expression's intensity:** Expression's intensity is one of the most critical facial expression characteristics (Adolphs and Tranel, 2004; Calder et al., 1996; Hess et al., 1997; Hoffmann et al., 2010; Motley and Camden, 1988; Rotshtein et al., 2010). Despite the process of detecting and recognising the presence of expression being possible, estimating the intensity of that expression is a more difficult process. The intensity of the expression increases gradually (not suddenly) until it reaches the apex as shown in Figure 1.4. Therefore, an AFER system needs to be sensitive to the subtle deformations of different intensities of a given expression.

Finding an optimal and generalised method or feature by which to distinguish among emotions under a wide range of face deformations is one of the most difficult and chal-

Fig. 1.4 Illustration of the expression's intensity. The image on the left has $\approx 0\%$ intensity, while the image on the right has $\approx 100\%$ intensity of happy expression. Those images are from the CK+ dataset (Lucey et al., 2010).

lenging tasks in computer vision due to the unknown variations in both the face and the expressions, as described above. As a result, AFER is an area of research which is in particular need of sensitive exploration in order to give machines the ability to interpret human emotions in a more general and effective way. The goal of ongoing research is to increase the robustness of the systems against different factors. Ideally, the aim is to develop a facial expression recognition system which simulates the remarkable capabilities of human visual perception.

1.4 Research Challenges

The following highlights some of the areas that currently draw research interest in AFER:

- **Face representation:** Representing a face image is one of the major problems in facial expression recognition. The choice of methods used for face representation significantly influence the level of expression recognition achieved. Shape-learning-based, and texture-learning-based techniques of modelling a face have achieved the highest success levels in AFER. Despite there being an enormous amount of methods used for face representation represented in the literature, their ability to recognise and their sensitivity to the massive variation of face deformation patterns is not clear (Benitez-Quiroz et al., 2017).

- **Facial feature points localization:** localizing the facial feature points is the process of using a facial landmark detector as an automated tool to predict the position of every predefined feature in every image directly. This process is an essential component for several expression analysis tasks, especially those with large datasets and high accuracy requirements, in which annotating each image manually is a time-consuming, laborious, and expensive process. The process of facial feature point detection is difficult due to the rigid (translation, rotation, and scaling) and non-rigid (age and expression) face deformations and as it requires an optimization in high

dimensions, where appearance can vary widely between the faces of individuals due to lighting conditions, image noise, resolution and intrinsic sources of variability such as identity, age and expression. Although the results of some state-of-the-art methods are comparable to a human's in some datasets, their generalizability to the real world conditions with a large range of shape variation remains unknown Wang et al. (2018).

- **Overlapping age with expression:** Research in psychology has shown that human ageing has a significant effect on facial expression recognition. This effect is simply due to the overlap between the age and expression features. AFER thus needs to be able to eliminate the impact of age appearance, such as wrinkles and furrows, on expressions meaning. The process of eliminating these effects is not easy since both age and expression features started from the same face image. Although a few studies have jointly and successfully modelled the appearances of both age and expressions, joint modelling of these different face deformations remains an open research area and more investigation is required to reduce the negative effects of that overlap.

- **Real and apparent age:** Recent research in computer-based age estimation systems proved that the apparent age of someone's face can be different from the chronological age as it has been shown that ageing, health, and lifestyle affect facial appearance (Antipov et al., 2016; Huo et al., 2016; Liu et al., 2015; Rothe et al., 2015, 2018; Uricár et al., 2016; Zhu et al., 2015). Results from those studies also showed that apparent age can be used as an alternative ground truth data to the real age for age estimation task. These explorations regarding the apparent age and its differences from the real age might increase the difficulties of modelling the age and expression. Therefore, the impact of apparent age on the performance of AFER needs to be investigated and modelled, as it might help to increase the accuracy of facial expression recognition performance.

1.5 Research Aims

This thesis's aim is to improve the state-of-the-art in the area of AFER. The general aim is to create a system to automatically analyse and recognise the facial expression with a high level of sensitivity to internal variations from the faces themselves including ageing, compound emotions, and expression's intensity, allowing precise measurements of AFER. The main focus is on the analysis of the human ageing's impact on the performance of AFER, aiming to capture and model the similarities and overlaps between age and expressions' appearances by jointly modelling age group estimation and age-specific facial expression recognition in a single framework using the same face features, since both

tasks start from the same face image. In so doing, we aim to achieve a more efficient and effective AFER system compared to the previous methods. The thesis also focuses on the analysis of the compound emotions effect on the performance of AFER, aiming to capture and model the partial similarities between the basic and non-basic expressions. The thesis also focuses on the analysis of the expression's intensity's influence on the performance of AFER, aiming to capture and model the expression category and its intensity.

1.6 Thesis Contributions

The following gives an overview of the main contributions of this thesis.

- **A novel face descriptor:** In this thesis, the BRIEF feature developed by Calonder et al. (2012) is introduced to develop a new face descriptor for describing the face image in general and facial expression in particular. We demonstrated that this model gives satisfactory performance on the task of facial expression recognition compared to the performance of alternative descriptors applied to the problem using the same dataset. We also show that the BRIEF-based face descriptor outperforms the alternative face descriptors in terms of generalisation capability to the new data set of facial expression with different characteristics. Using BRIEF-based face descriptor and some alternative face descriptor techniques we also figure out that ageing, compound emotions, and intensity have a significant effect on the face texture features and on texture-based AFER (see Chapter 4).

- **A fully automated facial expression localization (FEL) system:** In this thesis, the RFRV-CLM developed by Cootes et al. (2012); Lindner et al. (2015) is introduced to develop a fully automated FEL system that can detect the facial feature points automatically by placing a set of points along the face and its component's contours. Qualitative and quantitative results of five different facial expression datasets demonstrate that this model gives excellent agreement with ground truth data and outperforms the results of other detectors evaluated on the same datasets. During the model building, the 2052 images of the FACES dataset were annotated manually by the author of this thesis with 76 landmark points. These points are used to build an automatic facial features points detector to produce the 76-points model used on the test data. The produced points detector was used to automatically annotate the Lifespan and NEMO datasets with 76 points. We can provide our labelled landmark points for the three datasets to other researchers for further studies. Using the prosed FEL system, the results show that the problems under consideration have a significant influence on the face shape features and on the shape-based AFER (see Chapter 5).

- **Fully automated age-specific AFER using real and apparent age:** In this thesis, we show that by using the age information, in particular apparent age since some people might look younger or older than their real age, as prior knowledge to the expression recognition through using a weighted combination rule of a set of age group classifier and age-specific expression classifiers, we can significantly eliminate the influence of age-related features on the expression classification accuracy. We also show that our simple and novel system performs better than an equivalent system which ignores age, or is otherwise comparable, to the results found for alternative manifold-based and deep feature-based models recently applied to the problem. (see Chapter 6).

1.7 Thesis Structure

The rest of the thesis is organised into the following structure.

- **Chapter 2** reviews the related work on AFER system. First, it covers a wide variety of frameworks and approaches from the vast literature and reviews which methods are suitable for which circumstances and uses that to motivate our choice of approach. Second, it provides both a better understanding of the phenomena and challenges in AFER and a review of works which are relevant to the cases under consideration.

- **Chapter 3** sets out how we evaluate our contributions. First, we cover the different datasets that we used to evaluate our methods. Second, we discuss how a face image is represented for facial expression recognition. Thirdly, we cover the image features that we use to extract information from training images which are used for comparison and validation purposes. Finally, we cover the classifiers that we use to recognize the expression.

- **Chapter 4** gives the first of our contributions. It describes the development of a BRIEF-based face descriptor as well as the reasoning for choosing BRIEF. Experiments performed to optimize BRIEF's free parameters for face representation are evaluated in the task of AFER. A comparative evaluation of BRIEF and some other state-of-the-art face descriptor methods in the task of AFER is presented. The sensitivity of BRIEF and other face descriptor methods to the variations of age and expression, basic and non-basic expressions are presented as well.

- **Chapter 5** gives the second of our contributions. It describes the development of the fully automatic FEL system using RFRV-CLM as a FFPD as well as the reasoning for choosing RFRV-CLM. Experiments performed to optimize RFRV-CLM's parameters for the problem of facial expression point localization, along with a comparative evaluation of RFRV-CLM and some other state-of-the-art FFPD

methods are presented. The sensitivity of RFRV-CLM and other FFPD methods to the variations of age and expression with the basic and non-basic expressions , and expression's intensity are presented as well.

- **Chapter 6** gives the third of our contributions. It propose an algorithm of combining sets of age group and age-dependent expression classifiers in a single framework. It also studies the influence of apparent age on the performance of AFER.
- **Chapter 7** concludes the thesis with a summary of its contributions, limitations and suggests directions for future work.

Chapter 2

Background and Literature Review

This chapter reviews previous research related to the work presented in this thesis. In particular, it presents research in the area of AFER, aiming to show the limitations of existing methods, and gives an opportunity to select efficient methods that can be used for the present study. This overview starts by dividing the existing studies on AFER into conventional-based AFER, where handcrafted features are used for face and expression representation, and deep-learning based AFER, where deep features are used for the face and expression representation. While discussing the existing systems, particular attention is paid to the current studies on face representation methods as it is a key step in any AFER system because the classification accuracy is limited by the quality and relevance of the features used in the representation. Particular emphasis is paid also to the current studies in terms of analysing the impact of some challenges including human ageing, compound emotions, and expression's intensity on the performance of AFER as these are the main concerns of this thesis, and its proved that they have a negative impact on the success of facial expression recognition. Figure 2.1 describes the conceptual framework that is proposed in this thesis to analyse and compare the current state of the art in AFER studies.

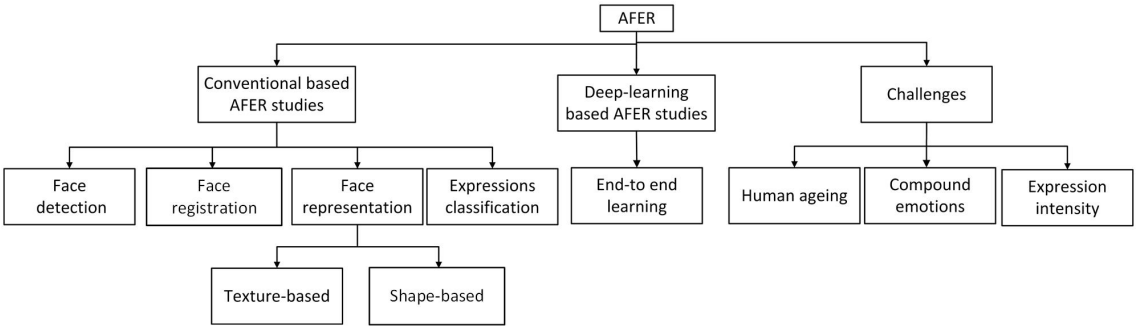


Fig. 2.1 The proposed conceptual framework used in this thesis for the analysis and comparison of existing systems.

2.1 Introduction

Suwa (1978) made the first attempt to build a computer-based (machine-based) system to automatically recognise and interpret the expression from a face image. That system was called an AFER system. Since then, huge progress has been made to increase the robustness of such systems. A complete description of previous work on AFER can be found in abundant literature, for example published by Fasel and Luettin (2003); Kumari et al. (2015); Martinez and Valstar (2016); Pantic and Rothkrantz (2000); Sariyanidi et al. (2015); Sumathi et al. (2012), whereas this chapter is limited to presenting and analysing the recent advances in AFER systems in order to put the work of this thesis into context.

2.2 Conventional-Based AFER Approaches

In conventional-based AFER approaches, handcrafted features are used for face and expression representation. Since the generic conventional-based AFER framework consists of four main modules, namely (1) detecting the face, (2) registering the face, (3) representing the face, and (4) classifying the expressions, this section reviews recent methods used for each module by decomposing existing systems into their fundamental components.

2.2.1 Studies on Face Detection

Face detection is the process of finding whether there is a face in a given image and its location. It is a major and critical requirement for facial expression recognition systems. This is because it segments out the appropriate face region for subsequent modules (registration, feature extraction, and classification). As such, the performance of the face detection module will significantly affect the overall result of the facial expression recognition process. Usually, most facial expression recognition systems assume that the input face has been correctly detected and cropped during the face detection step. A great deal of studies have been implemented to detect the face in an image of an arbitrary scene. The reader is referred to (Zafeiriou et al., 2015) for a more complete survey of recent advances in face detection. The two most popular approaches in the literature for face detection are the face detector proposed by Rowley et al. (1996, 1998) and the seminal work by Viola and Jones (2004).

In Rowley et al. (1996, 1998), the face is detected in two main steps, the first is the preprocessing step, in which the image is subsampled into windows and the intensity of each window is equalized to compensate for differences in camera input gains, and improve the contrast in some cases. In the second step, the preprocessed windows are passed through a neural network to decide whether the window contains a face. This network consists of three types of hidden units: 4 of 10×10 pixels subregion, 16 of 5×5

subregion, and 6 of 20×5 subregion as shown in Figure 2.2. Each of these types was chosen to allow the hidden units to detect local features that might be important for face detection.

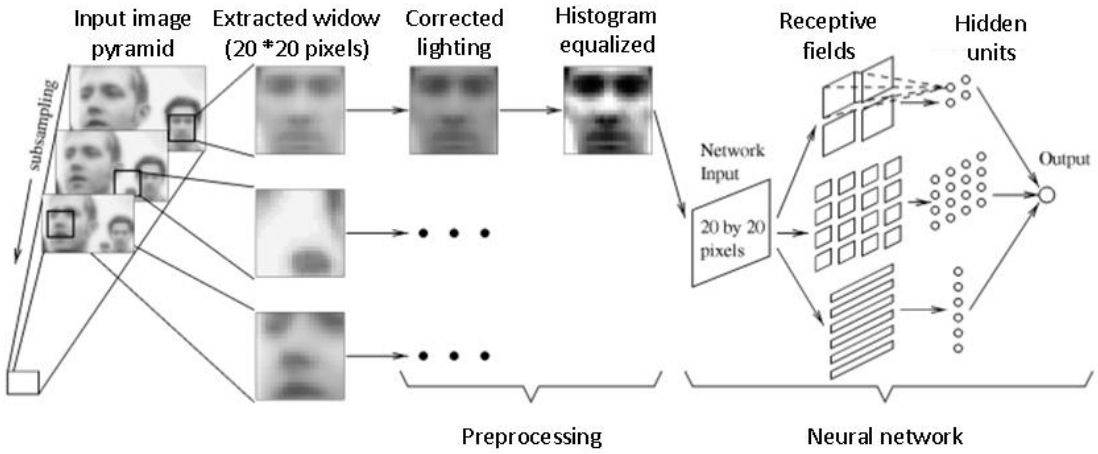


Fig. 2.2 Illustration of the face detection algorithm proposed by Rowley et al. (1998)

The face detector in the seminal work by Viola and Jones (2004) consists of four main steps. The first step is the use of Haar features. The second step is the use of the integral image algorithm to speed the process of Haar features calculation. The third step is the use of AdaBoost learning to select which Haar feature should be used and put them in a linear combination. Boosting is a method to find a highly accurate hypothesis by combining many "weak" hypotheses, each with moderate accuracy. The final step is the training of a cascade classifier which consists of a series of weak classifiers. Figure 2.3 shows the main components of the Viola and Jones face detector. An implementation of the Viola and Jones (2004) face detector can be found in the Open-CV library. Recently, this algorithm has been used with many facial expression recognition systems such as the systems developed by Dang et al. (2014); Lou et al. (2018); Majumder et al. (2013); Owusu et al. (2014); Sariyanidi et al. (2013); Shan et al. (2009).

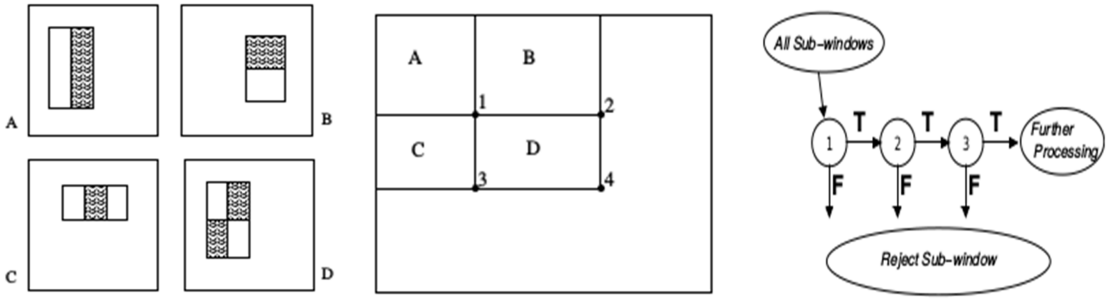


Fig. 2.3 Illustration of face detector method developed by Viola and Jones (2004). The Haar features (left), the integral image (middle), and the Boosting with cascade (right).

2.2.2 Studies on Face Registration

Face registration is the second critical requirement for a conventional-based AFER system, i.e. aligning the face to a common position and orientation. This is especially important in an uncontrolled context where subjects are free to move. Accurate registration is needed where expressions are natural (spontaneous expression) or with the problem of large displacement of the head, where just finding the face location in the image is not sufficient to extract accurate features. Figure 2.4 illustrates the negative effect on the feature matching due to misregistration. The goal of face registration is to find geometric transformations or deformations which reduce the inconsistency between two or more faces. In other words, the target of face registration is to minimize the differences in face shapes between individuals due rotation and scale. The process of face registration can be decomposed into two main steps: intra-subject registration and inter-subject registration. Intra subject registration eliminates the shape variation within one subject, that is, the variation caused by the head pose. Inter-subject registration aims to remove the differences in shape between many subjects. This is usually done by mapping a subject's face to a reference face. Face registration approaches can also be divided according on their output into a whole-face, part-face (eyes and mouth regions), and point-based approaches (Allaert et al., 2018). Numerous studies have been implemented to align the face in an image of an arbitrary scene. The reader is referred to (Jin and Tan, 2017) for a more complete survey of recent advance in face alignment in the wild.

Fig. 2.4 Illustration of the failure to match features extracted from two images (a frontal face (right) and an orientated face (left)) due to the head pose variations. The features extracted using a general work flow that is often used with facial features extraction of dividing the face into a regular grid of local patches from which features can be extracted.

2.2.3 Studies on Face Representation

Representing the face is another critical requirement for conventional based AFER systems, and is the process of locating and extracting a set of important features from the face image to effectively represent the face for use in further processes. The aim of representing the face, or feature exaction, is to reduce the dimensionality of the problem space, encode

the important features (information) and ignore irrelevant features (details) in order to increase the robustness of the system against other factors such as illumination variation and misalignments. Extracting an efficient feature from the face images is an important step for successful AFER as the classification accuracy is limited by the quality and relevance of the features used in the representation. The best features should maximize between-class variations while minimizing within-class variations. If an inaccurate representation is used, inaccurate prediction will result, even with the best classifier.

Different schemas have been proposed for representing the face for the purpose of expression recognition. The face can in general be described by shape and/or texture. Whilst texture (see section 2.2.3.1) encodes the gray level information, shape (see section 2.2.3.2) encodes the geometric information about the face . In some cases, and for compact representation of hybrid features, the area around the face component's contour can be incorporated into the shape modelling (Cootes et al., 2001; Gao et al., 2010; Matthews and Baker, 2004; Zeng et al., 2009). Recently many researchers have moved to use deep-learning based features for face representation (see section 2.3) (Mollahosseini et al., 2016; Yu and Zhang, 2015).

Each feature has different properties and levels of robustness against different factors. Although geometric features are robust to the variations in illumination since the intensity of the pixels is ignored, they are sensitive to tracking errors. Texture based features are less reliant on initialization and can encode micro patterns in skin texture that are important for facial expression recognition, but they are negatively affected by identity bias, and can be affected by illumination changes. Deep learning approaches have been shown to be useful in many applications. Training a deep network will typically require a huge amount of data, in an expression recognition task it is not uncommon to use some tens of thousands of images. Furthermore, the choice of parameters, topology and training is not yet understood, so implementing a deep learning strategy is still somewhat haphazard. In the following, recent advances in the uses of each feature will be presented.

2.2.3.1 Texture Representation

One way of representing a face is using texture-based feature methods in order to extract the texture features. These methods encode pixel intensity and texture information such as wrinkles, furrows and other patterns that are caused by face deformations associated with an expression. Since a face image contains texture, texture analysis methods have become an active topic in automatic face analysis tasks including face detection, face recognition, and facial expression recognition. One of the key issues in these applications is finding a robust representation/descriptor for several variations in facial appearance. This section reviews the current advances in texture measurement methods for facial expression analysis and recognition.

In recent years, the most notable trend in analysing face images in general and in particular facial expression is using low-level histogram representations based on local descriptors. Examples of these descriptors are the Gabor filter (Lyons et al., 1999), local binary pattern (LBP) (Ojala et al., 2002), local phase quantization (LPQ) (Ojansivu and Heikkilä, 2008), histogram of oriented gradient (HOG) Dalal and Triggs (2005), bag of words (BOW) (Lazebnik et al., 2006), quantized local Zernike moments (QLZM) (Sariyanidi et al., 2013), and binary robust independent elementary features (BRIEF) Calonder et al. (2012, 2010).

The Gabor filter representation is one of the first and most widely used features for facial expression recognition (Glodek et al., 2011; Littlewort et al., 2011; Tong et al., 2010, 2007). Using the Gabor representation, the input image is convolved with several filters of different scales and orientations to describe the spatial structure in the image. The magnitudes of the responses of the filters at each pixel location are combined as a feature vector. In spite of the robustness of this representation in extracting the local features in both the spatial and frequency domains, it suffers from identity bias, high dimensionality and high computational cost. This is because of the convolution of the input image with a set of filters of different scales and orientations. Figure 2.5 shows the Gabor filter bank with different scales and different orientations on the face image.

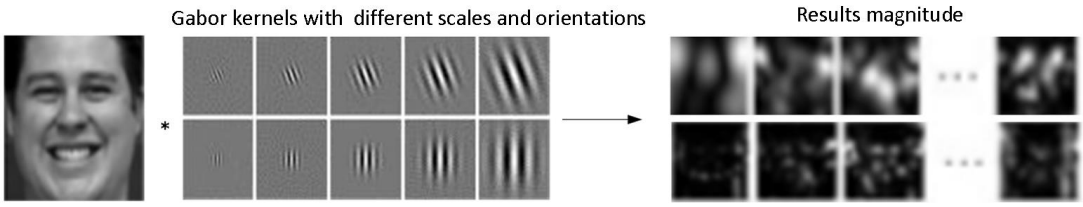


Fig. 2.5 Illustration of Gabor-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2015).

The drawbacks of Gabor representation have motivated researchers to find simpler features and representations such as LBP (Ojala et al., 2002). LBP works by labelling each pixel in the image with a binary number. This number results from thresholding the grey level intensity of neighbourhoods of each pixel with the intensity of the centre pixel. Thresholded values are coded as 0 or 1 and are read systematically to form a binary number which is used then to label each pixel with a decimal number, called an LBP code, which represents the local structure around each pixel. LBP has been shown to be extremely successful in face analysis tasks such as face recognition (Ahonen et al., 2004) and expression recognition due to its robustness to illumination changes (Sandbach et al., 2013; Shan et al., 2009; Valstar et al., 2011; Yang and Bhanu, 2011; Zhong et al., 2012). Shan et al. (2009) find out that the LBP representation outperforms the Gabor representation in a facial expression recognition application. Using LBP the texture is described with varying scales depending on the size of the local area, a larger area will

result in a longer descriptor. This will require more training data to get an accurate distribution. LBP also suffers from some loss of information when extracted from a large regions, as they ignore the pixels that remain inside the circular region due to the sampling points being extracted from the circumference of the circle. Figure 2.6 shows an example of face description using the LBP method.

Fig. 2.6 Illustration of LBP-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2015).

LPQ is another face representation method which was originally introduced by Ojansivu and Heikkilä (2008) as a blur-invariant texture descriptor. The LPQ descriptor is used to extract local phase information using the short term Fourier transform (STFT) computed over a rectangular neighbourhood of each pixel of the image. A histogram of the resulting code words is created and used as a feature in texture classification. LPQ obtained better results than LBP in facial affect recognition, probably owing to its using a larger circular region of 7 pixels diameter than the one used with LBP, and the dimensionality of the descriptor is smaller than for LBP (Cruz et al., 2011; Valstar et al., 2013). Figure 2.7 shows an example of face description using the LPQ method.

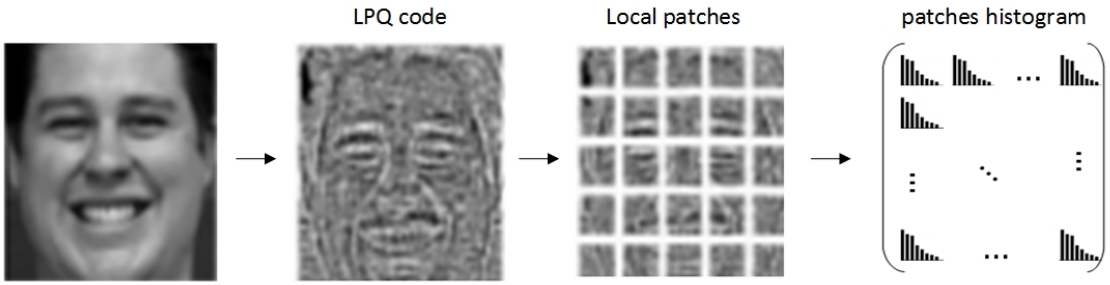


Fig. 2.7 Illustration of LPQ-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2015).

BOW (Lazebnik et al., 2006) is another texture representation method which is very widely used in many computer vision applications. It describes local neighbourhoods by extracting local features densely from fixed locations by dividing the image into sub-regions hierarchically and computing histograms of local features found inside each sub-region. Although this representation is computationally simple, it can have very high

dimensionality. This is because the computation of the visual word is based on a search of the visual vocabulary and depends on the vocabulary size and search algorithm used. Sikka et al. (2012) have successfully used the BOW representation in facial expression recognition and their results outperform the Gabor and LBP methods. Figure 2.8 shows an example of the BOW representation schema on the face image.

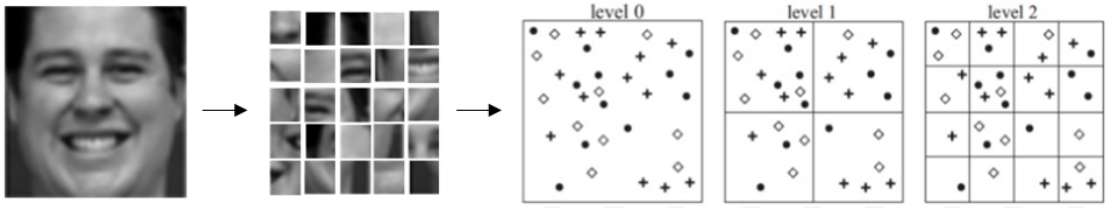


Fig. 2.8 Illustration of BOW-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2015).

Another approach of local representation is HOG proposed by Dalal and Triggs (2005). Using this approach, the image is represented by computing the local features of gradient magnitudes and the edges' orientations. Histograms of these local features are computed from blocks of the image and concatenated to construct a global histogram. Dahmane and Meunier (2011) and Li et al. (2017) have successfully used HOG features in expression recognition tasks and found that HOG is better than LBP owing to its robustness to affine transformations and to illumination changes. Figure 2.9 illustrates the HOG representation schema.

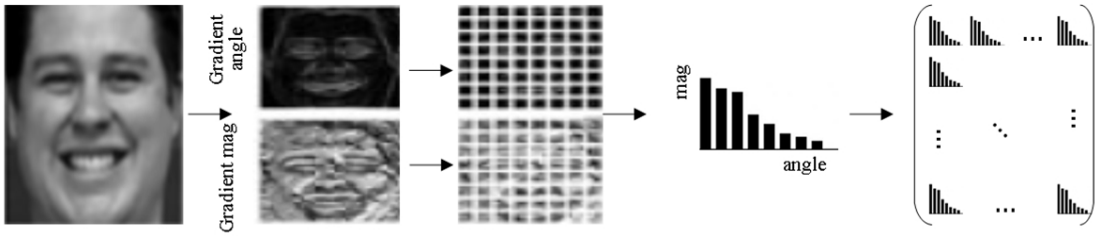


Fig. 2.9 Illustration of HOG-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2015).

Recently, researchers have used QLZM for face representation; QLZM describes a neighbourhood by computing its local Zernike moments. Each moment coefficient describes the variation at a unique scale and orientation (Sariyanidi et al., 2013). In general, moments describe numeric quantities at some distance from a reference point or axis. ZM is used as a global descriptor of the image with explicit shape such character or finger print by calculating the moments from the whole image, whereas in face image applications the method is used locally by calculating the moments around each pixel to obtain a new representation by exposing the intensity variation around each pixel. In (Sariyanidi et al., 2013), QLZM representation achieved higher performance than

the previous representations on the same dataset. Figure 2.10 illustrates the QLZM representation schema.

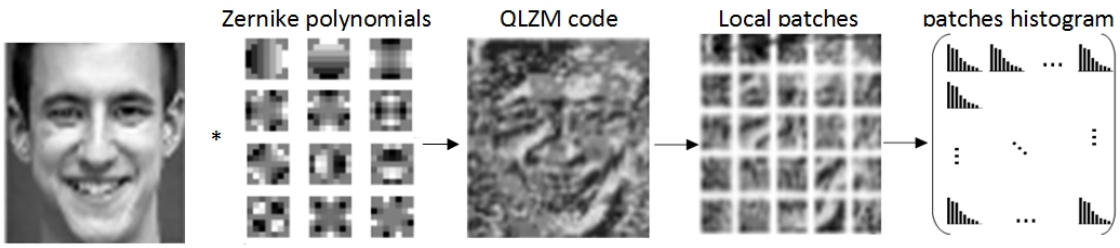


Fig. 2.10 Illustration of QLZM-based face texture representation. This picture is redrawn from (Sariyanidi et al., 2013) .

The above representations are evaluated in AFER tasks using different facial expressions datasets of different characteristics and the results are summarized in Table 2.1. These results show that the best result for recognizing facial expression was achieved using the QLZM representation with an accuracy of 96.1% on the CK+ data set. Despite these good results, the evaluation is limited to a small range of ages and expressions using the MMI (Pantic et al., 2005) and CK+ datasets (Lucey et al., 2010).

In addition to the above texture representation methods, there are many other methods are not tested clearly yet at the task of face description such as BRIEF developed by Calonder et al. (2012, 2010). BRIEF has recently shown remarkable performance in multiple domains including place recognition (Gálvez-López and Tardos, 2012), loop closure detection (Galvez-Lopez and Tardos, 2011), optic disc segmentation in retinal images (Mohammad et al., 2013; Mohammad and Morris, 2017), and real time facial expression recognition application on mobile phones (Alshamsi et al., 2016). Its applicability to multiple domains motivated us to investigate the possibility to use BRIEF features for describing the face and facial expression.

Having understood face representation using texture-based methods and seen the effectiveness of using texture features in automatic facial expression recognition, the next section will focus on face representation using shape-based methods and their effectiveness in face and facial expression modelling.

Table 2.1 A survey of some texture-based AFER systems (UN:unknown, P:point, RR:recognition rate, RMSE:root mean square error, AUs:action units, and BE:basic expression).

Reference	Detection	Registration	Representation	Classification	Dataset	Expressions	Performance
Tong et al. (2007)	Viola-Jones	Rigid (2P)	Gabor	DBN	CK	14 AUs	RR=0.87
Tong et al. (2010)	UN	Rigid (2P) Rigid (28P)	Gabor facial pointtextures	DBN	CK	14 AUs	RR=0.88
Glodek et al. (2011)	UN	UN	Gabor	SVM	AVEC'11	4	RR=0.55
Littlewort et al. (2011)	Viola-Jones	10P	Gabor	SVM	CK+ CERT	7 emotions 26 AUs	RR=90% RR=80%
Shan et al. (2009)	Viola-Jones	Rigid (2P)	LBP	SVM	CK MMI	6 emotions 6	RR=95.1% RR=86.9%
Valstar et al. (2011)	Viola-Jones	Rigid	LBP	SVM	GEMEP	12 AUs	F_1 =0.45
Yang and Bhanu (2011)	Viola-Jones	Rigid	LBP,LPQ	SVM	GEMEP-FERA	5 emotions	RR=0.84
Zhong et al. (2012)	Viola-Jones	Rigid (2P)	LBP	SVM	CK MMI	6 emotions	RR=89.9% RR=73.5%
Sandbach et al. (2013)	UN	Rigid (AAM)	LBP	MRF	DISFA	6 AUs	RMSE=0.34
Li et al. (2017)	Viola-Jones	Rigid (2P)	LBP	SVM	CASMEII SMIC-HS SMIC-VIS SMIC-NIR SMIC-subHS	8 emotions 3 emotions	RR=55.85% RR=57.93% RR=70.42% RR=64.79% RR=77.46%
Cruz et al. (2011)	Viola-Jones	Rigid (2P)	LPQ	SVR	AVEC'13		RMSE=13.61
Valstar et al. (2013)	UN	UN	LPQ	SVM	AVEC'11		RMSE=13.61
Sikka et al. (2012)	Viola-Jones	Rigid (2P)	BOW	SVM	CK+	7 emotions	RR=95.85%
Sariyanidi et al. (2015)	Viola-Jones	Rigid (2P)	HOG	SVM	GEMEP-FERA	5 emotions	RR=70%
Li et al. (2017)	Viola-Jones	Rigid (2P)	HOG	SVM	CASMEII SMIC-HS SMIC-VIS SMIC-NIR SMIC-subHS	8 emotions 3 emotions	RR=57.49% RR=57.93% RR=71.83% RR=63.38% RR=80.28%
Sariyanidi et al. (2013)	Viola-Jones	Rigid (2P)	QLZM	SVM	CK+	7 emotions	RR=96.1%

2.2.3.2 Shape Representation

Another way of representing the face is by using shape-based methods which extract the shape features. In these methods, the face's shape feature is found using a set of landmark points located around the face components such as the eyes, eyebrows, nose, mouth, and chin. There are different approaches to extracting the face shape features based on facial points. For instance, Pantic et al. (2012) and Lucey et al. (2007) have described the face image using the coordinates of 20 points and 74 points respectively, where the x and y coordinates are concatenated to form the feature vector that represents the face's geometry. Another way to extract the geometric features of the face images is to compute the distance between the landmarks instead of using the coordinates themselves. For example, Huang et al. (2010) computed the angles of opening and closing mouth or eyes, and extract an optimal set of triangular facial features to use for emotion recognition as described in Figure 2.11. Historically a widely used scheme to describe the face shape and appearance is to build an Active Shape Model (ASM) Cootes et al. (1995) or Active Appearance Model (AAM) Cootes et al. (2001) and then use these models to extract the face features. For instance, Lucey et al. (2007) extracted the shape and appearance features using ASM and AAM respectively. The identity bias is then reduced by subtracting a feature of emotion from the features of the neutral face as shown in figure 2.12.

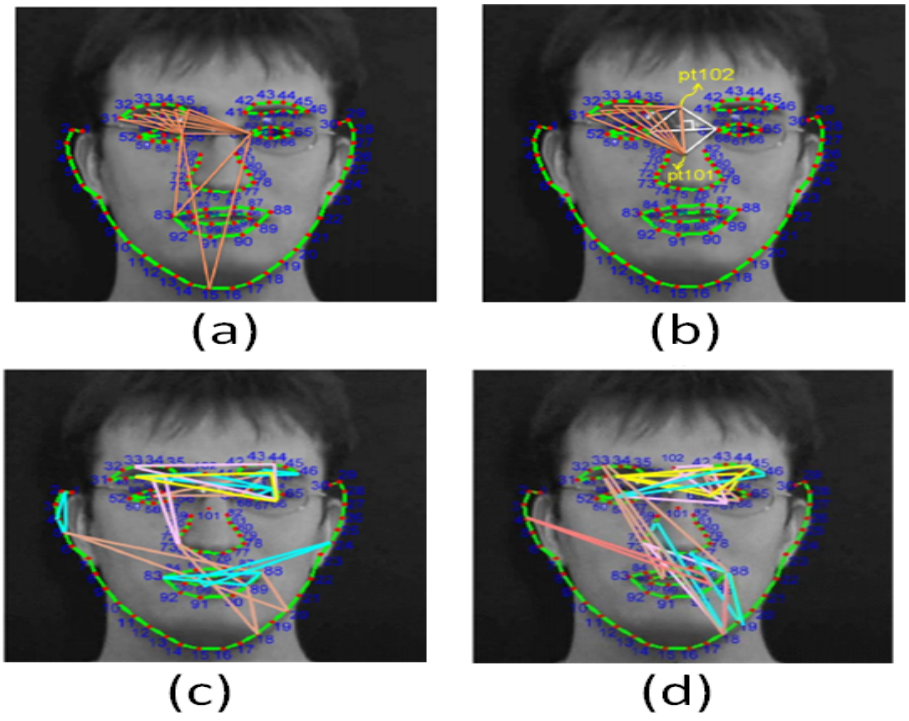


Fig. 2.11 Example triangles chosen by statistical analysis: (a) Example triangles of horizontal base, (b) Examples triangles of vertical base, (c) 12 optimal triangular features, and (d) The other 13 optimal triangular features. This figure is adapted from (Huang et al., 2010).

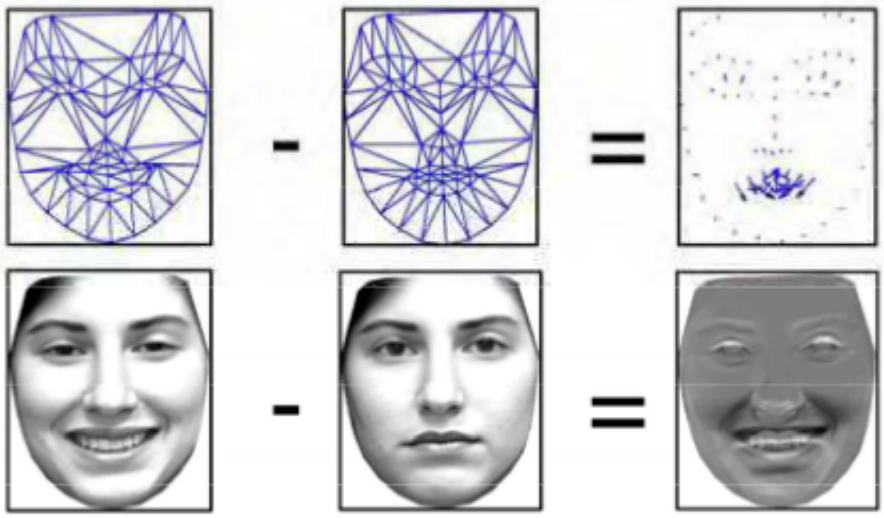


Fig. 2.12 This figure depicts a visualization of shape (row 1) and appearance (row 2) features. The peak and neutral frames for these different features can be seen in the left and the middle columns respectively. The difference between the emotion features and the neutral features can be seen in right column. This figure is adapted from (Lucey et al., 2007).

The use of the shape-based feature methods is dependent on the availability of the landmark points and their locations (x and y coordinates). Since the process of annotating each image manually with the required of landmark is time-consuming, laborious and expensive, an accurate and automatic facial features points detector (FFPD) is required which is difficult to accommodate in many situations. The next section reviews the existing methods used for FFPD.

2.2.3.3 Facial Feature Points Detection

FFPD is the process of using a facial landmark detector as an automated tool to predict the position of every predefined feature in every image. This process is a useful component for several expression analysis tasks, especially those involving large datasets and high accuracy requirements, in which annotating each image manually is time-consuming, laborious and expensive. The benefit of automation is due to the use of the detected points to extract both the shape and texture features from a patch around each point or from the whole face image to use for further processing such as expression classification and synthesis. Detecting the facial feature points is a challenge task owing to rigid face deformations (scaling, rotation, translation) and non-rigid face deformations (expression, ageing). Usually, the error of rigid face deformation is suppressed by registering the face as a whole entity. Then the error of non-rigid face deformation can be suppressed by using local registration (facial feature point detection) techniques.

FFPD or face alignment has a long history in computer vision, and many methods have been proposed by researchers to solve the problem. It started with ASM Cootes et al.

(1995), in which the pre-trained statistical shape model (SSM) from a set of images and their points is used to fit the shape model to the initial position of all the points in a search image. The ASM model has been expanded to the AAM Cootes et al. (2001), in which the model of shape variation is combined with a model of the appearance variation. ASM and AAM are widely and successfully cited for facial feature keypoint detection in the literature, especially in a controlled scenario (Anderson et al., 2013; Asthana et al., 2011; Fanelli et al., 2013; Hansen et al., 2011; Huang et al., 2012; Kinoshita et al., 2012; Martins et al., 2010, 2013; Tresadern et al., 2012, 2010; Tzimiropoulos et al., 2012; Tzimiropoulos and Pantic, 2013). These methods failed to localize the facial feature points in an uncontrolled scenario such as the presence of wide range of pose and expression appearances. This is because ASMs are very sensitive to the initial points' positions, particularly in facial images with a wide variation of appearances, e.g. due to ageing and expressions, which prevent putting the points on the right positions of the object edge due to the non relevant structures regarding the age and the expression which make the intensity patterns not being sufficiently distinctive across the subjects. AAMs are very sensitive to the texture variation between the training and testing images such as might be due to illumination and pose changes. With respect to facial images, this might relate to texture's variations due the presence of age and expressions' appearances. Although several generative (Baker and Matthews, 2004; Matthews and Baker, 2004) and discriminative (Liu, 2009; Saragih and Göcke, 2009; Saragih and Goecke, 2006, 2007) modifications have been made to the AAM, these methods have been shown to rely heavily on accurate initialization and data dependent experiments.

Motivated by the limitations of ASM and AAM, Cristinacce and Cootes (2006) have solved the problem by proposing the Constrained Local Model (CLM), where the SSM and the local texture model for each point are combined in a single model, in which the appearance variation around each point is considered independently and one response image can be computed for each point and used to predict the final positions for all the points. The shape model, generated from the training data, is then used to match to the predicted points, selecting the overall best combination of points. The CLM was extended in (Saragih et al., 2011) by proposing a fitting method, known as the Regularized Landmark Mean Shift (RLMS), which exhibited superior performance over AAM in terms of landmark detection accuracy and is considered to be among the state-of-the-art methods for the generic face fitting scenario. Figure 2.13 illustrates the two main steps of the CLM algorithm, where an exhaustive local search is first performed to obtain a response map for each landmark. Optimization is then performed over these response maps, which admits more sophisticated strategies compared to generic optimization methods that make no use of domain specific knowledge.

Recently the researchers' focus shifted to cascaded-regression-based models such as in (Asthana et al., 2013; Cao et al., 2014; Kazemi and Josephine, 2014; Lee et al., 2015;

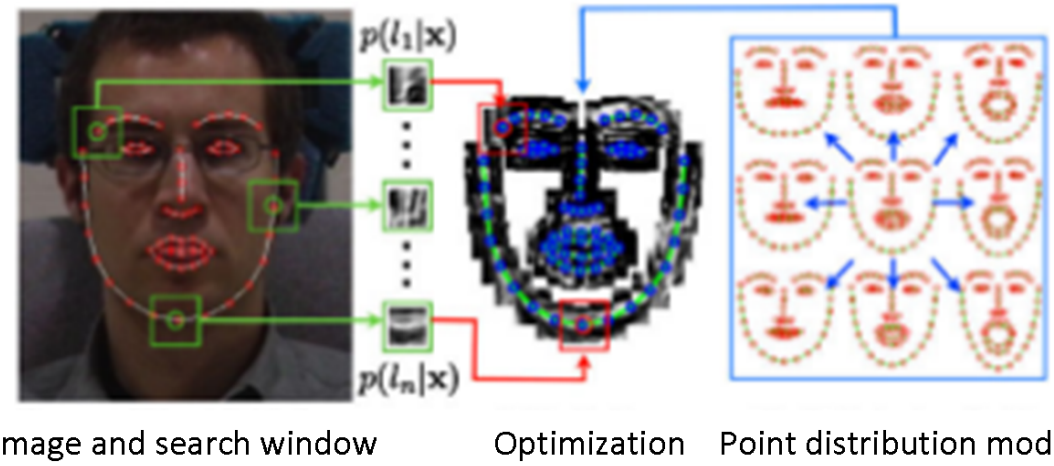


Fig. 2.13 Illustration of Constrained Local Model (CLM) search algorithm. This figure is adapted from (Saragih et al., 2011).

Ren et al., 2014; Tuzel et al., 2016; Valstar et al., 2010; Xiong and De la Torre, 2013) and deep-learning-based methods such as in (Bulat and Tzimiropoulos, 2016; Fan and Zhou, 2016; Trigeorgis et al., 2016; Xiao et al., 2016; Zhou et al., 2013).

Cascaded-regression-based methods have shown remarkable performance in facial feature keypoint detection especially with a large range of facial activities appearances (Wang et al., 2018). The key idea of these methods is the use of a regressor instead of a classifier to vote for the point position based on the information in nearby regions. Despite learning the regressor being more difficult than learning a classier, the regressor can provide more useful information about the target location, such as the distance of negative patches from the positive patch, while the classifier only determines whether the image patch is negative or positive. In recent years, random forests (RF) Breiman (2001) have proven to be an important tool for solving challenging computer vision tasks efficiently, such as anatomy detection and localization Criminisi et al. (2013, 2010), human pose estimation from single depth images Shotton et al. (2013), and object segmentation by Schroff et al. (2008).

Cootes et al. (2012) and Lindner et al. (2015) introduced random forest regression voting (RFRV) to the constrained local models (CLM) for the application of locating points on deformable objects and found that the RFRV-CLM framework outperforms the alternative discriminative methods of classification-based (Belhumeur et al., 2011) and boosted regression-based (Valstar et al., 2010; Yang and Patras, 2013) trained on the same datasets. It has been applied successfully to automatic landmark points localization in face (Lindner et al., 2015) and clinical (Bromiley et al., 2015a,b, 2016; Lindner and Cootes, 2015; Lindner et al., 2013) images. The reader is referred to Çeliktutan et al. (2013) and Wang et al. (2018) for a complete description and discussion of FFPD and its literature.

The success of the RFRV-CLM framework in many computer vision applications motivated us to introduce the RFRV-CLM method in this thesis to build an automatic FEL system to use for facial expression points localization and feature extraction.

2.2.4 Studies on Expressions Classification

The last module in any conventional-based AFER system is the expression classification module, in which a learning model is used to learn a description that is subsequently used for output prediction. Typically, the output from such a system is a label of the emotion or a label of an AU. The target label can be divided into frame-based labelling or sequence-based labelling. In frame-based labelling, a separate label is given for every frame, while in sequence-based labelling one label is given to all frames of the sequence. The performance of the system depends on the quality and the quantity of the training features and the method used for the learning. In general, there are three categories of learning methods, depending on what to learn (Alpaydin, 2010; Mitchell et al., 1997; Sutton et al., 1998). These categories are:

- **Supervised learning:** The model trains on labelled data to make predictions on the new, unseen data such as **classification** in which the class of a new example is categorized based on the pre-defined classes and **regression** in which a value is predicted instead of the class.
- **Unsupervised learning:** The model trains on unlabelled data to learn and find patterns to categorize and represent the data. Examples are clustering (find classes in the data) and dimensionality reduction (compress the data to best represent the patterns).

In this thesis, the focus is on supervised learning (regression and classification) for facial expression localization and recognition. A difficulty in these categories is how to find a method to provide a balance between memorising (using the training data to have knowledge) and generalising (transferring the properties of this knowledge to unseen data). Many possible classification techniques, or classifiers, have been successfully used for expression classification such as support vector machines (SVM) (Vapnik and Vapnik, 1998), dynamic Bayesian networks (DBN) Tong et al. (2008, 2010, 2007), neural networks (NN) and random forests (RF) (Breiman, 2001).

One of the most widely used classiers for AFER is the SVM (Vapnik and Vapnik, 1998) due to the fact that in binary classification problems the SVM guarantees maximum-margin separation with linearly separable and non-separable data (see Figure 2.14). SVMs were originally developed for binary classification. The method was then modified to deal with multiple classes classification. Currently there are two types of schemes for multiple classes SVM. The first scheme is implemented by constructing and combining several

binary classifiers (one-against-all), while the second scheme is performed by directly considering all data in one optimisation formulation (one-against-one). SVM offer specific advantages: the global optimality of the training algorithm, the existence of excellent data-dependent generalization, and its success in non-separable cases. These advantages made the SVM classifier a widely used tool for facial expression recognition as assessed in many papers (Cruz et al., 2011; Dahmane and Meunier, 2011; Glodek et al., 2011; Li et al., 2017; Littlewort et al., 2011; Lou et al., 2018; Sariyanidi et al., 2013; Shan et al., 2009; Sikka et al., 2012; Valstar et al., 2013, 2011; Yang and Bhanu, 2011; Zhong et al., 2012) (see Table 2.1). A complete description of the theory, parameter choosing, and implementation of SVM methods can be found in the tutorial by Chang and Lin (2011).

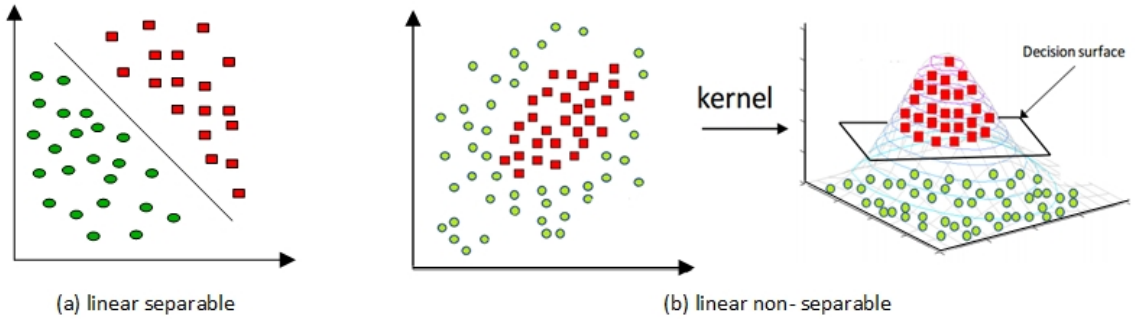


Fig. 2.14 Illustration of separating hyperplanes using SVM classifier in separable and non-separable cases

Despite the success of SVM, it does not extend naturally to problems with any number of classes (Crammer and Singer, 2001; Torralba et al., 2007). On the other hand, Breiman (2001) has discovered that the random forest decision trees work naturally (without modification) with any number of classes. For instance, in Shotton et al. (2011) and in Shotton et al. (2008) the random forest classifier has tested with ≈ 30 and ≈ 20 classes respectively. Moreover, an empirical evaluation in (Caruana et al., 2008) proved that forests have shown good generalization, even with the problem of high dimensional data. Furthermore, classification using RF has been applied successfully in a number of practical applications and seems to be promising as a classification scheme as assessed in many papers (Criminisi et al., 2009; Dapogny et al., 2015; Jia et al., 2016; Lepetit and Fua, 2006; Pu et al., 2015; Rogez et al., 2008; Shotton et al., 2011). The fundamental idea behind a random forest is to combine many decision trees into a single model. Decision trees were originally developed by Breiman (1984); Quinlan (1993) and have been around for a number of years. Breiman (2001) revived the method by discovering that ensembles of different trees in a single model tend to produce a much higher performance (a phenomenon known as a generalization) on the new data than the single decision trees. Each decision tree in the forest considers a random subset of features and only has access to a random set of the training data points. This increases diversity in the forest leading to more robust overall predictions. Each decision tree in the forest is used to make decisions. When RF is

used to make a prediction, it takes an average of all the individual decision tree estimates. Therefore, a random forest predication is better than a single decision tree prediction. See Figure 2.15 for an illustration of the RF method.

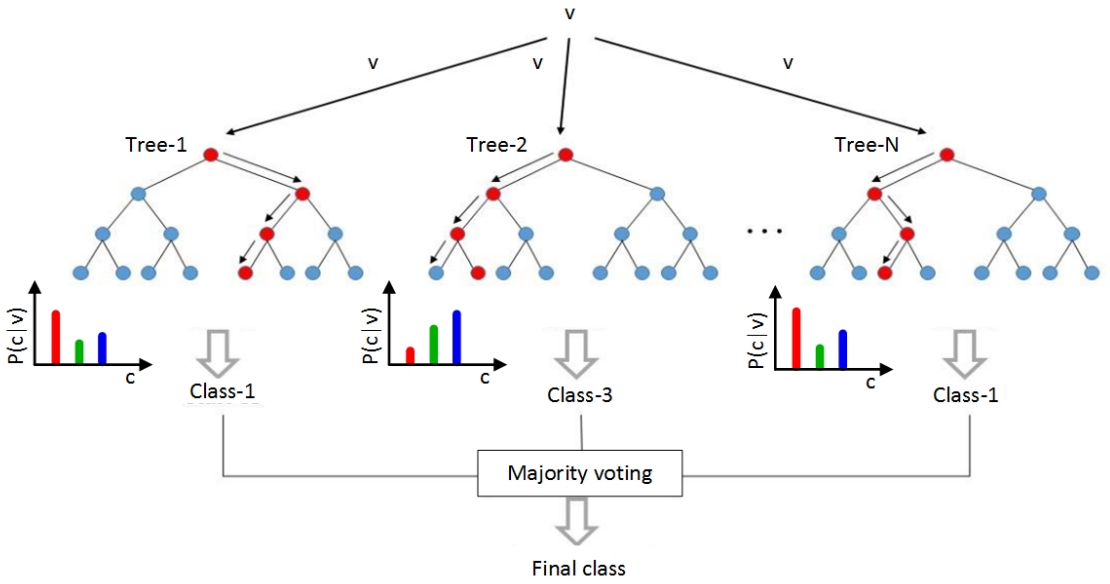


Fig. 2.15 Illustration of RF classification task: v is the input vector, c is the class type, and p(c|v) is the probability that vector v belonging to class c.

2.3 Deep-Learning Based AFER Approaches

Recently, deep learning (DL) has emerged as a general and alternative approach to the handcrafted features in many computer vision applications. The DL based AFER approach enables an end-to-end learning framework from a single image as described in Figure 2.16. In the figure, the input image is convolved with a collection of filters in the convolution layers to produce the feature map. The feature map is then combined in a fully connected network where the expression is recognised as belonging to a particular class (Walecki et al., 2017). Generally, convolutional neural networks contain three types of layers: convolution layers, max pooling (down-sampling) layers, and fully connected layers. The convolution layers take an input image or a feature map to produce a feature map that represents a spatial arrangement of the facial image. The max-pooling layers reduce the resolution and the dimensionality of the feature map by averaging or by max-pooling. The fully connected layers compute the class score or the class type (LeCun et al., 1989).

DL-based AFER approaches have become feasible owing to the large amount of data available to train the learning methods and the advances in graphics processing unit (GPU) technology. Recently, many facial expression recognition systems have used deep representation such as the systems developed by Breuer and Kimmel (2017); Ebrahimi Kahou et al. (2015); Graves et al. (2008); Hasani and Mahoor (2017); Jain et al. (2017); Jung et al.

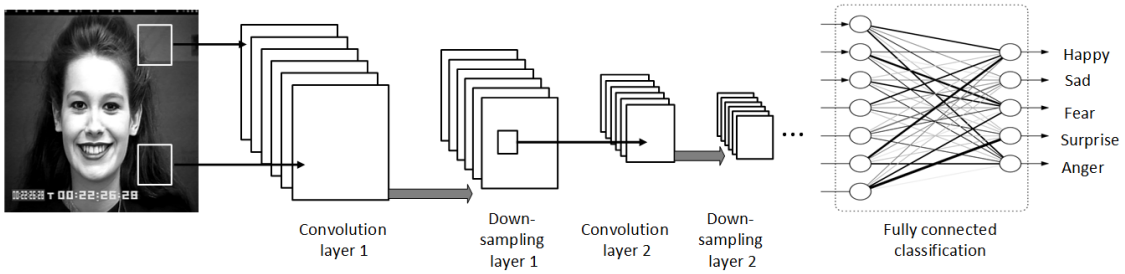


Fig. 2.16 Framework of the deep-learning-based AFER approach

(2015); Kim et al. (2017); Zhao et al. (2016). Although the usefulness introduced by the deep features in solving facial expression recognition problem has been demonstrated, they require a huge amount of data and power to implement and train (Liu and Deng, 2015; Shwartz-Ziv and Tishby, 2017). Since the datasets used in this thesis are small as described in Section 3.2, in this thesis the DL framework is described briefly since is not the focus of this thesis.

2.4 Challenges

This section gives a particular emphasis on the recent advances and challenges to the performance of AFER system. These challenges are human ageing, compound emotions and the expression's intensity.

2.4.1 Studies on Ageing and Expression Modelling

Although there is a small number of recognisable expressions as described in the previous sections, psychologists have recently observed that the age features (age-related structure) have significant effects on human understanding of those expressions (Ebner and Johnson, 2009, 2010; Hess et al., 2012; Houstis and Kiliaridis, 2009; Mary and Jayakumar, 2016; Wang et al., 2016). To increase the generality of the psychological studies' results, Minear and Park (2004) collected a lifespan dataset of human faces, with ages ranging from 18 to 94 years. Ebner et al. (2010) captured another dataset called FACES with six expressions and the age range of subjects from 19 to 80 years. Dibeklioğlu et al. (2012) also captured a dataset called NEMO, from subjects whose ages range from 8 to 76 years. Each subject recorded several videos of their expression changing from neutral to happy, spontaneously and also deliberately.

Using these datasets, Guo et al. (2013) have realised that the expression performed by old people is different from the expression performed by young people. In other words, "The expressions of older adults are not so exaggerated as the young people" (Guo et al., 2013). Another observation about the age effects on the facial expression recognition is that

the wrinkles, folds, and reduction of facial muscles' elasticity can change the appearance of the face and hence effect the expression' meaning as well.

Human age estimation and facial expression recognition are active research topics in computer vision and machine learning, and they are related to each other since both tasks start from the same face image and use the same feature' pattern. However, the facial expression recognition results are easily affected by the age pattern and vice versa. This is because both age and facial expression appearances overlap: they appear in a similar way, resulting in misclassification in both tasks. For example, the fold between the cheek and upper lip can appear in the happy expression of young people and the neutral expression of old people as described in 1.3 and hence the neutral expression of old people might be recognized as a happy expression and vice versa. Recently researchers have started to analyse and study the interrelationship between age and facial expression in order to achieve better results for both tasks. Some studies have focused on analysing the effect of facial ageing and facial expression features on the performance of the age estimation task (Guo and Wang, 2012; Osman and Yap, 2018). In this thesis the focus is on reviewing the current researches in AFER under the age pattern effect as the concern of this thesis is facial expression recognition and not age estimation. It is possible to divide the survey of this section into *real-age* based automated facial expression recognition and *apparent-age* based automated facial expression recognition, as recent studies found that the apparent age of someone's face can be different from the real age.

2.4.1.1 Real Age-Based AFER

Real age is the real chronological age of a person. To the best of our knowledge, there are very few studies modelling the real age and the expression together, for example published by Dibeklioğlu et al. (2015); Guo et al. (2013); Lou et al. (2018); Wang et al. (2016); Yang et al. (2018). Guo et al. (2013) were the first who studied and analysed the influence of age features on the performance of expression recognition. In their study, the subjects were divided into four age groups and he considered each expression in each age group as a separate class. For facial feature extraction, they manually labelled 31 fiducial points and applied Gabor filters at the location of those points and features were extracted from which a kernel SVM was trained to recognize expressions. Figure 2.17 illustrates the response of the Gabor filters for both age and expression. They then proposed to remove the age features such as wrinkles and furrows using image smoothing techniques as depicted in Figure 2.18.

Dibeklioğlu et al. (2015) studied the usefulness of using age features along with the other features in distinguishing between the posed and spontaneous smile expression. Experimental results using the NEMO dataset proved that using the age feature significantly helped to differentiate between the posed and spontaneous expression. Wang et al. (2016) proposed a propagandistic model using a Bayesian network (BN) to classify the expression

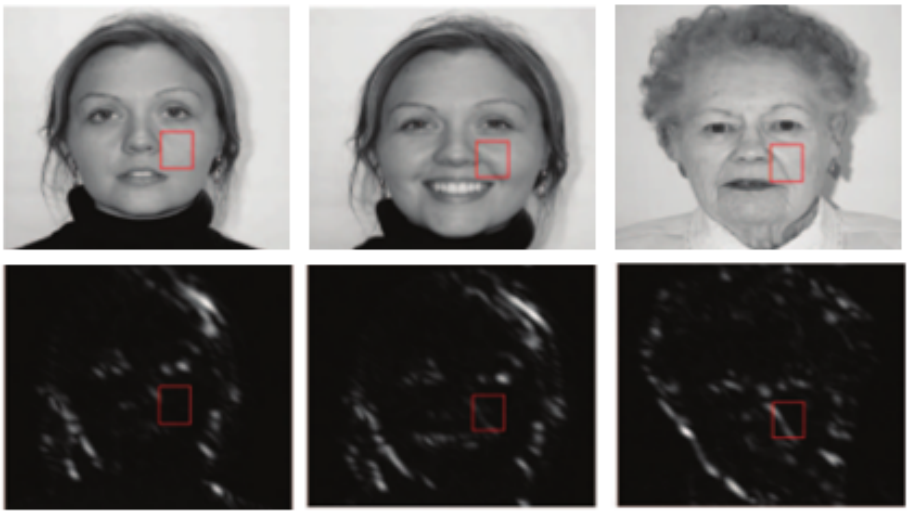


Fig. 2.17 Visualization of the Gabor filter response for expression and ageing. This figure is adapted from Guo et al. (2013)

Fig. 2.18 Visualization of ageing details removal from the face image. This figure is adapted from Guo et al. (2013).

with the help of age features. For facial expression feature extraction, they manually labelled 18 fiducial points and geometric features were extracted from which a BN was trained to recognize expressions. They proposed then to used multiple BNs to capture the spatial information of expression patterns. Experimental results on the FACES and LifeSpan datasets demonstrated that the proposed model of using geometric features only has successfully recognized the expression with comparable performance to the previous methods of using texture features only. In an attempt to improve the age estimation system's performance against the impact of expression related structure, two methods were very recently introduced by Lou et al. (2018) and Yang et al. (2018). These methods are evaluated also in recognizing the expression against the impact of ageing features. In Lou et al. (2018), a graphical model to jointly learn age and expression classes with a latent layer between the age and the expression is proposed as depicted in Figure 2.19. For feature extraction, the LBP features as described in Ahonen et al. (2006) are used. A multi class support vector machine (MC-SVM) is used for age and expression classification. Evaluation results for the age estimation task show an improvement in the performance when the age is jointly learnt with expression in comparison to the age estimation system

which ignores expression. The results also showed an improvement in the facial expression recognition performance when the age is jointly learnt with expression in comparison to the facial expression recognition system which ignores age.

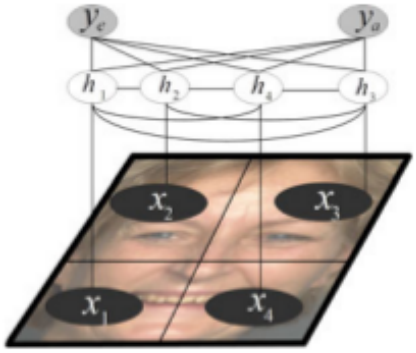


Fig. 2.19 The graphical model to jointly learn the age and the expression. This figure is adapted from Lou et al. (2018). x is the feature vector, h is the latent variables, y_a and y_e are the corresponding age and expression respectively.

In Yang et al. (2018), a deep multi task learning model is proposed which consists of two parallel columns composed of ConvNet and ScatNet, two fully connected layers, and an output layer as shown in Figure 2.20. ConvNet and ScatNet provide feature representations shared by the subsequent tasks. The multi-task learning formulation is employed to simultaneously learn to predict age and to classify expression.

All the methods proposed in this section are evaluated using three age-expression datasets: FACES (Ebner et al., 2010), LifeSpan (Minear and Park, 2004), and/or NEMO (Dibeklioğlu et al., 2012) datasets, and the results are summarized in Table 2.2.

Despite the usefulness that was brought by the previously mentioned studies to the AFER system to analyse the facial expressions across a large range of ages, they have some limitations. The first limitation is that the systems developed by Guo et al. Guo et al. (2013) and Wang et al. Wang et al. (2016) relied on the manually placed points to extract the texture features only as in Guo et al. (2013) or the geometric features only as in Wang et al. (2016), so was not automated and the demands of any automatic system are to generate those points automatically which is a difficult task especially in the presence of non-rigid face deformations related to age and expression.

Despite the usefulness that was brought by the previously mentioned studies to the AFER system to analyse the facial expressions across a large range of ages, they have some limitations (Dibeklioğlu et al., 2015; Guo et al., 2013; Lou et al., 2018; Wang et al., 2016; Yang et al., 2018). The first limitation is that the systems developed by Guo et al. (2013) and Wang et al. (2016) relied on the manually placed points to extract the texture features only as in Guo et al. (2013) or the geometric features only as in Wang et al. (2016), so was not automated and the demands of any automatic system are to generate those points automatically which is a difficult task especially in the presence of non-rigid face

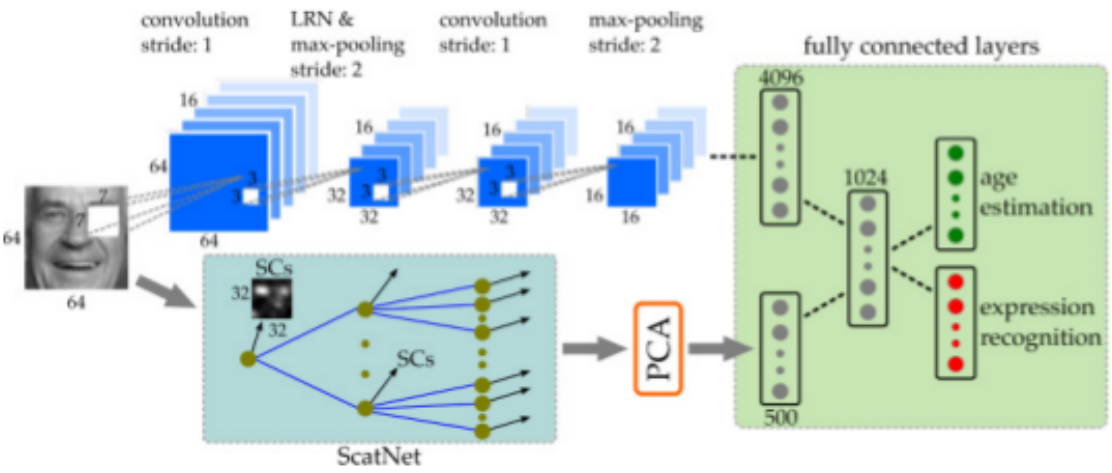


Fig. 2.20 An overview of the deep network model for age and expression estimation. This figure is adapted from Yang et al. (2018)

Table 2.2 An overview of the recent advances in expression recognition under the age effect.

Reference	Features	classifier	FACES	Lifespan	NEMO
Guo et al. (2013)	Gabor-filters representation	SVM	97.89	96.79	-
Wang et al. (2016)	Geometric representation	DBN	95.74	96.53	-
Lou et al. (2018)	LBP representation	MC-SVM	92.19	93.68	98.0
Yang et al. (2018)	Deep representation	CNN	95.13	96.32	-

deformations related to age and expression. The second limitation is that using the texture features only, such as those derived from the Gabor filter responses as in Guo et al. (2013) or LBP as in Lou et al. (2018), can encode micropatterns in skin texture that are important for age estimation and expression recognition but are negatively affected by identity bias. On the other hand, using the geometric feature only as in Wang et al. (2016) might be sensitive to registration error. Extracting deep features from a small dataset, in FACES datasets there are approximately 114 images per expression per age, is not enough to make a convincing conclusion since there is a concern that deep representation requires a huge amount of data train and power to implement Liu and Deng (2015); Shwartz-Ziv and Tishby (2017). For example, in Yang et al. (2018) the pre-trained model on the MORPH dataset of 55,134 images was used to extract the age and expression features for age estimation and expression recognition. Lower accuracy might be obtained if the deep learning model was trained on the FACES dataset only. The other limitation is that one way to eliminate the effect of age on the performance of expression recognition is to delete the age information using image smoothing techniques before expression recognition as in Guo et al. (2013). Using image smoothing techniques might lead to the loss of information related to age and expression since there is a concern that image smoothing may cause loss

of some image information. Another way is by using age information as prior information to the expression recognition as in Dibeklioğlu et al. (2015); Lou et al. (2018); Wang et al. (2016), in which the age information is required during training and testing as in Dibeklioğlu et al. (2015), or to avoid the error in age estimation and by subsequent on the expression recognition. In addition, all the previous studies are reliant on the real age (chronological age) for dividing the data into groups and then for training and testing the expression recognition classifiers; an error can occur if there is a difference between the real age (chronological age) and apparent age; since recently there is a concern that the apparent age of the individual might be different from his/her real age Antipov et al. (2016); Huo et al. (2016); Rothe et al. (2015); Uřičař et al. (2016); Zhu et al. (2015)

2.4.1.2 Apparent Age-Based AFER

Recently, apparent age (how old does a person look) has been used as a new measurement for the age estimation task. The only difference between real age and apparent age is that the label of the apparent age is provided by an assessor whereas the label for the real age is the chronological age of the person. The ChaLearn Looking at People (LAP) competition on apparent age estimation was the first study on apparent age estimation and was conducted by Escalera et al. (2015). The second edition of this competition was in (Escalera et al., 2016), where the organizers won the first place in the competition. In this competition, the organizers have collected a dataset of face images and developed a web service to use by other people to label each image in the dataset with an apparent age. After that, many researchers have used this dataset of the apparent age label for automatic age estimation (AAE) task and showed that the apparent age is different from the real age and has a significant effect on the AAE performance (Antipov et al., 2016; Huo et al., 2016; Liu et al., 2015; Rothe et al., 2015, 2018; Uricár et al., 2016; Zhu et al., 2015). Figure 2.21 illustrates age estimation using real and apparent age.

Fig. 2.21 Illustration of the real and apparent age prediction. This figure is adapted from Rothe et al. (2018)

Inspired by this idea, in this thesis the apparent age is introduced into AFER with the hope of presenting a novel and powerful method that can be used to eliminate the effect caused by ageing features on the performance. The underlying assumption is that the advantages that are brought by the apparent age to the age estimation task might by consequence enhance the performance of the facial expression recognition task against the impact of ageing since both tasks are correlated.

2.4.2 Studies on Compound Emotions Modelling

Most of the previous reviewed studies have focused on facial expressions defined by a single component emotion such as happy, anger, sad, surprise, fear, disgust plus the neutral as described above.

Recently the focus has moved from universal (basic) expressions to the non-universal (non-basic) expressions such as the pain recognition proposed by Lucey et al. (2012) and compound emotion classification proposed by Du et al. (2014).

Du et al. (2014) demonstrated that in addition to the 6-basic expressions there are 15-non basic expressions generated by combining two or more of the basic expressions. They found also that the Facial Action Coding System (FACS) of 22 compound emotions is different but consistent with the six basic categories. The aforementioned study demonstrated that those differences are sufficient to be able to distinguish between 22 compound emotions and that most of these categories are also visually discriminable from one another. Table 2.3 describes the 22 basic and compound emotions and their AUs.

The EmotioNet challenge developed by Benitez-Quiroz et al. (2017) is the first to assess the ability of computer vision algorithms in the automatic analysis of a large number of images of facial expressions of emotion in the wild. The challenge was divided into two parts. The first part tested the ability of current computer vision algorithms in the task of automatic detection of 11 AUs. The second part tested the ability of the algorithms in the task of recognizing 16 basic and compound emotion categories. The results of the challenge demonstrate that existing computer vision and machine learning algorithms are not ready to reliably solve these two tasks. In summary, the results of these studies suggested that research is needed to determine if current algorithms need modification, or whether a novel set of algorithms is required in order to be reliable in performing these tasks.

Table 2.3 Prototypical AUs observed in each basic and compound emotion category, adapted from Du et al. (2014). AUs used by a subset of the subjects are shown in brackets with the percentage of the subjects using this less common AU in parentheses. The underlined AUs listed in the compound emotions are present in both their basic categories. An asterisk (*) indicates that the AU does not appear in either of the two subordinate categories. This table is adapted from (Du et al., 2014)

NO.	Emotion Name	Prototypical and variant AUs
a	Neutral	-
b	Happy	12, 25 [6 (51%)]
c	Sad	4, 15 [1 (60%), 6 (50%), 11 (26%), 17 (67%)]
d	Fearful	1, 4, 20, 25 [2 (57%), 5 (63%), 26 (33%)]
e	Angry	4, 7, 24 [10 (26%), 17 (52%), 23 (29%)]
f	Surprised	1, 2, 25, 26 [5 (66%)]
g	Disgusted	9, 10, 17 [4 (31%), 24 (26%)]
h	Happily surprised	1, 2, 12, 25 [5 (64%), 26 (67%)]
i	Happily disgusted	10, 12, 25 [4 (32%), 6 (61%), 9 (59%)]
j	Sadly fearful	1, 4, 20, 25 [2 (46%), 5 (24%), 6 (34%), 15 (30%)]
k	Sadly angry	4, 15 [6 (26%), 7 (48%), 11 (20%), 17 (50%)]
l	Sadly surprised	1, 4, 25, 26 [2 (27%), 6 (31%)]
m	Sadly disgusted	4, 10 [1 (49%), 6 (61%), 9 (20%), 11 (35%), 15 (54%), 17 (47%), 25 (43%)*]
n	Fearfully angry	4, 20, 25 [5 (40%), 7 (39%), 10 (30%), 11 (33%)*]
o	Fearfully surprised	1, 2, 5, 20, 25 [4 (47%), 10 (35%)*, 11 (22%)*, 26 (51%)]
p	Fearfully disgusted	1, 4, 10, 20, 25 [2 (64%), 5 (50%), 6 (26%)*, 9 (28%), 15 (33%)*]
q	Angrily surprised	4, 25, 26 [5 (35%), 7 (50%), 10 (34%)]
r	Angrily disgusted	4, 10, 17 [7 (60%), 9 (57%), 24 (36%)]
19	Disgustedly surprised	1, 2, 5, 10 [4 (45%), 9 (37%), 17 (66%), 24 (33%)]
s	Appalled	4, 10, [6 (25%)*, 9 (56%), 17 (67%), 24 (36%)]
t	Hatred	4, 10, [7 (57%), 9 (27%), 17 (63%), 24 (37%)]
u	Awed	1, 2, 5, 25, [4 (21%), 20 (62%), 26 (56%)]

2.4.3 Studies on Expression's Intensity Modelling

Most systems designed to automatically recognize the facial expression seek to find a discrete expression of the six basic expressions or discrete facial action of the 44 AUs. Recently, researchers' focus has moved to find the expression's intensity or the facial action intensity. As stated before, each expression is a combination of several muscles' movements. In the case of continuous intensity, each movement is specified with a small score when the movement is subtle and a large score when the movement is pronounced.

Studies proved that expression's intensity plays an important role in the meaning of the expression. For instance, Prkachin and Solomon (2008) found that the intensity of the expression has an important role in determining the pain level in the application of shoulder pain detection. They used the summation of the intensity of four AUs to determine the pain intensity level. These AUs are brow lowering, levator contraction, orbital tightening and eye closure.

Some of the past work in the field of facial expression recognition has introduced methods used for recognition of facial expression intensity level (Chang et al., 2006; Gunes and Piccardi, 2009; Hess et al., 1997; Kim and Pavlovic, 2010; Koelstra et al., 2010; Li et al., 2013; Lien et al., 2000; Shan et al., 2006; Valstar and Pantic, 2012; Yang et al., 2009).

In (Hess et al., 1997), the author used the difference between facial images presented at different stages and neutral or relaxed expression to determine the expression intensity as relative to the difference corresponding to the neutral frame.

In (Chang et al., 2006; Li et al., 2013; Lien et al., 2000), the AFER system can recognize facial expressions using the dynamic features through a probabilistic graphical model. In (Lien et al., 2000), the system used Hidden Markov Models (HMMs) to classify the expression over time using the AU or the AU combination that maximized the likelihood of the extracted facial feature. In (Chang et al., 2006), an embedded manifold is applied on the face features to embed the face representation from the high-dimensional space into a low-dimensional space and the learned model is used for testing. Figure 2.22 illustrates a 3D expression manifold. In (Li et al., 2013), a dynamic Bayesian network (DBN) is used to simultaneously and coherently represent the facial activity. This DBN is constructed using three different levels of facial activities Bottom level: facial feature points, middle level: facial Action Units (AUs), and top level: six basic facial emotions. The main drawback of using a probabilistic model is the requirement for a huge amount of data of facial expressions. For instance, (Chang et al., 2006) found that the optimal dataset for completely modelling the six basic facial expressions should contain $O(10^2)$ persons and each person has $O(10^3)$ images.

An alternative method to exploit the dynamic features in facial expression recognition is through the use of spatio-temporal descriptors such as local binary patterns from three

orthogonal planes (LBP-TOP) (Zhao and Pietikainen, 2007) and local phase quantization from three orthogonal planes (LPQ-TOP) (Jiang et al., 2014, 2011) (without using probabilistic modelling). The way to use these spatio-temporal descriptors is to exploit both the spatial and temporal features by applying them independently to the three orthogonal planes (XY plane: appearance, XT planes: horizontal motion, and YT plane: vertical motion) in a video volume of the whole face region or local face region (see Figure 2.23).

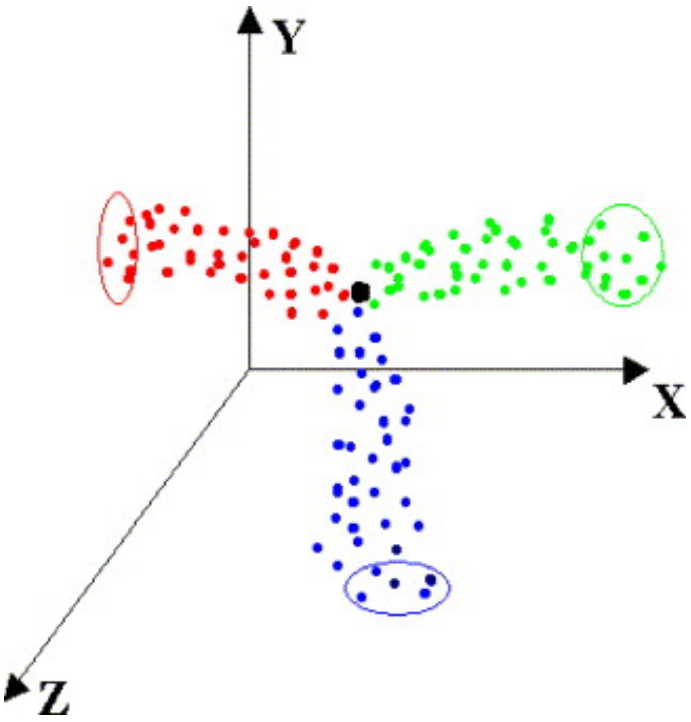


Fig. 2.22 Illustration of the 3D expression manifold. The centre is the neutral frame. The further a point is away from the centre point, the higher is the intensity of 3 expressions. This figure is adapted from Chang et al. (2006)

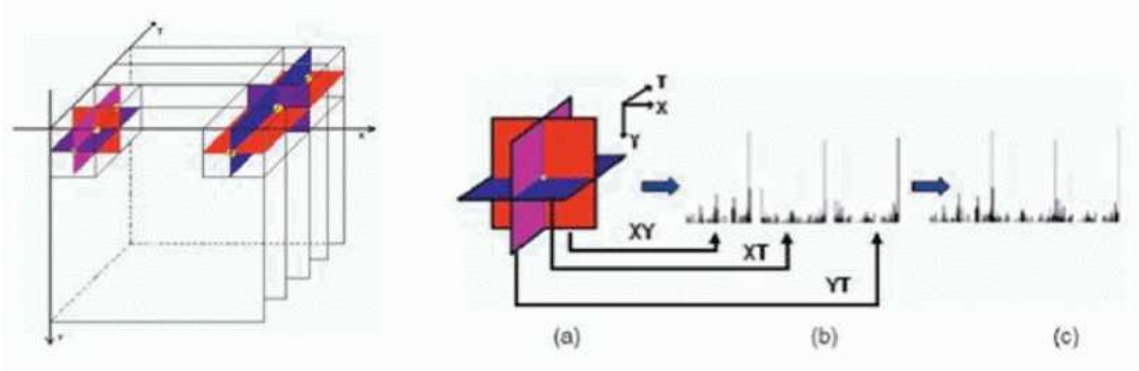


Fig. 2.23 Left: Three planes in spatio-temporal space to extract neighbouring points. Right: Concatenated histogram from three planes. This figure is adapted from Jiang et al. (2011)

2.5 Discussion

This chapter has presented an in-depth description of different aspects of AFER. A review of the techniques used in AFER was then presented followed by techniques to model several challenges in AFER. According to the literature review, several areas can be highlighted requiring more investigations and this thesis has contributed to filling some of these gaps.

Firstly, in Section 2.2.3.1, a number of texture analysis methods used for the purpose of facial expression representation have been reviewed. It can be noticed that these representations have become very popular in recent years and highly useful for the task of facial expression recognition, due to not only their robustness to generic image processing challenges such as illumination and registration error but also to the implementation simplicity. Although the robustness of these representations to discrimination among several action units and basic expressions within a limited range of ages and spatial intensity based on the datasets used for evaluation has been demonstrated, the generalisation capability of these representations under a wide range of variations in facial activities (face deformation) such the variations related to the problems under study are not obvious. These deformations might reduce the representation ability and subsequently lead to poor performance of the AFER system. In other words, a major problem with the texture representation is that they stem from a limited amount of facial effect activities (3-7) emotion and (4-26) AUs (see Table 2.1). Additional work still needs to be done to investigate if the existing face descriptors are sufficient as they are, some modifications are required, or a novel set of algorithms is required to represent these facial activities. A further improvement in the aspect of texture-based AFER system could play a large role to achieve the goal. BRIEF features, as we propose to use in this thesis, might contribute to this gap of the literature.

Secondly, in 2.2.3.2, many facial features points localisation methods have been reviewed. It can be noticed that these representations have become very popular in recent years and highly useful for the task of facial expression localisation and feature extraction due to their robustness to the variations in illumination since the intensity of the pixels is ignored. Very recently, Wang et al. (2018) proved that although the results of some state-of-the-art methods of FFPD are comparable to the human in some datasets, their success in some real-world scenarios is not clear; scenarios such as the intrinsic variations from people themselves such as those due to age and expression, and extrinsic variations from the environment such as pose, and distance to the camera (image resolution). As a result, a further investigation into this aspect of automatic facial expression localisation could play a significant role to achieve the target. RFRV-CLM, as proposed to be used in this thesis, will examine to contribute to this gap of the literature.

Thirdly, Section 2.4.1 reviewed the studies that modelled the age and expression together for facial expressions recognition. Although these studies have addressed this issue successfully using the texture features only such as Gabor filter as in Guo et al. (2013) and LBP as in Lou et al. (2018), geomtric feature only as in Wang et al. (2016), or deep features as in Yang et al. (2018) the problem is not studied in depth, and more investigations are required to explore the possibilities that can be used to model the age and the expression. For example, to investigate the effect of the human ageing on the face shape features or analyse the benefit from using apparent age on the accuracy of AFER system. A comprehensive study is proposed in this thesis to study the problem deeply using all the face information including both the geometric and appearance features.

Moreover, although a very limited number of studies have addressed the compound emotions problem as described in Section 2.4.2, this issue is still open research and more investigations are required to investigate the possible features or methods that can be used for compound emotion representation and classification to make the solution to automatically recognize expression more general and effective.

Finally, the expression's intensity is also still open research and more investigations are required to make the solution more reliable for real world applications. This is because the intensity play an important role in the meaning of human expression.

Chapter 3

Data and an Overview of the Proposed Study

3.1 Introduction

The research problems and the literature review which were described in Chapter 1 and Chapter 2 respectively have demonstrated that intrinsic variations from people themselves such as ageing, compound expressions, and expressions's intensity have introduced many difficulties to the task of facial expression classification. Therefore, the aim of this thesis as described in Section 1.5 is to construct an AFER system that can recognize the facial expressions under a wide range of face deformations, namely a wide range of age, expressions, and expression's intensities.

Building such a system requires a comprehensive analysis and study to understand the inner details and associated problems in order to explore the possibilities that can be used for successful AFER. These investigations require a comprehensive facial expressions dataset that contains a broad range of variations of face appearances. These data need to be processed and prepared: the face object in each image in the dataset needs to be detected and annotated. The detected/annotated face need to be processed (modelled) in order to obtain a suitable representation (face parameters) for use with machine learning methods to distinguish among expression categories. In order to recognize the facial expression from the face image, a way of representing/modelling the face as a compact set of parameters is needed. These parameters can be the shape parameters, texture parameters, or the combination of the shape and the texture. Finding these parameters can be done by using texture-based, shape-based, or appearance-based models.

With reference to the facial expressions and the research problems of this thesis, both the texture and shape face features vary widely across the images and they are correlated. These variations in texture and/or shape may be related to the difference in one or more variables such as age, gender, identity, expression type, and the intensity of the expression.

For a comprehensive understanding, investigation and analysis about these differences and the problems under study, both the shape and the texture features need to be modelled separately and in combination to obtain the inner details of the face.

Therefore, in this thesis, we divided our study into three parts. The first part uses the texture measurements and machine learning methods to measure the influence of the problems under study on texture features and to explore the possible ways of using textures feature for facial expressions modelling. The second part uses the shape measurements and machine learning methods to measure the influence of these problems on shape features and to explore the possible ways of using shape feature for facial expressions modelling. Based on the observations that are obtained from parts one and two of our study, the third part focus on eliminating the effect of age on the accuracy of facial expression recognition In the following, the data and the methodologies or hypotheses used throughout the project are briefly described.

3.2 Datasets

There are many high quality facial expression datasets available to the public McDuff et al. (2013). However, since our target is to understand and analyse the effect of some problems including human ageing, compound emotions and expression's intensity on the performance of facial expression recognition, the dataset needs to be representative and reflect the conditions of those problems. Therefore, a specific facial expression dataset with specific characteristics is used with each problem. These datasets are the age and expression datasets, compound emotions dataset and expression's intensity dataset. The details of each dataset is described below. Although each dataset of each problem is quite representative, the data is very small for instance, there are 114 images per expression per age group in FACES dataset and 100 images per expression in compound emotion dataset. That why we restricted ourselves in this thesis to use handcrafted features, and it prevented us from using Deep learning since the data is not enough to obtain a convincing conclusion and since there is a concern that deep learning need a huge amount of data to train, validate, and test and deep learning is data-based learning. Therefore, deep learning is beyond the scope of this thesis.

3.2.1 Age and Expression Datasets

For analysing and modelling human ageing, we used three datasets: FACES (Ebner et al., 2010) , Lifespan (Minear and Park, 2004) and NEMO Dibeklioğlu et al. (2012) which are designed for research into both age estimation and facial expression classification.

The FACES dataset contains 171 people showing six expressions (anger, disgust, fear, happy, neutral and sad) with uniform distribution of the expression classes. The ages of the

subjects range from 19 to 80. In total there are 37 different ages unevenly distributed. The subjects are divided into three groups according to their age: young: 19-31 years old (M = 24.3 years, SD = 3.5), middle-aged: 39-55 (M = 49.0 years, SD = 3.9), and older: 69-80 (M = 73.2 years, SD = 2.8). The faces in this dataset are frontal with fixed illumination mounted in front and above the faces. Two examples of each expression were recorded for each individual. In total the dataset consists of 2,052 frontal face images (Ebner et al., 2010).

The Lifespan dataset contains face images from people of different ethnicities showing eight different expressions with different subset sizes; neutral (N=580), happiness (N=258), surprise (N=78), sadness (N=64), annoyed (N=40), angry (N=10), grumpy (N=9), and disgust (N=7). The ages of the subjects range from 18 to 93 years and in total there are 74 different ages Minear and Park (2004).

The NEMO dataset consists of 400 male and female subjects and 1240 videos. The age of the subjects ranges from 8 to 76 years. Each subject recorded several videos of their expression changing from neutral to happy, spontaneously and deliberately (Dibeklioğlu et al., 2012). See Figure 1.2 for an example from the age and expression dataset.

3.2.2 Compound Emotions Dataset

For compound emotions experiments, we use the compound expression dataset that was developed in Ohio State University which has 6 basic and 15 non-basic expressions plus the neutral expression with ground-truth labels. This data has been collected to study the effect when several emotions are exhibited at once. This dataset was collected from 230 (male and female) subjects showing 22 distinct expressions. Most ethnicities and races were included: Caucasian, Asian, African American, and Hispanic individuals are represented in the dataset (Du et al., 2014). See Figure 1.3 for an example from the compound emotion dataset.

3.2.3 Expression's Intensity Dataset

We use the Cohn–Kanade (CK+) dataset for analysing the intensity of the expression. The CK+ dataset is one of the most comprehensive datasets in the current facial expression research community. The dataset consists of 123 university students aged from 18 to 30 years, 65% of whom were female, 15% were African-American and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were based on a description of prototypic emotions. Image sequences consist of the first frame (containing a neutral expression) and the last frame (containing the expression at peak intensity). The target displays were digitized into 640 by 490 pixel arrays with 8-bit precision for gray scale values (Lucey et al., 2010). See Figure 1.4 for an example from the Cohn–Kanade dataset. Table 3.1 summarizes the datasets and their characteristics.

Table 3.1 Description of the datasets use throughout the project.

Dataset	Age Range	Examples	Expressions	Persons	Modality	Landmarks
FACES	19 - 80	2052	6	171	Image	76
Lifespan	18 - 93	1006	8	230	Image	-
NEMO	8 - 76	1240	2	400	Video	-
Compound emotions	av:23	2200	22	230	Image	78
CK+	18 - 30	327	7	118	Video	70

3.3 Data Setting

Once the data is collected, the face in each image needs to be detected and annotated.

3.3.1 Face Detection

Usually, in any face analysis application before any processing algorithm is applied to the whole face image, the location of the face in that image should be found by using a face detection algorithm. In this thesis, the face detector introduced by Viola and Jones (2004) is selected since it is probably the best known and widely used nowadays for face detection.

3.3.2 Manual Annotation

Manual annotation is the process of labelling and identifying the key points of an object in the image. The choice of these points is an essential step of automatic shape modelling since this relies on these points and their consistency across the samples of the dataset. There are several important factors that should be considered in the process of image annotation, including the number of points, the positions of the points, and the relationship between the set of the points. The number of the points is varied depending on the accuracy requirements. For example, for a high-level task such as facial expression understanding, a large number of points are required, from 60 to 80. The position of these points should be chosen carefully. The aim is to identify landmarks that can be placed repeatedly and reliably across many images, and that will represent the variation in the face due to changes of expression.

In the case of facial expression recognition, positioning of the points is more important around the mouth and eyes areas since these areas are subject to the maximum variations for different expressions. The variations with facial expressions' annotations are not only due to variabilities of framing of the person in the image, such as position, scaling, and rotation, but also due to the expressions' variabilities (anger, disgust, fear, happy, sad, and surprise). In addition to these variabilities, there are some factors that might hinder the process of manual annotation, such the presence of age-related patterns and their overlap

with the expression pattern, spontaneous (genuine) or posed (faked) expressions, showing two expressions (compound expressions) simultaneously, and showing the expression in a different way depending on the person's personality such as showing an expression in different intensities. These hindrances will be discussed in the Chapter 5 in more detail.

In this thesis the 2052 images from the FACES dataset were manually annotated with 76 landmark points and we automatically detected the points of the Lifespan and NEMO datasets using our model trained on the FACES dataset described in Chapter 5. Each frame of each sequence of the CK+ data set was annotated with 68 landmark points. In addition to these 68 points, and for the requirements of our prosed FEL, We manually added two points in the centre of the eyes to make 70-points in total. For the compound emotion dataset, 78 points were recorded by the author. Figure 3.1 shows an examples of manual annotations for three datasets.

Fig. 3.1 Landmark annotation of FACES dataset (top left), compound emotions dataset (top right), and CK+ dataset(bottom).

3.4 Facial Expression Modelling and Problem Setting

This section covers the handcrafted face representation methods used in this thesis for face features modelling. DL is not used in this thesis since the data is very small in number and is not enough to obtain a convincing conclusion. For instance, there are 114 images per expression per age group in the FACES dataset and 100 images per expression in the compound emotion dataset.

3.4.1 Face Texture Modelling

Once the data is prepared, we start our study by using texture measurement methods in which the grey/colour information of the face image is used for finding a suitable face representation which can be used for further processing, such as to make measurements or to classify patterns. The problem with the texture features of facial images, with the presence of an extensive amount of skin texture deformations related to the expressions, age and intensity patterns, is that the face texture has no explicit/global texture variations to get useful description directly. For that reason, the texture needs to be modelled locally using local description techniques, in which each pixel is evaluated locally with the neighbouring pixels in order to capture the micropatterns of the face. This can be done by using histogram representations based on local descriptors as reviewed in Section 2.2.3.1 and it is proposed to use them in the thesis since they have the potential to satisfy our requirements.

In this thesis, three methods are selected for texture measurements and for representing the face texture features. These methods are: Binary Robust Independent Elementary Features (BRIEF) (Calonder et al., 2012, 2010), local binary pattern (LBP) (Ojala et al., 2002), and Quantized Local Zernike Moments (QLZM) (Sariyanidi et al., 2013). The BRIEF method is used to introduce a new face descriptor using the BRIEF feature, whereas the LBP and QLZM methods are used for both comparison and validation purposes of the BRIEF-based face descriptor. The LBP method is chosen because LBP is a widely used technique in face applications and in particular in facial expression recognition and it is very similar to the BRIEF technique. The QLZM method is chosen because it is a widely used technique in computer vision applications such as character recognition (Khotanzad and Hong, 1990), face recognition (Ono, 2003; Singh et al., 2011) and facial expression recognition (Sariyanidi et al., 2013). In (Sariyanidi et al., 2013), the author found out that describing the face images using QLZM outperforms the LBP (Ahonen et al., 2006), Gabor (Littlewort et al., 2011), and BOW (Sikka et al., 2012) in a facial expression classification task. In addition, the performance of QLZM in an AFER task with 96.1% accuracy (Sariyanidi et al., 2013) outperformed the performance of HOG with 70% accuracy (Sariyanidi et al., 2015). The superior performance of the QLZM method encourages us to choose it in our comparative study.

Chapter 4 presents comprehensive details about the texture measurements performed in this thesis regarding the problems under study, along with the methods used, motivations, experiments, and findings.

3.4.2 Face Shape Modelling

We extended our study by modelling face shape features to analyse, understand and explore the impact of the problems under study on face shape features and hence on the performance of face components localisation and shape-based AFER. Face shape modelling is the process of using the geometric information to find a suitable face representation which can be used for further processing, such as to make measurements, segment the object or to classify the object. Similar to the texture feature, the problem with the shape features of facial images, with the presence of an extensive amount of rigid and non-rigid face deformations related to the expressions, age and intensity patterns, is that the face shape has no explicit/global shape variations to capture the precise facial feature points' poses. For that reason, the shape needs be to modelled locally using local discriminative techniques, in which each shape point is modelled locally and independently within the patch around it in order to capture the best position of each point in the face image. This can be done by using FFPD methods reviewed in Section 2.2.3.2 to localise the shape features of the input face image.

In this thesis, we restrict ourselves to random forest regression voting constrained local models (RFRV-CLM) Cootes et al. (2012); Lindner et al. (2015, 2013) to develop an automated facial expression localization FEL system. RFRV-CLM is chosen since has the potential to satisfy our requirements. RFRV-CLM attempts to model the points' poses directly by learning a regressor from an image patch feature around the points to predict the target pose from the nearby regions of the face image. It also achieved remarkable performance in many computer vision application. The RFRV-CLM requires hand-annotated images for training the model and can handle variations in points' poses. A sample of images were analysed to understand the variation in shape and orientation of the face. Using this information a set of points was placed to describe the key shape characteristics of the facial expression. Therefore, each image is annotated with n landmark points as described in Section 3.3.2.

Chapter 5 presents comprehensive details about the shape measurements performed in this thesis regarding the problems under study, along with the methods used, experiments, and findings.

3.5 Expression Classification

In this thesis the SVM and RF classifiers are selected for expression classification throughout the project. The SVM classifier is chosen for the comparison and validation of the BRIEF-based face descriptor since most of the previous face descriptor schemes which were applied to the task of facial expression classification were evaluated using the SVM classifier. With regard to the parameter selection of SVM, as suggested by Hsu et al. (2003), we carried out grid-search on the hyper-parameters using 10-fold cross-validation. The parameter setting producing the best cross-validation accuracy was picked.

In the rest of the thesis, the RF classier is used since it outperforms the SVM in many computer vision tasks as reviewed in Section 2.2.4. See (Criminisi et al., 2012) for more complete details about the performance of RF in several applications and its comparison with SVM and boosted methods. In this thesis the random forest classifier is selected for classifying the data into discrete classes. RF is an ensemble of random decision trees as we described in Section 2.2.4. Each tree uses a subset of features to classify the data. In each node the best split to classify the data is found by calculating the information gain I for each feature which is calculated using the entropy of the features $H(X)$. The equations below illustrate a simplified measure of the information gain I and the entropy $H(X)$:

$$I_j = H(X_j) - \sum_{i \in \{L,R\}} \frac{|X_j^i|}{|X_j|} H(X_j^i) \tag{3.1}$$

where j is the node, X is the features, and i is the leaf.

$$H(X) = -\sum_{c \in C} p(c) \log p(c) \tag{3.2}$$

where $p(c)$ is the probability of data in each class c. During training, information that is useful for predication during testing will be learned for all the leaf nodes and stored. In the forest, all the trees are learned independently. During testing, each test point v is pushed starting from the root through all trees until it reaches the corresponding leaves. All the trees will make predictions and then will be combined into a single forest predication and the average is calculated by

$$p(c|v) = \frac{1}{T} \sum_{t=1}^{T} p_t(c|v) \tag{3.3}$$

where T is the number of trees in the forest. All forest experiments presented in this thesis were run with $T = 100$.

3.6 Discussion

In this chapter we have discussed the specific details about how we represent facial expressions. Firstly, we have described the five different facial expression data sets that are used in our study to evaluate our models. Secondly, we have covered the face image features that we used to extract the face image information relevant to facial expression. Finally, the classifier that was used for expression classification is described. For easy understanding, we describe the methods of each chapter in the chapter itself. Figure 3.2 illustrates an overview of the proposed methods and problem settings of our study.

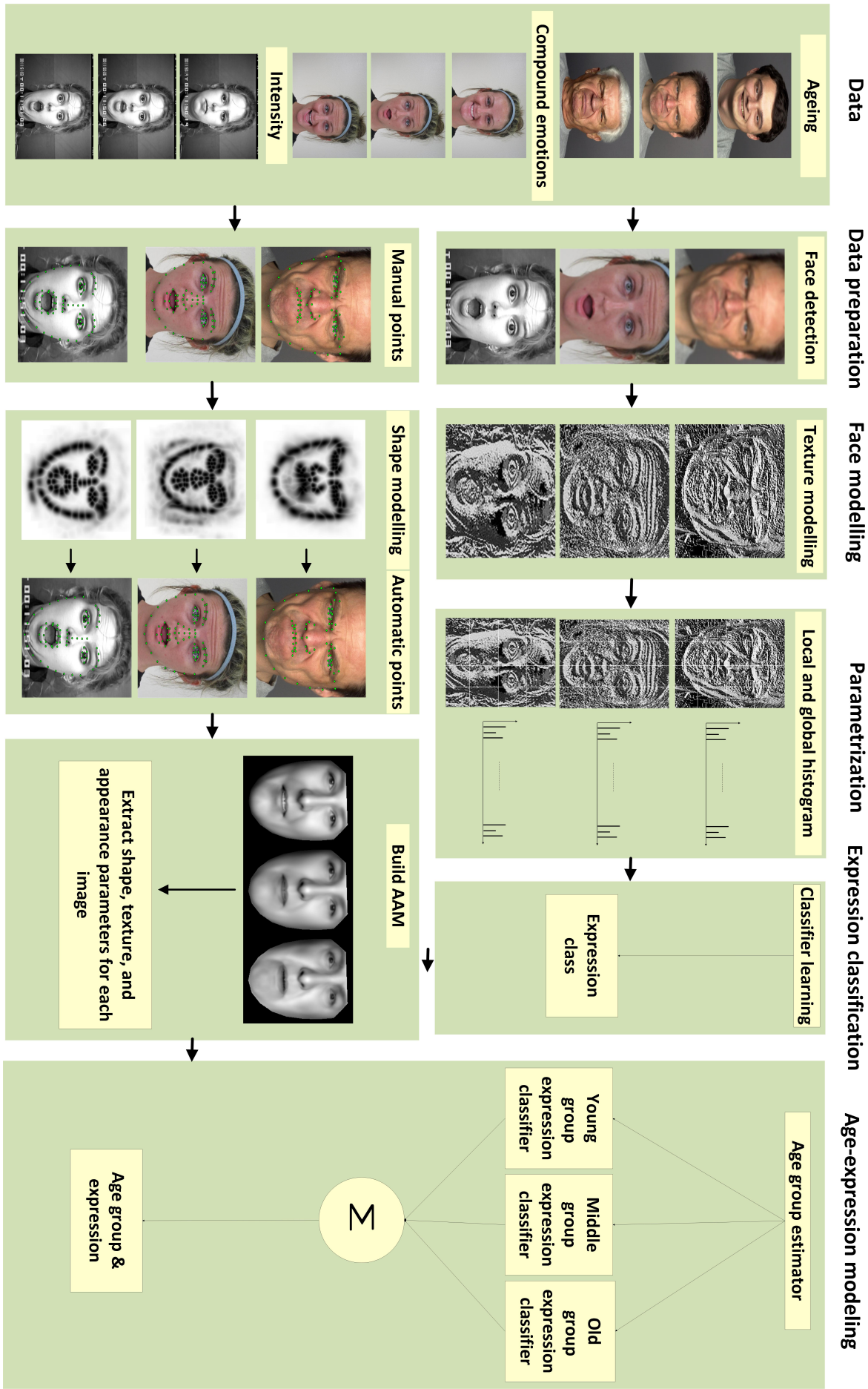


Fig. 3.2 Overview of the proposed methodologies in this thesis.

Chapter 4

Development and Comprehensive Evaluation of BRIEF-Based AFER

4.1 Introduction

As stated in Section 2.2.3.1, a robust face descriptor is essential for face image analysis in general and AFER in particular, where the descriptor is used as a tool to extract the texture features of an image to be used for further measurements such as segmentation, classification, etc. This chapter introduces the BRIEF features as a new face descriptor. BRIEF features are used in a similar way to the proposal of the well-known face descriptor approach useing LBP developed by Ahonen et al. (2006) to account for the discriminative information in various regions of the face such as around the mouth and eyes since these areas contribute most to different variations of expressions. *The first aim* here is to examine the development of a BRIEF-based AFER system, along with an analysis of which combination of BRIEF's free parameters is suitable for describing the face image. *The second aim* is to present a comparative study between the BRIEF-based face representation, and some alternative face representation methods including LBP-based and QLZM-based. This comparative study is used to first evaluate the ability and sensitivity of a texture-based AFER system to accurately discriminate among facial expressions in the presence of a wide range of age and emotions patterns, and secondly to measure the influence of those patterns on the texture features and thus their impact on the performance of texture-based AFER. For easy understanding, this chapter starts by giving our motivations for selecting the BRIEF method for facial expression representation in 4.2 and then gives an explanation of the BRIEF algorithm in 4.3.2. In Section 4.3.2, we introduce our BRIEF-based AFER system. Section 4.4 presents results of the proposed system of BRIEF-based face descriptor in AFER tasks under the effect of the problems under study with a comparison made against the results of alternative methods evaluated on the same datasets. Finally, Section 4.5 concludes the chapter by providing a summary of the key findings.

4.2 Motivations

This section presents the motivations for selecting BRIEF features for face texture description. BRIEF has proven to be highly discriminative, and it is very efficient both to compute and to store, making it suitable for demanding image analysis tasks. It belongs to a broad class of local binary descriptors. The most similar descriptor to BRIEF is LBP, the main difference between the two is that instead of comparing the pixels on the circumference of a circle against the one at the centre, BRIEF selects a subsample of pixel pairs at Gaussian weighted random locations from the described area. This property gives it the ability to visit the pixels inside the circle and to use a window with a larger diameter than LBP uses with less loss of information than LBP (see Figure 4.1(a) and (b)). The other difference lies in the ability of BRIEF to have more than one pixel contributing to a single descriptor bit. Therefore, BRIEF's descriptor length is half of the LBP's descriptor, despite being generated by the same number of pixels. For instance, 17 pixels (the centre pixel and the 16 neighbours) based on LBP contribute to a 16-bit descriptor, while 16 pixels based on BRIEF contribute to an 8-bit descriptor (see Figure 4.1 (c) and (d)). Using a large window and short descriptor can help to represent the texture locally and globally with a short feature vector.

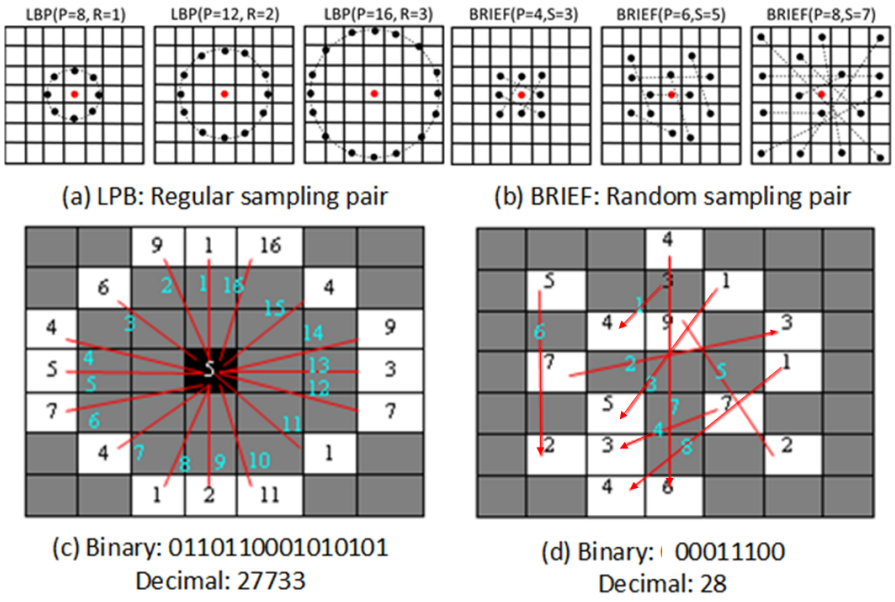


Fig. 4.1 Comparison between the LBP (left) and BRIEF (right) descriptors: (a) and (b) describe the regular and random sampling pattern used by LBP and BRIEF respectively; P, R, and S refers to the number of sample pairs, the radius of the circle and the size of the window respectively; (c) and (d) show examples on 7×7 window of how the different ways in which LBP and BRIEF are computed result in different descriptor lengths. (White cells, red lines and green number on the red lines refer to the intensity of the pixel, the pairs to compare and the sequence of pair wise comparisons to generate the binary value respectively.

Furthermore, Chai et al. (2013) showed that LBP's code is very sensitive to noise. Conversely, we found out that BRIEF is less sensitive to noise. Figures 4.2 and 4.3 illustrate the sensitivity of LBP and BRIEF descriptors to noise respectively.

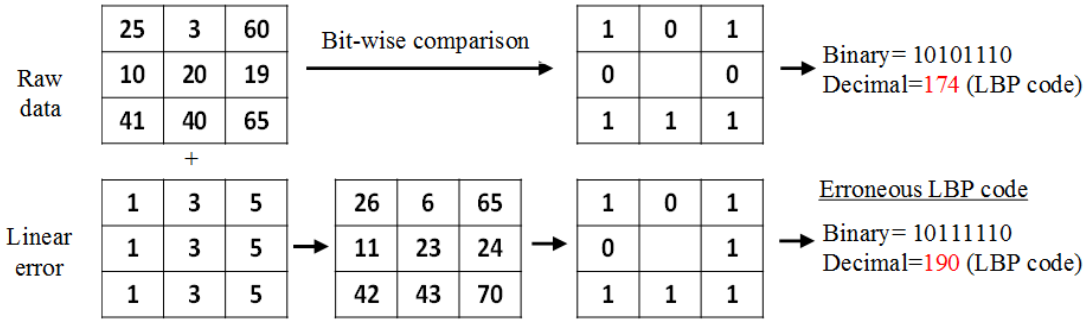


Fig. 4.2 Illustration of the sensitivity of Local Binary Pattern (LBP) to noise. This figure is redrawn from (Chai et al., 2013).

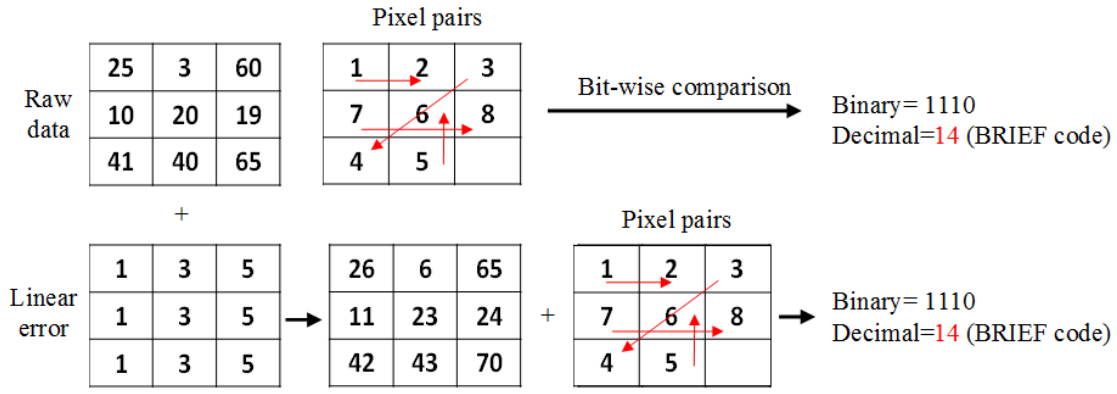


Fig. 4.3 Illustration of the sensitivity of Binary Robust Independent Elementary features (BRIEF) to noise.

Moreover, Heinly et al. (2012) performed comparative experiments among Binary Robust Independent Elementary Features (BRIEF), Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011), and Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011). The experiments were performed to evaluate a comparable performance on pairs of images under various transformations including image blur, exposure, JPEG compression, combined scale, and rotation, using the Oxford dataset and simple rotation, simple scaling and illumination changes using their own dataset. They found that the BRIEF method outperformed ORB and BRISK methods due to the fixed pattern regarding the scale and the orientation and the accurate representation for the gradient of image patch.

In addition to these characteristics, BRIEF has recently shown remarkable performance in multiple domains including place recognition (Gálvez-López and Tardos, 2012), loop closure detection (Galvez-Lopez and Tardos, 2011), and optic disc segmentation in retinal

images (Mohammad et al., 2013; Mohammad and Morris, 2017). Applicability to multiple domains can be attributed to the simplicity of the BRIEF method, with significant research in feature extraction and a very good compromise between distinctiveness and computation time, as well as its suitability for appearance recognition.

A strength that BRIEF brings to the table is the possibility to get pairs of pixels at random positions based on the Gaussian distribution which proved to be superior to other sampling patterns in terms of recognition rate (Calonder et al., 2012, 2010). Moreover, the ability of BRIEF to map a pair of pixels to a single descriptor bit results in a short feature vector, and it is independent of changes in illumination which help to give a good sampling of the whole area of the patch.

In summary, these findings motivated us to investigate the suitability of using BRIEF features for the problem of facial expression recognition in particular and for describing the face in general. Our target is to use the BRIEF features for the AFER task. Having that in mind, we must take into account also the evaluation of using the BRIEF features on the effects of previously mentioned challenges (human ageing, compound emotions, and intensity), since they are the focus of this thesis.

4.3 Method

This section describes the original BRIEF method and the BRIEF-based face descriptor method which is proposed in this thesis.

4.3.1 Binary Robust Independent Elementary Features (BRIEF)

Binary Robust Independent Elementary Feature (BRIEF) Calonder et al. (2012, 2010) is one of the best performing texture descriptors. BRIEF's basic idea is based on the hypothesis that an image patch can be effectively classified on the basis of a small number of pairwise intensity comparisons. The results of these tests are used to train a classifier to recognise image patches from different viewpoints. BRIEF quantifies the texture of an image patch as a binary string. Bits of the string are computed by comparing the values of pairs of pixels within the patch. The formal definition of BRIEF is given in Equations 4.1 and 4.2. A test τ is defined on patch S of size $s \times s$.

$$\tau(S;x,y) = \begin{cases} 1 & \text{if } I(S,x) < I(S,y), \\ 0 & otherwise, \end{cases} \tag{4.1}$$

Where $I(S,x)$ and $I(S,y)$ are the pixel's intensities in a smoothed version of patch S at locations x and y. Choosing a set of n_d (x,y) location pairs uniquely defines a set of binary tests. The BRIEF descriptor is then defined as the n_d-dimensional bitstring, corresponding to Equation 4.2:

$$f_{n_d}(S) = \sum_{1 \le i \le n_d} 2^{i-1} \tau(S; x_i, y_i) \tag{4.2}$$

In the calculation of BRIEF, two factors need to be considered. First is the smoothing kernel used to smooth the image patch and second is the spatial arrangement of the pixel pairs. Smoothing is introduced to suppress noise, thus increasing the stability and repeatability of the descriptor. For the spatial arrangement of the pixel pairs, there are five methods proposed by Calonder et al. (2010) for selecting the points pairs as shown in Figure 4.4. One of these methods (G II in Figure 4.4) is sampling the points from the patch S randomly based on an isotropic Gaussian distribution $(0, S^2/25)$ which achieves the best results in terms of matching results and recognition rate. For this reason, in all further experiments presented in this chapter, it is the one we will use.

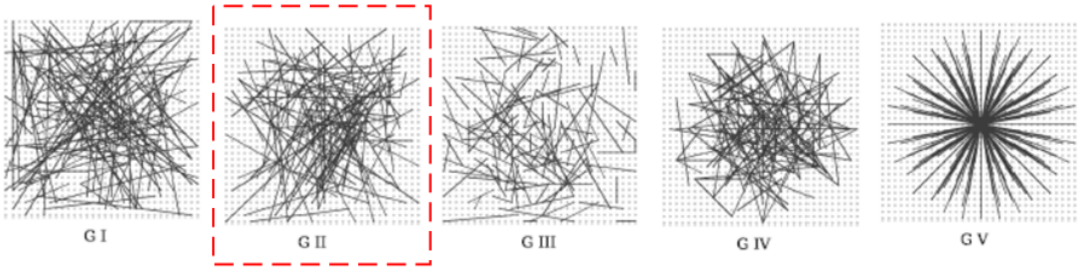


Fig. 4.4 Five different approaches to sampling patterns. This figure is adopted from (Calonder et al., 2010).

BRIEF was originally used to match points in images with different viewpoints. In this thesis, we propose BRIEF as a texture descriptor method used to measure and describe face texture in general and facial expression in particular.

4.3.2 BRIEF-Based Face Descriptor

BRIEF features are widely used for texture description and classification, in which the occurrence of local texture values (BRIEF code) is used to generate one histogram to represent an image. This histogram is then used for classification among several classes either by computing the histogram's similarities or by training a classifier to use on the test image. Applying the same procedure for facial images results in losing some spatial information, in particular information from locations such as the areas around the eyes and the mouth. The reason for that loss is just because with a holistic approach the texture descriptor aggregates over the whole image area. An alternative approach that retains the information of spatial location is to use BRIEF features derived from distinct regions to build several local descriptors and then concatenate them. Therefore, in the proposed descriptor the BRIEF features are used locally folowing the proposal of Ahonen et al.

(2006) to account for the discriminative information in spatial locations of the face image. The processes of the proposed face descriptor are:

- Each pixel in the face image is evaluated locally using Equation 4.2 with a defined number of sample pairs P and overlapping patch size S.
- The generated BRIEF image is then divided into grid of non-overlapping regions W.
- A histogram of each region is then calculated. A histogram H of the labelled region $I(x,y)$ can be defined as the following.

$$H_i = \sum_{x,y} I(f(x,y) = i), i = 0,...,n-1 \tag{4.3}$$

in which n is the number of different labels produced by the BRIEF operator using Equation 4.2, $I(x,y)$ is the label value. This histogram contains information about the distribution of the local micropatterns over the face image.

- For efficient representation, the histograms of all patches $W_0,...W_{m-1}$ (W is the size of the patch and m is the number of patches) are then concatenated in a one histogram to form a global representation of the face image as in Equation 4.4.

$$H_{i,j} = \sum_{x,y} I\{f_i(x,y) = i\} I\{(x,y) \in W_j\}, i = 0,...,n-1, j = 0,\ldots,m-1 \tag{4.4}$$

The final histogram represents a combination of three levels of local descriptions: pixel description, patch description (local histogram), and an image description (global histograms). This representation/description will be used to discriminate one texture from another for facial expression recognition. Figure 4.5 illustrates the proposed face descriptor.

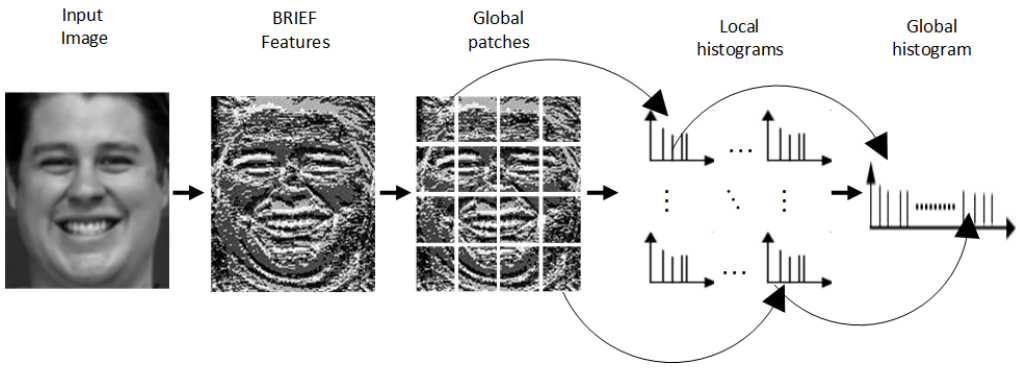


Fig. 4.5 The framework of the proposed face descriptor using BRIEF features.

4.3.3 BRIEF's Free Parameters Description

BRIEF initially used 256-pixel pairs to form a texture descriptor. In the present work and since the idea is to use BRIEF features for building several local face descriptors, small patches and a small number of sample pairs will be used. The optimum values of those variables and their impact on the facial expression recognition task are not obvious. Therefore, a grid search to find the most suitable values of BRIEF's free parameters to describe the face image, in particular, is required. In the following, the proposed face descriptor's free parameters will be explained in more detail followed by their optimal values in Section 4.4.1.1.

Smoothing kernel size: The first parameter is the kernel which is used to smooth the image patch. Smoothing is introduced to suppress noise, thus increasing the stability and repeatability of the descriptor. Smoothing using a 9×9 Gaussian kernel with $\sigma = [1-3]$ is recommended by Calonder et al. (2010).

Patch Size: The second parameter is the division of the image into regions including local overlapping regions S, which are used to evaluate each pixel based on BRIEF, and global non$-$overlapping regions W which are used to build the final histogram. A large number of small regions will produce a very local descriptor with high time consumption, whereas using large local regions causes more spatial information to be lost. Larger sized grids of global regions will result in fewer samples being derived from a given sized face image and a sparser histogram, and therefore worse recognition. We choose to divide the image with a grid of $W_0, ...W_{m-1}$ equally sized square windows. The values of S and W are the subject of the optimization experiment in section 4.4.1.1.

Spatial Arrangement of the (x,y) pairs: The third important parameter of the BRIEF descriptor is the arrangement of the sampling pairs as explained in Figure 4.4. We randomly select the pixel pairs (using an isotropic Gaussian distribution with parameters $(0, S^2/25)$) from within the patch. The pairs locations are predefined during initialization, and then the same locations are used for analysis of all texture classes.

Descriptor length: The fourth parameter of BRIEF is the descriptor length. Choosing an arrangement of pixel pairs from the patch S that produces a large number of different labels makes the descriptor longer and the learning operation slower. Using a small number of labels makes the feature vector shorter but also means losing more information. For instance, when finding the value of a pixel from 8 pairs (16 pixels in total), the descriptor is 8-bits and its range is $[0, 2^{8-1}]$, for 16 pairs (32 pixels in total and 16-bits) and its value range is $[0, 2^{16-1}]$ and so on. In the original work, 256 pairs were used to form a BRIEF descriptor. Since our plan is to use the descriptor in a histogram framework, the use of a 256 bit descriptor will result in a very large and sparsely populated histogram. Therefore, in our implementation a much shorter BRIEF descriptor is used. And as will be shown later, a smaller number of pixel pairs is sufficient and able to demonstrate good performance.

Experiments reported in section 4.4.1.1 will show how these parameters (descriptor length, local patch size to find BRIEF features and global patch size to build the histogram) can be empirically tuned and optimized. We denote a parametrized BRIEF as $BRIEF(P,S,W)$ having P random pairs, a local square patch S and a global square patch W. To summarize, the proposed face description is performed in two steps as shown in Algorithm 1.

Algorithm 1 . BRIEF-Based Face Descriptor.

Input : A face image I of width w and height h, BRIEF-patch-size S, and pairs P, histogram-patch-size W.
Output : A histogram H.
Initialization : $P = 0$, a BRIEF image $\hat{I} = 0$, $H = 0$.

Step 1:- Calculate the BRIEF image $\hat{I}$ for the input face image I
for $i \rightarrow 0$ **to** w **do**
  for $j \rightarrow 0$ **to** h **do**
    1: From the input face image, crop a patch of size $S \times S$ with length equal to $((S-1)/2))$ around the current pixel(i,j).
    2: if *P is empty* **then**
    Generate P pixel pairs from within the current cropped patch $S \times S$ using an isotropic Gaussian distribution with $(0, S^2/25)$.
    else got to 3:
    3: Compute the pixels's BRIEF value from Equation 4.2 within the current patch $S \times S$ and pairs P.
    4: Transfer the result of BRIEF value for each pixel from the face image I to BRIEF image $\hat{I}$.
  end
end
Step 2:- Calculate the histogram H
for $i \rightarrow 0$ **to** w **do**
  for $j \rightarrow 0$ **to** h **do**
    1: Divide the result BRIEF image $\hat{I}$ into cells of size $W \times W$.
    2: Compute the histogram of the BRIEF values occurring over each cell using Equation 4.3
    3: Concatenate histogram of all cells using Equation 4.4., giving the feature vector of histogram H for the face image
  end
end

4.4 Experimental Evaluation

This section describes experiments performed for evaluating the performance of the proposed face descriptor using BRIEF features. The performance was evaluated in the

AFER system described in Figure 4.6 which consists of (i) face detection using the Viola and Jones detector (Viola and Jones, 2004), (ii) BRIEF feature representation and extraction using the method described in Algorithm 1, and (iii) expression recognition using kernel SVM (Chang and Lin, 2011; Vapnik and Vapnik, 1998). For fair comparison with the previous facial representations methods, faces were detected automatically using the Viola and Jones detector, and were downsampled to 150 * 150 pixels based on the location of the eyes and SVM classifier was used for expression classification.

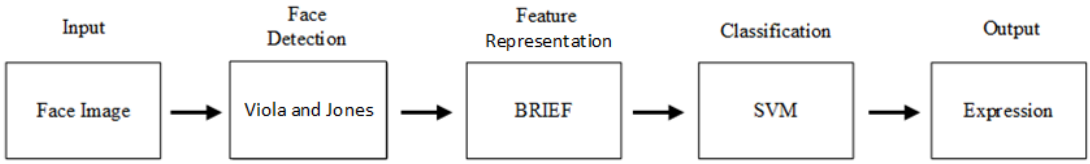


Fig. 4.6 BRIEF-based AFER system diagram

Datasets: The proposed descriptor is evaluated using three different facial expression datasets, including the FACES dataset for ageing and expression described in 3.2.1, the compound emotions dataset for compound expressions described in 3.2.2, and the CK+ dataset for expression and its intensity described in 3.2.3.

Evaluation Metric: For expression classification, both the average of all expressions classes with the standard error and per-class (confusion matrix) classification accuracy between the ground truth label and the predicted label are reported. To avoid over-fitting and identity bias issues, 10-fold cross validation (person independent) experiments are applied. For validation purpose of the texture-based AFER, simple and complex cases of facial expression recognition are considered. A comparison of BRIEF-based performance with the performance of some other methods is also reported.

4.4.1 Evaluation in Simple Cases

In this thesis, the simple case refers to the model that is trained and tested using 6-basic expressions and a limited range of ages of static images.

4.4.1.1 Experiment 1 - BRIEF's Free Parameters Optimization

Case Description: The objective of this experiment is to find the optimal values of BRIEF's free parameters that give the best accuracy in the task of facial expression recognition. Three essential design parameters which must be considered when extracting the BRIEF features are the number of sampling pairs P and the patch sizes (local patch size S and global patch size W). We would expect these parameters to directly affect the size of the feature vector and the runtime of the model. Their effect on the accuracy of the model is less obvious. In order to maximise the throughput of our classifier (system), it is

necessary to minimise the size of the feature vector, while maintaining acceptable accuracy. The parameters of P are set to a sample pair of $(1,2,...,9)$, S and W are set to sizes of $(5^2,7^2,...,21^2)$ pixel. The rationale behind running the experiment with an exhaustive combination of parameters was to ascertain the degree to which the performance of the algorithm depends on the chosen parameters. The optimal parameters are then used to extract the expression features in future experiments.

Using the peak frames of 327 sequences of six facial expressions from CK+ dataset, similar to the other approaches in the literature, a sensitivity analysis of BRIEF-based AFER was carried out in order to estimate the optimum values of the BRIEF free parameters including: pixel pairs P, the local window size S and global window size W, along with their sensitivity to the face representation and hence their impact on expression recognition performance.

Results: Figure 4.7 shows the mean recognition rates of AFER using BRIEF features as a function of the windows size S and number of sample pairs P. These results demonstrate that a small P value such as 5 or 6 is sufficient to obtain good accuracy and a very short feature vector. We found that building the histogram from W equal to S gives better results. Overall, the sensitivity analysis shows that BRIEF representations are not very sensitive to parameter changes, and a small P value can be chosen to keep the dimensionality relatively low. In this thesis the BRIEF(6,11,11) is chosen since it achieved the best performance.

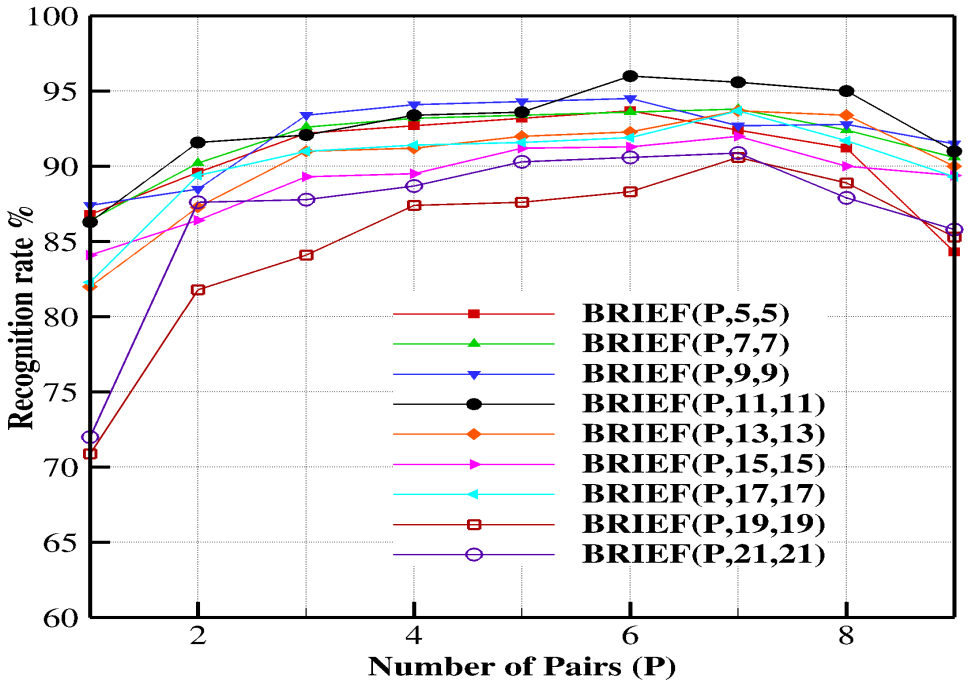
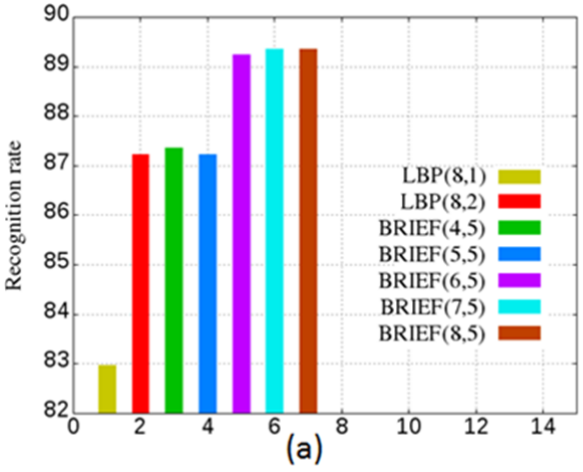


Fig. 4.7 The mean recognition rate for the BRIEF-based AFER as a function of local window size S, global window W, and sample pairs P ($BRIEF(P,S,W)$)

4.4.1.2 Experiment 2 - Comparison to LBP Face Representation

Case Description: The objective of this experiment is to validate and compare the performance of BRIEF features to the LBP features in the task of facial expression recognition. For both methods, we trained and tested the system using images of 118 subjects from the CK+ dataset. Both methods are evaluated using similar configurations regarding the descriptor size and matching windows.

Results: Figure 4.8 illustrates the performance of BRIEF versus LBP. Going from the worst to the best performance of the BRIEF and LBP descriptors, it can be seen that the overall worst performance is provided by LBP(8,1) with 8-bit descriptor length. It is followed by LBP(8,2) which provided better results than the LBP(8,1) due to the larger radius of window matching with the same descriptor length. It is followed by BRIEF(4,5) with a 4-bit descriptor length and BRIEF(5,5) with a 5-bit descriptor length, both of which provided comparable performance to LBP despite their shorter descriptors' length. It is finally followed by BRIEF(6,5), BRIEF(7,5) and BRIEF(8,5), all of which provided better performance than LBP with descriptor-bits length shorter than or the same as LBP (see Figure 4.8 b). The rationale behind this comparison was to demonstrate that the advantage of BRIEF over LBP is that BRIEF can provide comparable or better performance than LBP with a shorter bit string.



Method	Bits	Bins
LBP(8,1)	8-bit	[0,255]
LBP(8,2)	8-bit	[0,255]
BRIEF(4,5)	4-bit	[0,15]
BRIEF(5,5)	5-bit	[0,31]
BRIEF(6,5)	6-bit	[0,63]
BRIEF(7,5)	7-bit	[0,127]
BRIEF(8,5)	8-bit	[0,255]

(b)

Fig. 4.8 The mean recognition rate for the BRIEF-based AFER versus LBP-based AFER as a function of window size S and sample pairs P using CK+ dataset: (a) the performance of AFER and (b) bits and bin numbers required for constructing the descriptor and the histogram respectively.

4.4.1.3 Experiment 3 - Comparison to Other Face Representation

Case Description: The objective of this experiment is to compare the performance of BRIEF-based AFER to the performance of some other face representation methods reported by Sikka et al. (2012) and Sariyanidi et al. (2013) of using Gabor (Littlewort et al., 2011),

LBP (Shan et al., 2009), BOW (Sikka et al., 2012), and H-QLZM (Sariyanidi et al., 2013) on the same datasets with the same configurations of the training and testing (the peak frame of CK+ dataset with leave-one-subject-out evaluation).

Results: The results of this experiment are summarized in Table 4.1. These results show that the performance of the proposed method significantly exceeds that of Gabor and LBP with an improvement on the mean recognition rate of over 4.2% and 13.6 % respectively. We obtained comparable performance to the BOW and QLZM but with a smaller standard deviation from the mean recognition rate, suggesting that the BRIEF representation is more stable.

Table 4.1 Comparison of the BRIEF method with other methods tested on CK+ dataset.

Methods	Classifier	Recognition rate %
LBP (Shan et al., 2009)	SVM polynomial	82.4±2.3
Gabor (Littlewort et al., 2011)	SVM Linear	91.8±2.0
BOW (Sikka et al., 2012)	SVM Linear	95.9±1.4
H-QLZM (Sariyanidi et al., 2013)	SVM RBF	96.1±1.6
BRIEF (present work)	SVM RBF	96.0±0.7

Tables 4.2, 4.3, and 4.4 show the confusion matrices of expression recognition using LBP, QLZM and BRIEF features respectively on the CK+ data set. Results in these tables show that the BRIEF-based AFER system outperforms the LBP-based and QLZM-based systems in most of the emotions.

Table 4.2 Confusion matrix for expression classification results using LBP feature on CK+ dataset.

Data	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	95	2.9	1.1	0	0	1
Disgust	9.7	97.9	11.3	0.1	0	0
Fear	1.9	4	94.1	0	0	0
Happy	0	0	0	100	0	0
Sad	0	3	17	0	80	0
Surprise	0	0.3	3.4	4.3	2	90

Table 4.3 Confusion matrix for expression classification results using QLZM feature on CK+ dataset.

Data	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	100	0	0	0	0	0
Disgust	7.2	91	0	0	0	1.8
Fear	0	3.9	87.5	0	8.6	0
Happy	0	0	0	100	0	0
Sad	0	0	14.3	0	85.7	0
Surprise	0	0	0	0	0	100

Table 4.4 Confusion matrix for expression classification results using BRIEF feature on CK+ dataset.

Data	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	98	1	1	0	0	0
Disgust	2	98	0	0	0	0
Fear	0	0	97.4	0	0	2.6
Happy	0	0	0	100	0	0
Sad	0	0	13.4	0	86.6	0
Surprise	0	0.6	2.2	0	1.2	96

4.4.2 Evaluation in Complex Cases

Having understood the ability of the BRIEF feature with the optimal parameters and its effectiveness concerning the satisfactory recognition rate using small feature vectors in simple cases of facial expression recognition, we move to investigating more complex cases. In this section, the sensitivity and robustness of the BRIEF-based AFER with the optimal parameter are validated in more complex cases of facial expression including human ageing and compound emotions of a rich set of 22-compound emotions patterns (6-basic with 15-non-basic plus the neutral expression). Furthermore, a comparative study among the BRIEF-based, LBP-based, and QLZM-based AFER systems on the problems under study is presented.

4.4.2.1 Experiment 4 - Ageing Effect on Texture-Based AFER

Case Description: The objective of this experiment is to analyse and understand the effect of human ageing on the expression's appearance and hence on the performance of the AFER. The second objective is to assess the robustness of texture face descriptors including BRIEF, LBP, and QLZM and their ability to accurately discriminate among the expressions in the presence of an extensive range of age patterns. The idea here is that the presence of

ageing features which manifest themselves in a similar way to some expression features might make the system confuse the expression categories due to the similar texture patterns which will lead to poorer performance in the AFER task. For example, the fold between the cheek and upper lip can appear in the happy expression of young people and the neutral expression of old people and thus they have a very similar texture code as shown in Figure 4.9.

BRIEF(6,11,11) parameter settings, which resulted in a 96.0% successful recognition rate is selected since it yielded a good trade-off between performance and vector length (see Figure 4.7). We split the subjects from the FACES dataset (described in 3.2.1) into three age groups. We trained three age-group-specific expression classifiers (one for each age group) and one age-agnostic classifier (by combining the training sets of each age) on BRIEF features using the selected parameters. We then tested each classifier on each of the age group test sets.

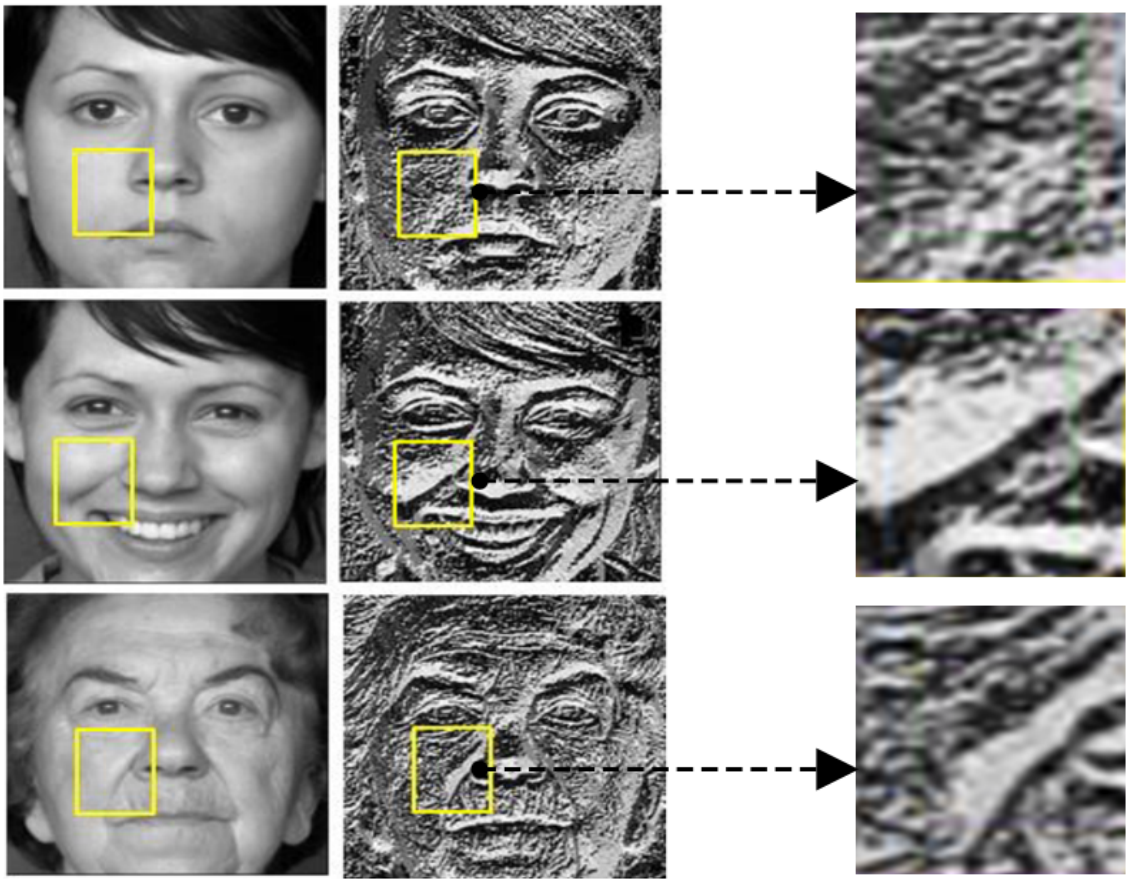


Fig. 4.9 Illustration of the similarity between the expression features of young people with the happy expression (second row) and the ageing features in old people with a neutral expression (third row). Original image (left) and BRIEF response (middle and right)

Results: The performance of the age-specific and age-agnostic classifiers is summarised in Table 4.5. These results show that:

- Performance is best on the age-group for which the system was trained, and degrades as the age difference increases, concluding that ageing has a significant effect on the system performance.
- Performance on the older group is worse than that on the young and middle-aged groups, suggesting that it is harder to extract texture expression features accurately from older faces than younger.
- BRIEF performance using the optimal parameters is stable to some extent with the FACES dataset compared to the CK+ dataset in the previous experiments.

Table 4.5 Expression classification results BRIEF-based AFER for age-specific and age-agnostic models.

	Test				
Train	Groups	Young Group	Middle Group	Old Group	All Groups
	Young Group	99.8	79.2	65.4	62.5
	Middle Group	81.7	96.2	73.3	60.2
	Old Group	64.2	80.0	83.5	60.8
	All Groups	55.05	58.42	53.3	90.1

Table 4.6 summarizes the results of Table 4.5 concerning the age effects in three cases:

- Within age group (the average accuracies over the diagonal three elements from Table 4.5).
- Across age group (the average accuracies over all off-diagonal elements from Table 4.5).
- Mixed age group (the average accuracy when all age groups are mixed without separation).

Going from the worst to the best performance of the classifier, it can be seen that the overall worst performance is provided by across age group with 73.96% since training using one age group of the data cannot provide all the expected variation from other groups. The second best performance is provided by mixed all ages with 90.1%. Finally the best performance is provided by within age group with 93.1% since the training and the testing use the same age group data.

We then compare these results to the previous results published by Guo et al. (2013) and Lou et al. (2018). This comparison shows that:

- The best performance of within age group is obtained by the method developed by Guo et al. (2013) with 97.85% using Gabor features.
- The best performance of across the age groups is obtained by the method developed in this thesis with 73.96% using BRIEF features, suggesting that the feature developed here is less effected by the ageing features.
- The best performance on all the age groups mixed is obtained by the method developed by Lou et al. (2018) with 92.1% using LBP features, suggesting that the approach developed here is less effected by the ageing features.

Table 4.6 Comparison to previous work on FACES dataset.

Reference	Within age group	Across Age Group	Mixed All Groups
Guo et al. (2013)	97.85	64.04	88.80
Lou et al. (2018)	90.05	-	92.1
present work	93.1	73.69	90.1

Table 4.7 illustrates the results of comparative experiments among BRIEF, LBP, and QLZM on the FACES dataset using the same configurations. These results demonstrate that:

- Performance of the three methods is best on the age-group for which the system was trained, and degrades as the age difference increases.
- Performance of the three methods on the older group is worse than that on the young and middle-aged groups.
- The performance of BRIEF is best on the all groups.

Table 4.7 Comparative experiments among BRIEF, LBP, and QLZM on FACES dataset of age-specific and age-agnostic models.

Method	Young Group	Middle Group	Old Group	All Groups
LBP	96.6±0.3	90.0±0.6	77.8 ±1.3	82.1± 0.5
H-QLZM	91.4±0.5	87.5± 0.7	78.4 ± 1.1	86.3± 0.6
BRIEF(present work)	**99.8 ±0.3**	**96.2±0.6**	**83.5±1.7**	**90.1±0.6**

Tables 4.8, 4.9, and 4.10 show the confusion matrices of the emotions recognition system on the FACES data using LBP, QLZM, and BRIEF features respectively.

Table 4.8 Confusion matrix for expression classification results using LBP-based AFER on FACES dataset.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	81	6.6	4.7	0	5.4	2.3
Disgust	0	73	0	0	0	27
Fear	2	0	81	0	17	0
Happy	0	0	0	91	4.5	4.5
Neutral	0	0	2	0	90	8
Sad	0	0	5.7	0	17.3	77

Table 4.9 Confusion matrix for expression classification results using QLZM-based AFER on FACES dataset.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	94.1	5.9	0	0	0	0
Disgust	0	94.1	0	0	0	5.9
Fear	0	0	82.4	0	11.7	5.9
Happy	0	0	0	100	0	0
Neutral	11.8	0	5.8	0	82.4	0
Sad	11.8	0	0	0	23.5	64.7

Table 4.10 Confusion matrix for expression classification results using BRIEF-based AFER on FACES dataset.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	100	0	0	0	0	0
Disgust	0	100	0	0	0	0
Fear	0	0	88.2	0	5.9	5.9
Happy	0	0	0	100	0	0
Neutral	0	0	0	0	100	0
Sad	5.8	0	0	0	11.8	82.4

These results show that the emotions are less confused using the BRIEF descriptor. For instance, 27% (see Tables 4.8), 5.9% (see Tables 4.9), and 0% (see Tables 4.10) of the disgust expression are recognized as the sad expression using LBP, QLZM, and BRIEF features respectively. The reason for that confusion is due to the similarity between the extracted features using those face descriptors as illustrated in the first column of Figure 4.9. In this figure, the BRIEF's code of the happy face of young people is similar to some extent to the BRIEF's code of the neutral face of old people.

4.4.2.2 Experiment 5 - Compound Emotions Effect on Texture-based AFER

Case Description: In this experiment, the variation of the 22-compound emotions is considered. This consideration is beneficial for more validation of the optimal parameters and the texture method's ability to recognise a rich set of 6-basic and 15-non-basic expressions plus the neutral expression instead of just the 6-basic expressions in previous experiments. The second benefit is to analyse and understand the effect of the compound emotion on the performance of the texture-based AFER system. The idea here is that the 22 compound expressions have large texture pattern similarities among them which might lead to poorer performance of the classifier in discriminating among them. These similarities are due to the fact that each compound emotion is generated by combining two of the basic emotions, resulting in emotions which have partial similarities in muscles' deformations as shown in Figure 4.10. In the figure, the appearance of the happily-surprise emotion is partly similar to the appearance of both the happy and the surprise expressions. These similarities are highlighted by red and blue rectangles respectively.

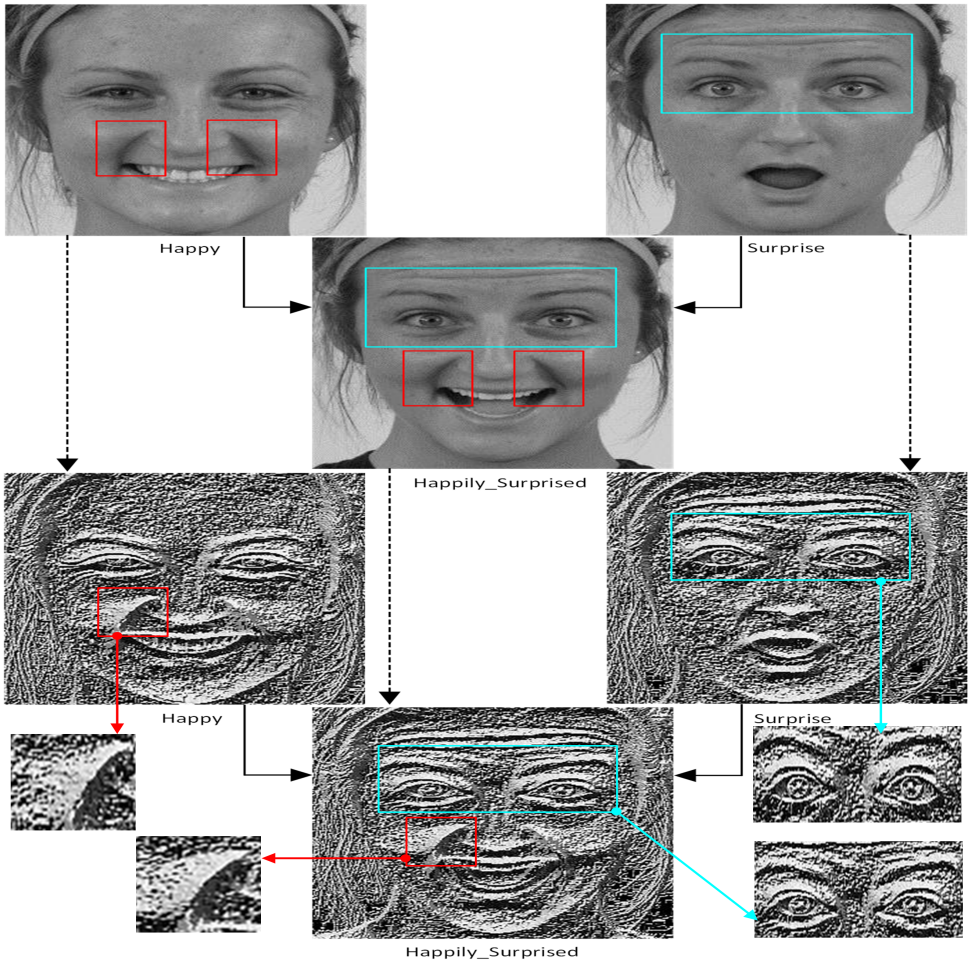


Fig. 4.10 Illustration of the similarity between the texture features among basic (happy and surprise) and compound (happily surprised) expressions. Original image (three top) and BRIEF response (three bottom).

Using a set of 2,200 images from the compound emotion dataset described in Section 3.2.2 and the selected BRIEF (6,11,11) parameters, three SVM classifiers were trained on the 6-basic emotions only (each one trained on different features including: LBP, QLZM, and BRIEF) and another three SVM classifiers were trained on the whole dataset of 6-basic and the 15-non-basic (22-compound) expressions.

Results: The results of all the classifiers are summarized in Table 4.11. These results show that:

- Performance of the three methods is best on the 6-basic expressions and that performance degrades as the compound expression are included and the number of expressions increases.
- QLZM descriptor shows the best performance on 22 compound emotions dataset.
- The performance of texture-based AFER using the three descriptors is not enough to capture the variation of 22 class of expressions and hence to discriminate among them.

Table 4.11 Performance of the LBP, QLZM, and BRIEF method using compound emotion datasets.

	6-basic	6-basic + 15-non-basic(Compound Emotions)
LBP	80.5±1.1	39.8 ±1.7
H-QLZM	82.7 ±3.3	45.5±2.4
BRIEF (present work)	84.3±2.7	42.2 ±1.9

Tables 4.12, 4.14, and 4.13 show the confusion matrices using the classifiers trained on LBP, QLZM, and BRIEF features on the 22 emotions respectively. Those confusion matrices reveal that there are huge confusions among most of the expressions in all the tables of all the features. The simple reason for the low performance and the huge confusions is due to the large similarities between the 6-basic emotions and 15-non basic emotions in the Action Units (AUs) used to generate them (see Table 2.3 in Chapter 2).

Table 4.12 Confusion matrix and accuracy of LBP-based AFER system applied to 22 emotions (a:neutral, b:happy, c:sad, d:fearful, e:angry, f:surprised, g:disgusted, h:happily surprised, i:happily disgusted, j:sadly fearful, k:sadly angry, l:sadly surprised, m:sadly disgusted, n:fearfully angry, o:fearfully surprised, p:fearfully disgusted, q:angrily surprised, r:angrily disgusted, s:disgustedly surprised, t:appalled, u:hate, v:awed).

classes	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
a	90	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.6	0	6.4
b	0	80	0	0	0	0	0	11.2	9.8	0	0	0	0	0	0	0	0	0	0	0	0	0
c	20	0	30	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40
d	10	0	0	50	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	30
e	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40
f	10	0	0	0	0	60	0	0	0	0	0	10	0	0	10	0	0	0	0	0	0	10
g	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	10	20	0	50
h	0	20	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	30	0	0	0	0	0	0	70	0	0	0	0	0	0	0	0	0	0	0	0	0
j	0	0	0	10	0	0	0	0	0	30	0	20	0	20	10	0	0	0	0	0	0	10
k	0	0	0	0	10	0	0	0	0	0	16	0	10	0	0	0	0	0	0	0	30	34
l	0	0	0	0	0	20	0	0	0	0	0	40	0	0	0	0	10	0	0	0	0	30
m	0	0	0	0	0	0	0	0	0	10	10	10	20	0	0	0	0	10	0	10	0	30
n	0	0	0	0	0	0	0	0	20	0	0	10	0	40	0	0	0	0	0	0	0	30
o	10	0	0	30	0	10	0	0	10	0	0	0	0	0	20	10	0	0	0	0	0	10
p	0	0	10	0	0	0	10	0	20	10	0	0	0	20	0	10	0	0	0	0	0	20
q	0	0	0	0	10	0	0	0	0	0	0	0	0	10	0	0	50	0	0	0	0	30
r	0	0	0	0	20	0	10	0	0	0	0	0	0	0	0	0	0	0	0	30	10	30
s	10	10	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	50	0	0	20
t	0	0	0	0	10	0	10	0	0	0	0	0	0	0	0	0	0	0	0	10	20	40
u	0	0	0	0	20	0	0	0	0	0	30	0	0	0	0	0	0	0	0	10	10	30
v	0	0	0	0	20	0	0	0	0	0	30	0	0	0	0	0	0	0	0	10	0	40

Table 4.13 Confusion matrix and accuracy of QLZMA-based AFER system applied to 22 emotions (a:neutral, b:happy, c:sad, d:fearful, e:angry, f:surprised, g:disgusted, h:happily surprised, i:happily disgusted, j:sadly fearful, k:sadly angry, l:sadly surprised, m:sadly disgusted, n:fearfully angry, o:fearfully surprised, p:fearfully disgusted, q:angrily surprised, r:angrily disgusted, s:disgustedly surprised, t:appalled, u:hate, v:awed) using BRIEF face descriptor.

classes	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
a	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
b	0	70	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0
c	10	0	20	0	0	0	0	0	0	0	0	10	0	0	10	0	0	0	0	0	0	50
d	0	0	0	30	0	10	0	10	10	0	0	0	0	10	10	0	0	0	0	0	0	20
e	0	0	0	0	40	0	10	0	0	0	0	0	0	0	0	0	10	0	0	0	0	40
f	0	0	0	0	0	70	0	0	0	0	0	10	0	0	10	0	0	0	0	0	0	10
g	10	0	10	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
h	0	20	0	0	0	0	0	70	0	0	0	0	0	0	10	0	0	0	0	0	0	0
i	0	10	0	0	0	0	10	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0
j	0	0	10	0	0	10	0	0	0	20	0	0	0	30	10	10	0	0	0	0	0	10
k	0	0	10	0	0	0	0	0	0	0	10	0	10	0	0	0	10	0	0	10	10	40
l	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	10	0	0	0	0	40
m	0	0	0	0	0	0	0	0	0	20	10	10	10	0	0	0	0	10	0	10	0	30
n	0	0	10	0	0	10	10	0	10	10	0	0	0	40	0	0	0	0	0	0	0	10
o	0	0	0	40	0	10	0	10	10	0	0	0	0	10	20	0	0	0	0	0	0	10
p	0	0	0	10	0	0	10	0	30	0	0	0	0	0	0	50	0	0	0	0	0	0
q	0	0	0	0	10	0	0	0	0	0	0	10	0	0	0	0	50	0	0	0	0	30
r	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	50	0	30	0	10
s	0	0	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0	60	0	0	20
t	0	0	0	0	10	0	20	0	0	0	0	0	0	0	0	0	0	10	0	30	10	30
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	30	50
v	0	0	0	0	30	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	**55**

Table 4.14 Confusion matrix and accuracy of using BRIEF-based AFER system applied to 22 emotions (a:neutral, b:happy, c:sad, d:fearful, e:angry, f:surprised, g:disgusted, h:happily surprised, i:happily disgusted, j:sadly fearful, k:sadly angry, l:sadly surprised, m:sadly disgusted, n:fearfully angry, o:fearfully surprised, p:fearfully disgusted, q:angrily surprised, r:angrily disgusted, s:disgustedly surprised, t:appalled, u:hate, v:awed).

classes	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**	**v**
a	88	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
b	0	87	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	5
c	17	0	67	0	8	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
d	17	0	0	8	0	8	0	17	0	8	0	8	0	8	17	0	0	0	0	0	0	9
e	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
f	10	0	0	0	0	48	0	0	0	0	0	25	0	0	8	0	0	0	0	0	0	9
q	0	0	8	0	8	0	17	0	0	0	8	0	8	0	0	8	0	0	0	33	8	2
h	0	17	0	0	0	0	0	57	8	0	0	0	0	0	8	0	0	0	0	0	0	10
i	0	8	0	0	0	0	0	0	64	0	0	0	0	0	0	8	0	0	0	0	0	20
j	8	0	8	8	0	0	0	0	0	18	8	17	0	33	0	0	0	0	0	0	0	0
k	0	0	17	0	25	0	0	0	0	0	42	0	8	0	0	0	0	0	0	0	8	0
l	0	0	0	0	0	0	0	8	0	0	0	59	0	8	0	0	25	0	0	0	0	0
m	0	0	33	0	8	0	0	0	8	17	8	0	26	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	25	8	1	8	25	25	0	0	8	0	0	0	0	0
o	0	0	0	0	0	0	0	8	0	0	0	8	8	0	56	0	0	0	0	0	0	20
p	8	0	0	0	0	0	0	17	8	0	0	8	8	0	8	43	0	0	0	0	0	0
q	0	0	8	0	8	0	0	0	0	0	0	33	0	0	1	0	42	0	0	8	0	0
r	0	0	0	0	25	0	25	0	0	0	0	0	17	0	0	0	0	8	0	0	25	0
s	0	0	8	0	0	0	0	8	0	0	8	0	8	0	2	0	0	8	50	8	0	0
t	0	0	0	0	25	0	17	0	0	0	8	0	17	0	0	0	0	8	0	8	17	0
u	0	0	0	0	58	0	8	0	0	0	8	0	0	0	1	0	0	0	0	8	17	0
v	8	0	0	0	0	0	0	17	0	0	0	25	0	0	25	0	0	0	0	0	0	25

4.5 Discussion

In this chapter, texture features were used to describe a face image, build a texture-based AFER system, and measure the influence of the ageing and compound emotion appearances on the performance of AFER. The objective was first to introduce BRIEF features as a new face descriptor for the application of AFER. The second objective was an assessment of the development of the texture-based AFER system in two challenges regarding the expression: the ageing effect and compound emotions.

A parameter sensitivity analysis of BRIEF's free parameters was performed. It showed insensitivity to the choice of BRIEF's representation to the patch size and the descriptor length. This means that few calculations are needed for the BRIEF descriptor. We implemented the proposed approach and compared it to different face representation methods that were tested on the CK+ dataset. The experimental results clearly show that the BRIEF-based AFER system outperformed the LBP-based one in terms of the recognition rate and the size of the representation. For example, our method achieved a classification rate of 96.0% with a histogram of integers ranged in [0,63], while the performance using the LBP method was 82.4% with a histogram of integers ranged in [0,255]. We also obtained comparable performance to the BOW and QLZM but with smaller standard deviation from the mean recognition rate suggesting that the proposed model is more stable.

Having ensured that the developed AFER system based on BRIEF features worked satisfactorily with automatic expression recognition in general, the second objective of this chapter was to apply BRIEF and further test for more complex cases which were recognising the expression across the extensive range of ages and recognising 22-compound expressions. The results indicated that BRIEF achieved satisfactory performance on a range of facial expression recognition datasets with different characteristics. That might be becasue BRIEF selects a subsample of pixel pairs at Gaussian weighted random locations from the described area which gives a good sample of the area. These results also indicate that the texture measurements of three different face descriptors methods proved that the age and compound emotions have a significant effect on the system performance. In the case of the ageing, the best result was obtained using age-dependent classifiers. In the case of classifying 22-compound emotions, despite the good results using QLZM compared to other methods with the same configuration, the texture features presented in this chapter are not sufficient to discriminate among 22-compound emotions and more investigations are required. Figures 4.11, 4.12, 4.13, 4.14, and 4.15 show examples of the BRIEF code images (BRIEF response) from the three datasets.

Moving on from texture measurements regarding the effect of the problems under study on the performance of facial expression understanding, the next chapter deals with another type of measurements, that of shape analysis and shape measurements.

Fig. 4.11 Example results of the BRIEF descriptor on the CK+ dataset: original images (first two rows) display first 8 frames of one subject displaying the surprise emotion in continuously increasing intensity from neutral to peak, BRIEF code value (last two rows).

Fig. 4.12 Example results of the BRIEF descriptor on CK+ dataset: original images (first two rows) display second 8 frames of one subject displaying the surprise emotion in continuously increasing intensity from neutral to peak, BRIEF code value (last two rows).

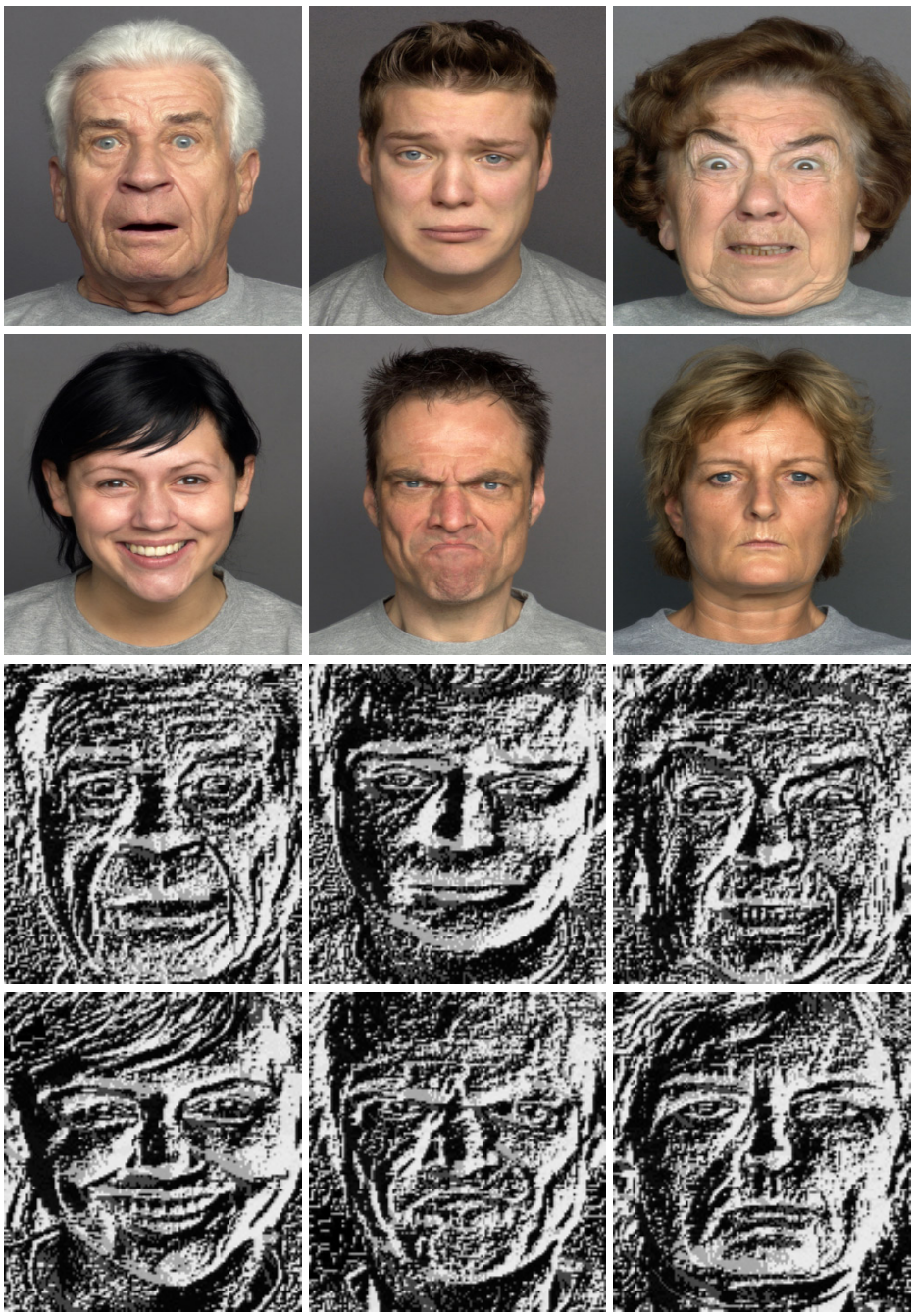


Fig. 4.13 Example results of the BRIEF descriptor on FACES dataset: original image (first row), BRIEF code value (second row).

Fig. 4.14 Example results of the BRIEF descriptor on the compound emotion dataset: original images (first 3 rows) display 11 compound emotions form left to right neutral, happy, sad, fearful, angry, surprised, disgusted, happily surprised, happily disgusted, sadly fearful, sadly angry respectively, BRIEF code value (last 3 rows).

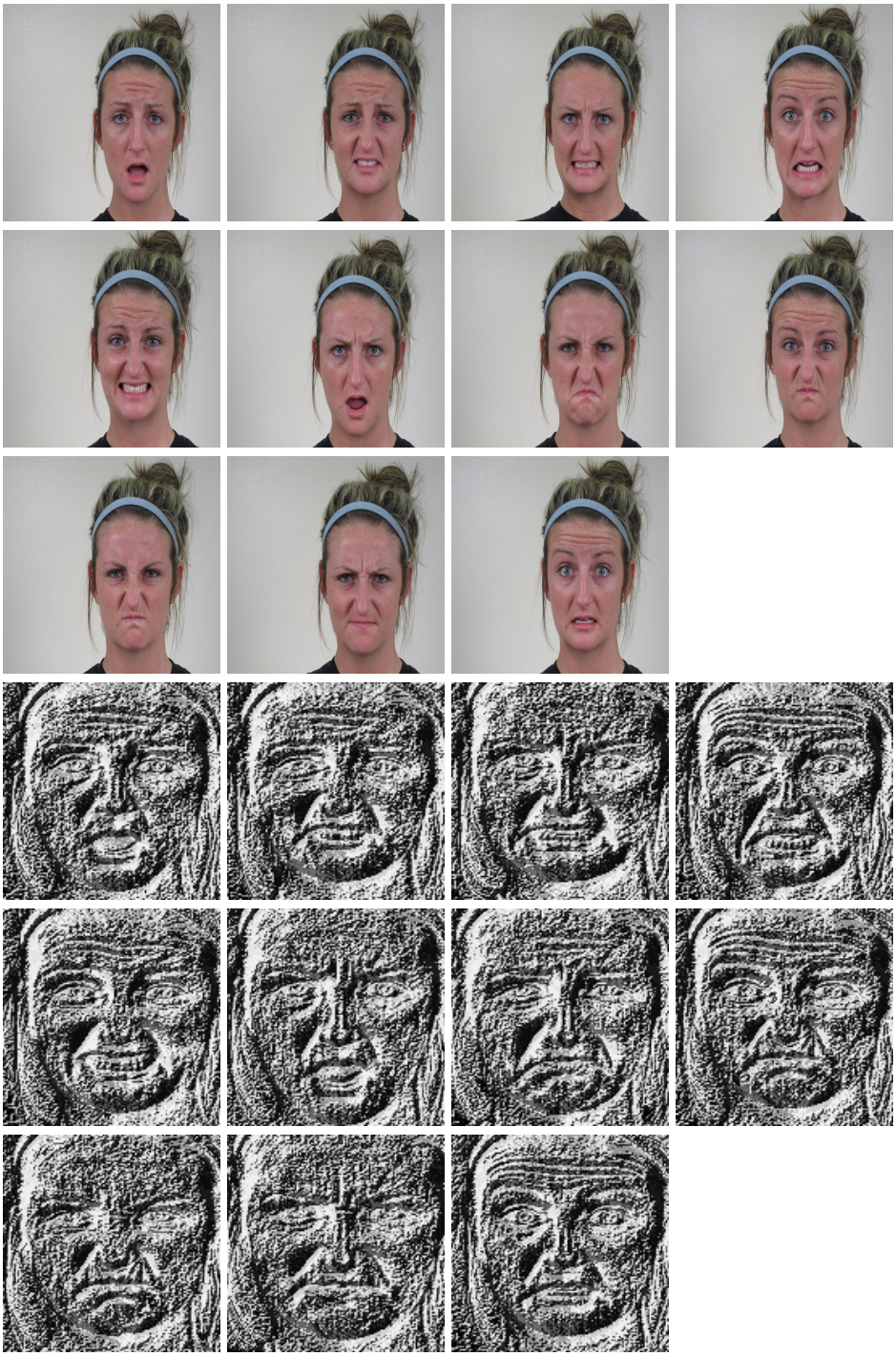


Fig. 4.15 Example results of the BRIEF descriptor on the compound emotion dataset: original images (first 3 rows) display 22 compound emotions form left to right sadly surprised, sadly disgusted, fearfully angry, fearfully surprised, fearfully disgusted, angrily surprised, angrily disgusted, disgustedly surprised, appalled, hatred, and awed respectively, BRIEF code value (last 3 rows).

Chapter 5

Development and Comprehensive Evaluation of RFRV-CLM Based FEL and AFER

5.1 Introduction

As stated in Section 2.2.3.2, a robust and automatic facial feature point detector is essential and would be widely used for facial image analysis in general and AFER in particular in which the detector is used as a tool to localize the geometrical structure of the target object and is used then for further analysis such as feature extraction and pattern recognition. This chapter describes a system for fully automatic FEL using the application described by Cootes et al. (2012) and Lindner et al. (2015) of using RFRV in CLM. The first objective of this chapter is to examine the proposed FEL, along with an investigation of which are the best parameters of the RFRV-CLM framework to fit the problem of facial expression localisation of faces exhibiting wide variability in facial appearances. The second objective is to test the strength of the proposed detector to accurately localize the target facial feature points with the presence of a large range of shape deformations including human ageing, compound emotions, and expressions's intensity. The third objective is to measure the influence of the problems' patterns on the shape features and hence their impact on the performance of shape-based AFER.

For ease of understanding, this chapter starts by giving our motivations for selecting the RFRV-CLM framework for facial expression localization in 5.2 and it then gives an explanation of the RFRV-CLM algorithm in 5.3. In Section 5.4, we introduce our automated system for Automatic FEL. Section 5.5 presents results of the proposed system in FEL and AFER tasks respectively under the effect of the problems under study with the comparison made against the results of alternative methods evaluated on the same datasets. Finally, Section 5.6 concludes the chapter by providing a summary of the key findings. To

the author's knowledge, no previous work studied the sensitivity of landmark localisation to age.

5.2 Motivations

This section presents the motivations for proposing the RFRV-CLM framework for the problem of automatic facial expressions localisation.

RFRV-CLM is one of the shape regression-based methods. The key idea of regression-based methods is the use of a regressor instead of a classifier to vote for the point position based on the information in nearby regions. In spite of learning a regressor being more difficult than learning a classifier, a regressor can provide more useful information about the target location, such as the distance of negative locations (patches) from the positive locations (patches). This satisfies our requirements of localizing patches that are related to age and expressions variations from the face image. Conversely a classifier can only determine whether the image patch is negative or positive. Cootes et al. (2012) and Lindner et al. (2015) introduce RFRV to the constrained local models (CLM) and found that the RFRV-CLM framework outperforms alternative discriminative methods (classification and boosted regression) trained on the same datasets, and it has been applied successfully to the automatic landmark points localization in several applications and shown remarkable performance (Bromiley et al., 2015a,b, 2016; Lindner et al., 2015; Lindner and Cootes, 2015; Lindner et al., 2013).

These findings motivated us to look into the generalisation capability of the RFRV-CLM method for the problem of expression localization. The underlying hypothesis is that automatically finding the optimal position of facial feature points in the input image with small mislocalization error will help to obtain robust and accurate results for further analysis such as feature extraction and facial expression recognition as the target of our thesis is the building of AFER.

5.3 Method

This section describes the RFRV-CLM framework. The RFRV-CLM algorithm consists of two main parts. The first part is random forest regression voting (RFRV), in which a RF regressor is used to search patches and detect the location of a face region in an image. The second part is the use of constrained local models (CLMs) to fit the most likely points to the observed shape model from the training data.

5.3.1 Random Forest Regression Voting (RFRV)

The principle idea of the RFRV algorithm is that the correct answer can be obtained via the majority votes of a number of independent regressors. In other words, the problem is solved by combining predictions of a set of decision trees. Each tree is trained by finding features that best split a given sample of data. The resulting leaf node from a decision tree contributes to the final decision, reached by majority voting. In relation to the problem of point localization used in this thesis, a set of random forest regressors are trained independently to localize points in an image. For each point i in each image I, a random displacement d_i is generated by sampling within the range $\pm d_{max}$. Image patches of size w^2_{patch} are then sampled at these displacements, and features f_i (Haar$-$features Viola and Jones (2001)) are extracted from those patches (see Figure 5.1 left). To make the localizer pose-invariant and to avoid the error due to the inaccurate initialization of pose during model fitting, the process are repeated with random scales and angles. Each tree is then trained on the pairs (f_i, d_i) of a bootstrap to estimate the most likely positions of the target point. At each node, a decision tree is trained by finding features f_i and threshold t that best divide a given sample of data into two compact groups by minimizing the following equation.

$$G_T(t) = G(\{d_i : f_i < t_f\}) + G(\{d_i : f_i \geq t_f\}) \tag{5.1}$$

Where f_i is the feature from sample i and $G(..)$ is a function to find the threshold that best splits the features based on the variance in displacement d_i resulting from the split. The aim is to minimize the entropy in the branches when spitting the nodes using

$$G(d_i) = N \log |\Sigma| \tag{5.2}$$

where N is the number of the displacement in d_i and Σ the covariance matrix. This process is terminated at a maximum depth D_{max} of the trees or minimum number of samples N_{min} at the node and repeated to generate a forest of n_{trees}. The tree outputs a displacement from the given patch of the image, based on learned displacements from the training set, along with a weight of the prediction (see Figure 5.1 right). For each leaf, the mean and standard deviation of the displacements of all training samples arriving at that leaf are stored during training, and this will give the response during testing. Predictions are made using a single vote per tree and all the votes are accumulated to create the response image. The response images is then passed to an optimiser (CLM), to shift the shape model points to the best fitting points given by the random forest.

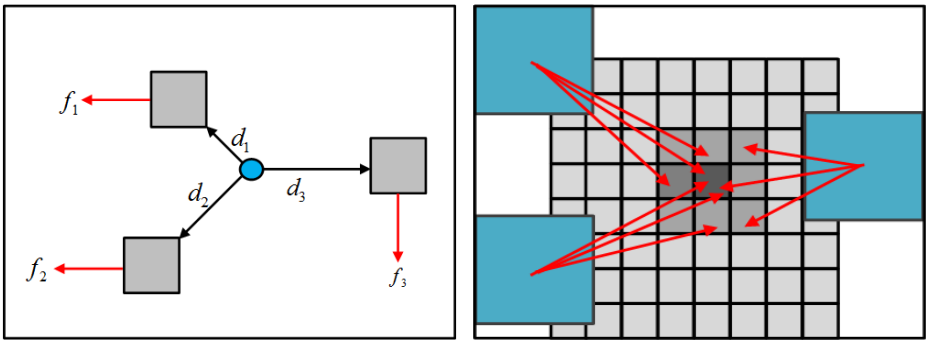


Fig. 5.1 Patches sampled at random displacement d_i (left) and predicted displacements of a random forest (right).

5.3.2 Constrained Local Model (CLM)

CLM is a model for matching a set of points to an image, where the global shape constraints are combined with local models of the pattern of intensity. Given a set of training images with manual annotations x_l of a set of n landmark points, where $l = 1 \ldots n$, a statistical shape model is built by applying principal component analysis (PCA) to the aligned facial shape vectors (Cootes et al., 2001), creating a model with the form:

$$x_l = T_\theta(\hat{x}_l + P_l b + r) \tag{5.3}$$

where x_l represents the point's l position in the reference frame, $\hat{x}_l$ represents the mean shape in the reference frame; P_l is a matrix of the set of eigenvectors corresponding to the highest eigenvalues, which describe different modes of variation; b is a set of parameter values of the shape model; r allows small deviations from the model; and T_θ applies a global similarity transformation with parameters θ.

To fit the model to new image, the best quality of fit Q is found by optimizing the parameters $P = \{b, \theta, r_l\}$

$$Q(P) = \sum_{l=1}^{n} C_l(T_\theta(\hat{x}_l + P_l b + r_l)) \;\; s.t \;\; b^T S_b^{-1} b \leq M_t and |r_l| < r_t \tag{5.4}$$

where C_l is a cost image for the fitting of landmark points l, S_b is the covariance matrix of shape model parameters b, M_t is a threshold on the Mahalanobis distance, and r_l is a threshold on the residuals. M_t is chosen using the cumulative distribution function (CDF) of the χ^2 distribution so that 98% of the samples from a multivariate Gaussian of the appropriate dimension would fit. This ensures a plausible shape for a model parameter.

5.3.3 RFRV in the CLM Framework

The idea of using RFRV in the CLM framework is to use RFRV described in Section 5.3.1 to vote for the best position for every landmark point in order to provide the cost image

C_l in Equation 5.4. During training of the RFRV-CLM the shape model from Equation 5.3 is used to estimate the global pose, θ, of the object in each image by minimising $|T_s(\hat{x}) - x|^2$. Each image is resampled into a standardised reference frame by applying $I_r(i,j) = I(T_\theta^{-1}(i,j))$. The model is scaled so that the width of the reference frame of the mean shape is a given value, w_{frame}. A RFRV is then trained for every point in each image as described in Section 5.3.1 to be used during the fitting of the model to the new image

That fitting of the RFRV-CLM model to new image starts by estimation of the pose parameters b and θ on the new image. The reference frame is used for the search operation in order to allow for the variations in both scale and pose across the dataset. The initial estimate of shape and pose parameters are used to sample the region of interest of the image into the reference frame and an area around each landmark is searched within $\pm d_{search}$ displacements to predict the position using the pre-trained model as described in Section 5.3.1. The response image R is computed at every area of the search for each point independently. After that an optimization for shape model parameters is performed, in which all the response images of all the landmark points are combined with the global shape model of trained SSM find the best quality of fit Q over the shape parameters $P = b, \theta, r_l$

5.4 Fully Automatic Expressions Localization System

The architecture of the proposed FEL is composed of two main stages. The first stage is a global model or global search, which finds the approximate positions of the eye centres using the RFRV technique described in 5.3.1, while the second stage is a local model or local search which uses the estimated landmark values that are obtained from the global model to locate all the facial key points using the RFRV-CLM described in 5.3.2, by searching in a Coarse-to-Fine RFRV-CLM fashion. Figure 5.2 illustrates the training and testing phases of the proposed FEL system.

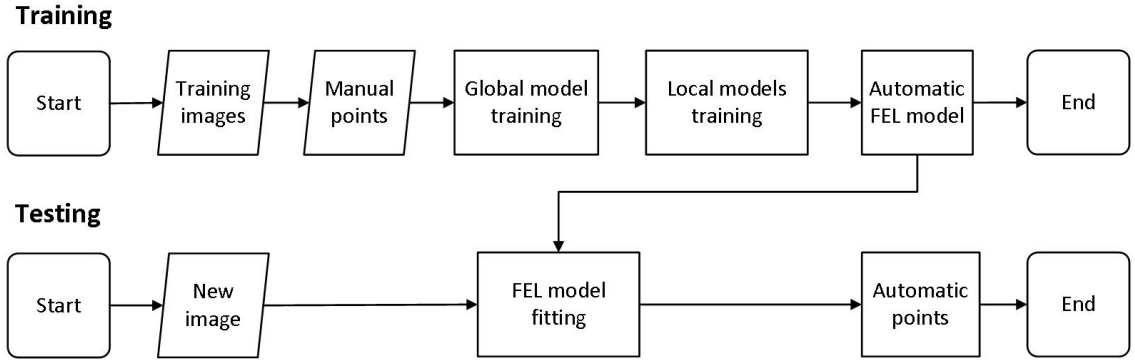


Fig. 5.2 Flow-chart giving an overview of the proposed automatic facial expression localization (FEL) system. See main text for details.

5.4.1 Global Models

Two reference points with a bounding box are set to capture the object of interest. The training phase of the global model runs as follows: for each image in the training set, a bounding box with two reference points corresponding to the eyes'centres is set to determine the face area required to be captured and its position, orientation and scale. These two points are chosen because they have approximately a constant separation across individuals and hence this allows us to capture the same area for all the images in the dataset (Dodgson, 2004). For each image many samples i are sampled with arbitrary displacement in angle, scale, and position, giving the detector the ability to be invariant to the local search range. For each sample i, Haar features f_i are extracted at the random displacement d_i within the bounding box. A RF is then constructed in which each tree is trained independently on a bootstrap sample of pairs of (f_i, d_i) as described in 5.3.1 to learn the functional dependency between the centre of the patch and the true position, producing a full object (face) detector to be used on the test images.

The testing phase of the global model on a new image runs as follows: a sliding window approach is used to scan the new image and patches of several combinations of scales, angles, and pose are sampled. Features f_i are then extracted at each combination and sent to the detector to make predictions of the true centre of the reference frame, and the response images R is then obtained. Once all the response images R are calculated for every combination of scale, angle and pose, all the maxima will ranked and the predictions with the most votes are used to get the position of two points from the reference frame corresponding to those in the training set. The final output from the global model is a bounding box with two points used to initialize the local model. Figure 5.3 illustrates the 10 best fits of the output of the global model detector.

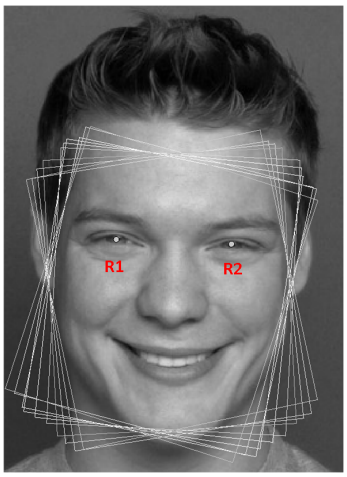


Fig. 5.3 Visualization of 10 best fits of the global model detecting the approximate position of two reference points of the eyes centres.

5.4.2 Local Model

Each image in the training set is labelled with n landmark points, x_l, where $l = 1 \ldots n$. The shape of the face can be encoded as:

$$x = (x_1, ..., x_n, y_1, ..., y_n)^T \tag{5.5}$$

Then, from the points and the images, an SSM is trained as described above in Equation 5.2. Within the local model, a set of single local detectors are trained independently for each point using the method described in 5.3.1 to localize points in a new image. During training, for every point l in each image I, a random displacement d_j is generated within the range $\pm d_{max}$. Image patches of size w^2_{patch} are then sampled at these displacements, and features f_j are extracted from those patches. To make the localizer pose-invariant and to avoid the error due to the inaccurate initialization of pose during model fitting, this process is repeated with random scales and angles. Each tree in the forest is then trained on the pairs (f_j, d_j) to estimate the most likely positions of the target point. At each node, a decision tree is trained by finding features f_j and threshold t that best splits a given sample of data into two compact groups as described in Section 5.3.1.

During testing, the process of matching the model to a new image runs as follows: for every point around the initial estimate from the previous model or from manual initialization, a set of patches are sampled randomly. Features are extracted at each location in the grid and fed to the regressor to predict the most likely matching position. This process is performed independently for each point. The shape model from Equation 5.2 is then used to regularise the results via a series of searches, finding the parameters which maximize the total votes as in Equation 5.4. The number of the search iterations N_{search} depends on the searching range $\pm d_{search}$ as well as the error between the initialization of the model and the target points. Figure 5.4 shows the superposition of vote accumulation images of the 76-point face model (left) and the corresponding captured points (right).

5.4.3 Coarse-to-Fine RFRV-CLM

To improve the accuracy and efficiency of the local models, the local search is implemented using RFRV-CLM in a coarse-to-fine multi stage framework with increasing resolution of frame width by changing the number of pixels in the reference frame. In the first stage, the lower resolution of the reference frame is used to roughly estimate the position of every landmark point. The second stage uses a higher resolution of the reference frame to refine the results of the first stage and so on. This process contributes to learning a good initial model and searches for the optimal model smoothly to avoid missing fairly good intermediate models in subsequent procedures (see Figure 5.5).

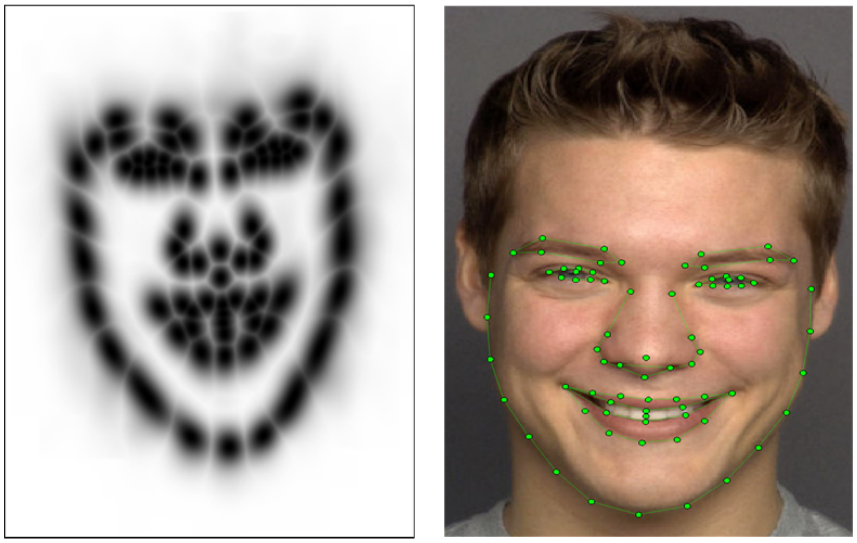


Fig. 5.4 Fully automatic FEL model: superposition of 76 local models votes (left), and final automatic points' positions outlining the face components (right).

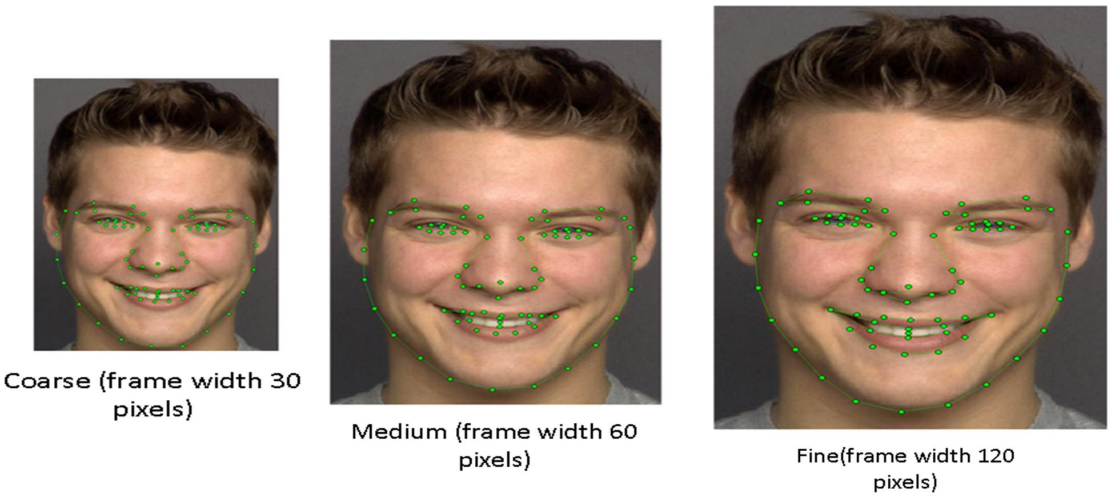


Fig. 5.5 Illustration of three local stages of RFRV-CLM searches with the model iterating over the various frame width (w_{frame})

5.4.4 Combined Global and Local Models

Once the Global and Local models are trained and combined, the fully automated facial landmark annotation system is started by putting a dense annotation of n points onto the test image. The single global detector is then started to predict the most likely position of the object of interest in the input image, producing a bounding face box and two reference points r_1 and r_2. The predicted poses regarding r_1 and r_2 from the global model will then be used to initialize the local models which start from the first stage with low resolution to estimate the position of every landmark point from the model and then use the second stage of higher resolution to refine the solution to give the optimal location coordinate for every model point, ending in a series of automatic points.

5.5 Experimental Evaluation

A series of experiments was performed in order to evaluate the performance of the RFRV-CLM framework in FEL and AFER tasks against the effects of the problems under study.

5.5.1 Evaluation in Facial Expression Localization

A series of experiments was performed in order to evaluate the performance of the RFRV-CLM framework in FEL task. The first experiment is focused on parameter optimization in order to find which are the best RFRV-CLM free parameters for the problem of facial expression localization. Then several experiments are performed for more validation of the method and the optimal parameters under a large range of facial activities and their overlap with the expressions variation that might hinder the process of placing the points around the target contour correctly and hence might lead to poor ability by the automatic detector to accurately detect the key points.

Database: The proposed FEL is evaluated using five different facial expression datasets, including the Cohn−Kanade dataset for expression and its intensity, the FACES, LifeSpan, and NEMO datasets for ageing and expression, and the compound emotions data set for compound expressions (see Section 3.2 for more details about the datasets).

Evaluation Metric: The accuracy of the landmark localization is tested by comparing the locations of automatically detected points against the manual annotation (ground-truth) across each test image. Following the common practice in Cristinacce and Cootes (2008), the errors are given as the Euclidean distance between the detected point and the corresponding manual annotation and displayed as a Cumulative Density Function (CDF) of the mean point-to-point error as a percentage of the Inter-Ocular Distance (IOD), defined as:

$$e_n = \frac{1}{nd_{eyes}} \sum_{i=1}^{n} |p_a^i - p_m^i| \tag{5.6}$$

Where n is the number of model points, p_m^i is the manually annotated location of point i, p_a^i is the automatically detected facial point location, and d_{eyes} is the distance between the pupil centres, defined as the distance between the centres of the eyes i.e. $|\hat{x}_{lefteye} - \hat{x}_{righteye}|$. All the experiments are entirely dataset-dependent and subject-independent for the training and testing. Some experiments are dataset-independent for the training and testing. We further estimated the error in millimetres, using a mean IOD in humans around 63 mm (Dodgson, 2004) .

Global model training and testing during the training of the global model, we sample 336 patches in 20 random positions (within 50% of the reference frame position) for 15 random scales (in the range of $\pm 15\%$) and angles (in the range of $\pm 15°$). Including the true pose. We then trained a RF consisting of five trees, and used random subsets of size

500 when there were more than 500 samples to be processed at a node. The stopping criteria for node splitting were either a tree depth of 10 or fewer than 5 samples per node. During testing, we used the output detector to scan every test image at seven orientations ranging from -30 to 0 and at a range of scales such that the height of the bounding box is 30% to 60% of the image height. Figure 5.3 illustrates the 10 best fits of the output of the global model detector.

5.5.1.1 Experiment 1 - RFRV-CLM's Free Parameters Optimization

Case Description: Usually in the case of facial expression analysis regarding the localization of points, a large error in the distance between the automatically labelled point and the manually labelled point is unacceptable, since subtle changes in facial expression will be not extracted due to the errors in facial points localization. Therefore, in this experiment, we test, validate and optimize the parameters of the proposed method in order to find the optimal parameters that give the best and acceptable error for the problem of facial expression point's detection.

There are two important sets of parameters for the RFRV-CLM to optimize. The first type is the parameters of the RFs: the number of trees, n_{trees}; the number of the random features to be considered at each node, n_{feat}; the minimum number of training examples at each node, N_{min} and the maximum depth of each tree, D_{max}. The second type of parameters are those relating to the extraction of data from the image: the width of the reference frame, w_{frame}; the size of the sampled patch w_{patch}, the range used to generate random displacement for extracting Haar-like features, d_{max} and the search range for each search; d_{search}.

Since the RFRV-CLM is applied using a multi-stage framework, first, experiments are performed to optimize the first stage parameters. These are used then to initialize and optimize the second stage parameters and so on. Every experiment is initialized with the global model to detect a face region and the two reference points. In every experiment the model was trained on 50% of the data, producing an n-points model then is tested on the other 50%. The training and testing data was then interchanged and the experiment repeated.

The optimal parameters were found by performing grid-search experiments of a series of two-fold-cross-validation of three repeats on the FACES dataset in which w_{frame} is set to a frame width of (10, 20, 30, ..., 160), w_{patch} is set to a patch size of (12, 15, 18, 21, 24), n_{trees} is set to (1, 2, ..., 9, 10). The N_{min}, n_{feat}, D_{max}, d_{max}, and d_{search} parameters are fixed to 1, 500, 25, 12, and 7 respectively with random displacements in scale (in the range of 22 %) and rotation (in range of 13°) as they showed a constant performance (Lindner et al., 2013).

Although the datasets described in this thesis covered a wide range of expressions, the FACES dataset is chosen for the optimization experiment since it covers a wide range

of variations including subjects, age, gender, and expressions. In addition, all of these variations are equally distributed in this dataset. To be sure of that selection, the SSM are built from the images of each dataset with the corresponding manual landmarks. Then the modes of variations are extracted from each shape model. The number of modes represent 99% of the data is summarized in Table 5.1.

Table 5.1 Modes of variations of the statistical shape models of three datasets

Database	Subjects	Images	Landmarks	Modes
Ageing (FACES)	171	2052	76	**59**
compound (CE)	100	2200	78	51
intensity (CK+)	123	180 sequences	70	31

Results: Figures 5.6, 5.7, and 5.8 show complete results of w_{patch}, w_{frame}, and n_{trees} of the three stages respectively. These results indicate that:

- Varying the patch size has a little effect on localization performance and the best quality of fit with the minimum point-to-point-error is obtained with 21 pixels for the three stages as illustrate in Figure 5.6.
- Varying the frame width has a significant effect on the performance as shown in Figure 5.7. For instance, with the optimal patch size of 21 pixels, the optimal frame width of stage 1 is 30 pixels, the optimal frame width of stage 2 is 60 pixels using a series of 30-60, and the optimal frame width of stage three is 120 pixels using a series of 30-60-120 and any additional frames do not show significant improvement.
- In the earlier stages of the models one or two trees are sufficient and any further trees do not contribute to any improvement as illustrated in Figure 5.8. For the third stage of the model best performance of quality of fit is obtained with seven trees and any increase in the number of trees does not show any improvement in the performance.
- the parameters have similar performance in all the stages in which varying the frame width significantly effect on the performance compared to the patch size and number of trees. Table 5.2 describes the optimal parameters of the three stages.

Figure 5.9 shows the effectiveness and summary results of using the RFRV-CLM method and the optimal parameters in several combinations of single and multi-stages on the FACES dataset. These results reveal that the three stage model with a 30-60-120 series of increasing frame width gives the best results with a wide range of convergence across 6-basic expressions and a wide variety of different ages. The mean point-to-point error between the manual and automated points for this data set (FACES dataset) was within 3.4% (2.14 mm) on 99% of all the data.

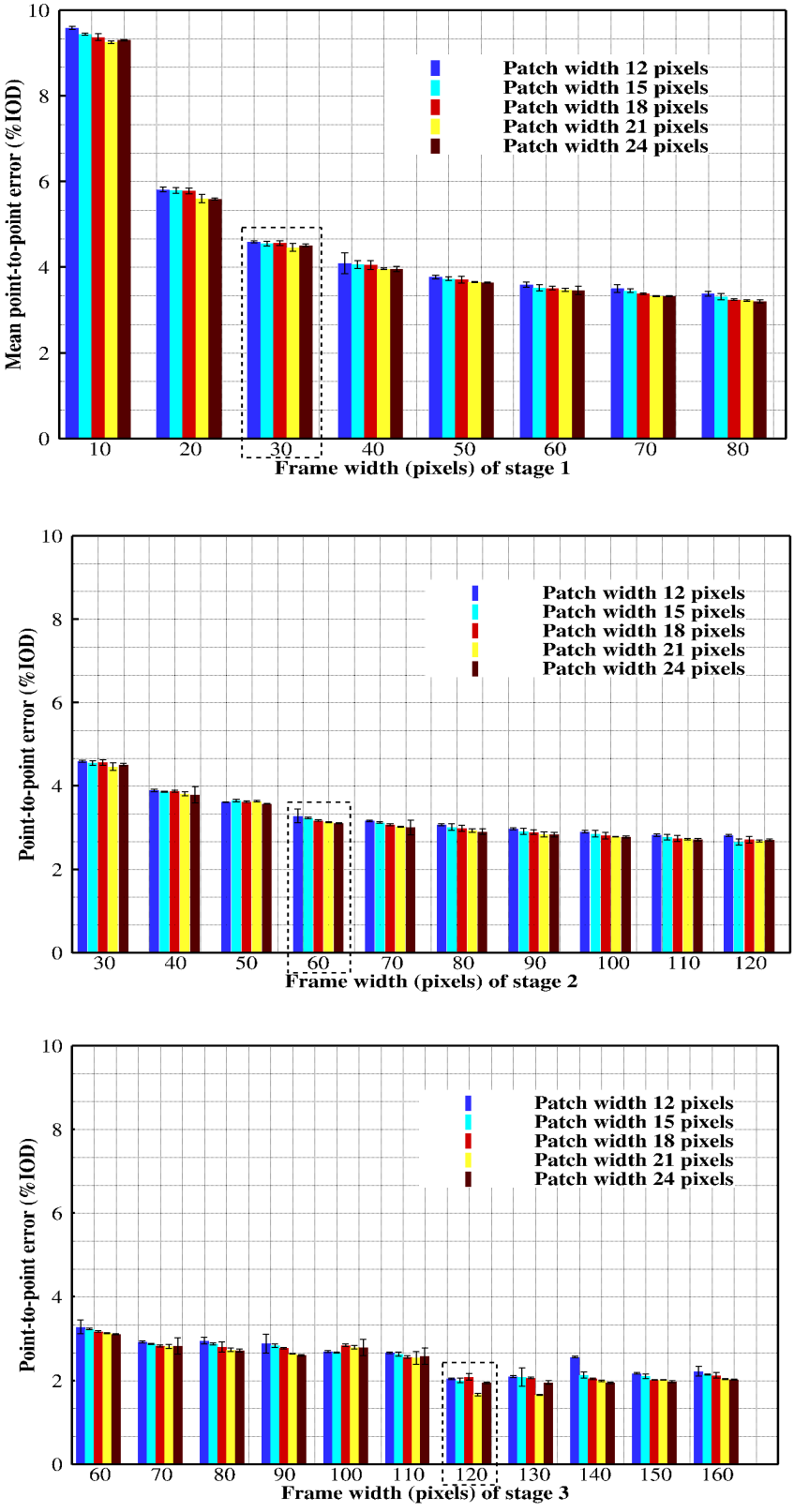


Fig. 5.6 Patch size and frame width optimization results with 1-stage (top), 2-stages (middle), and 3-stages (bottom). Patch size 21 pixels shows the best performance for the 3 stages with frame width 30, 60, 120 pixels of stage 1, stage 2, and stage 3 respectively. Stage-2 is initialized by stage-1's results. Stage-3 is initialized by stage-2's results. Performance is given as a point-to-point error as a percentage of the IOD. Error bars are given as a standard deviation of the repeat of three runs.

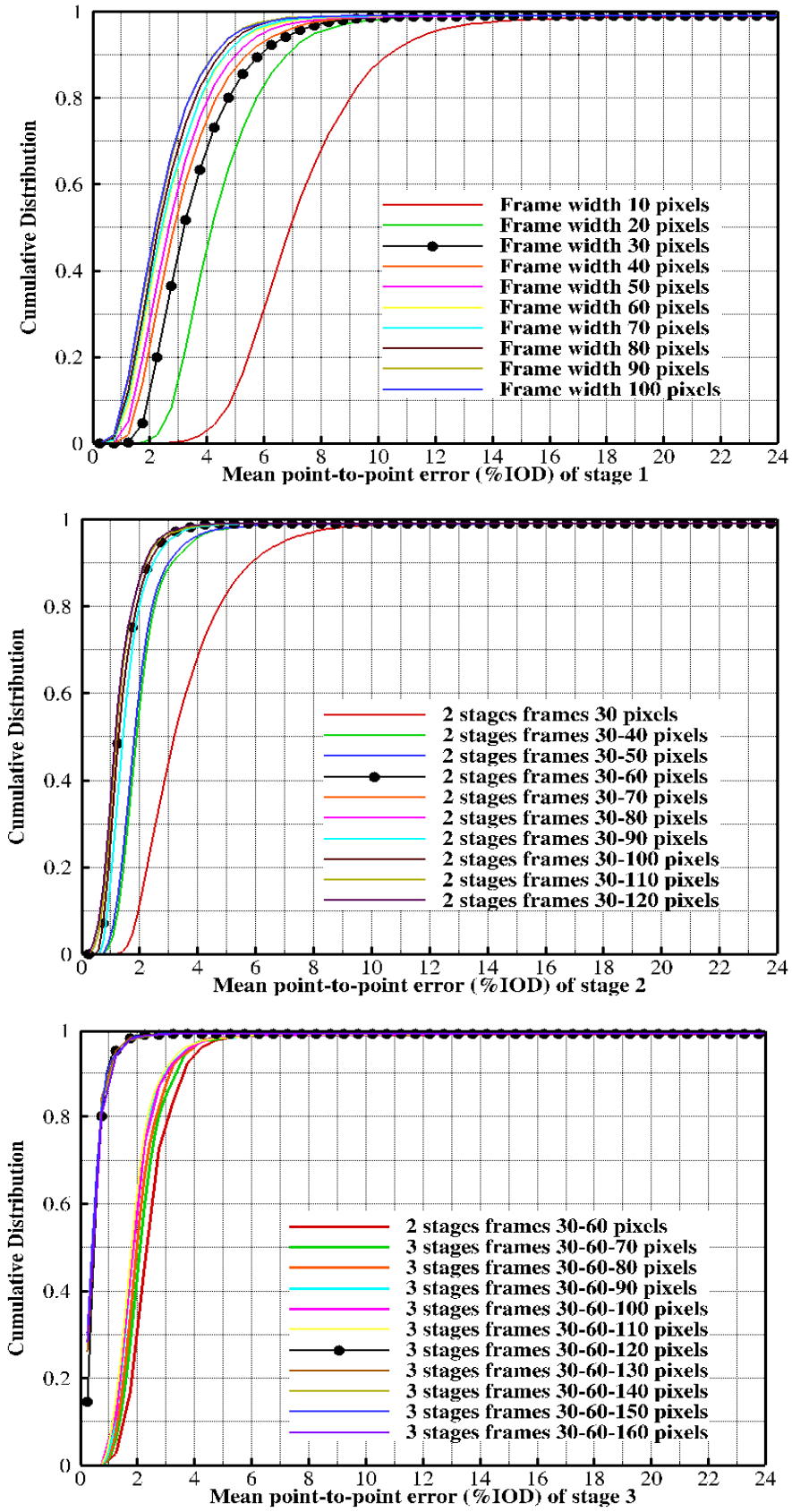


Fig. 5.7 Frame width optimization results with 1-stage (top), 2-stages (middle), and 3-stages (bottom). Frame width of 30 pixels, 60 pixels, and 120 pixels show the best performance for stage-1, stage-2, and stage-3 respectively. Stage-2 is initialized by stage-1's results. Stage-3 is initialized by stage-2's results. Performance is given as a point-to-point error as a percentage of the IOD.

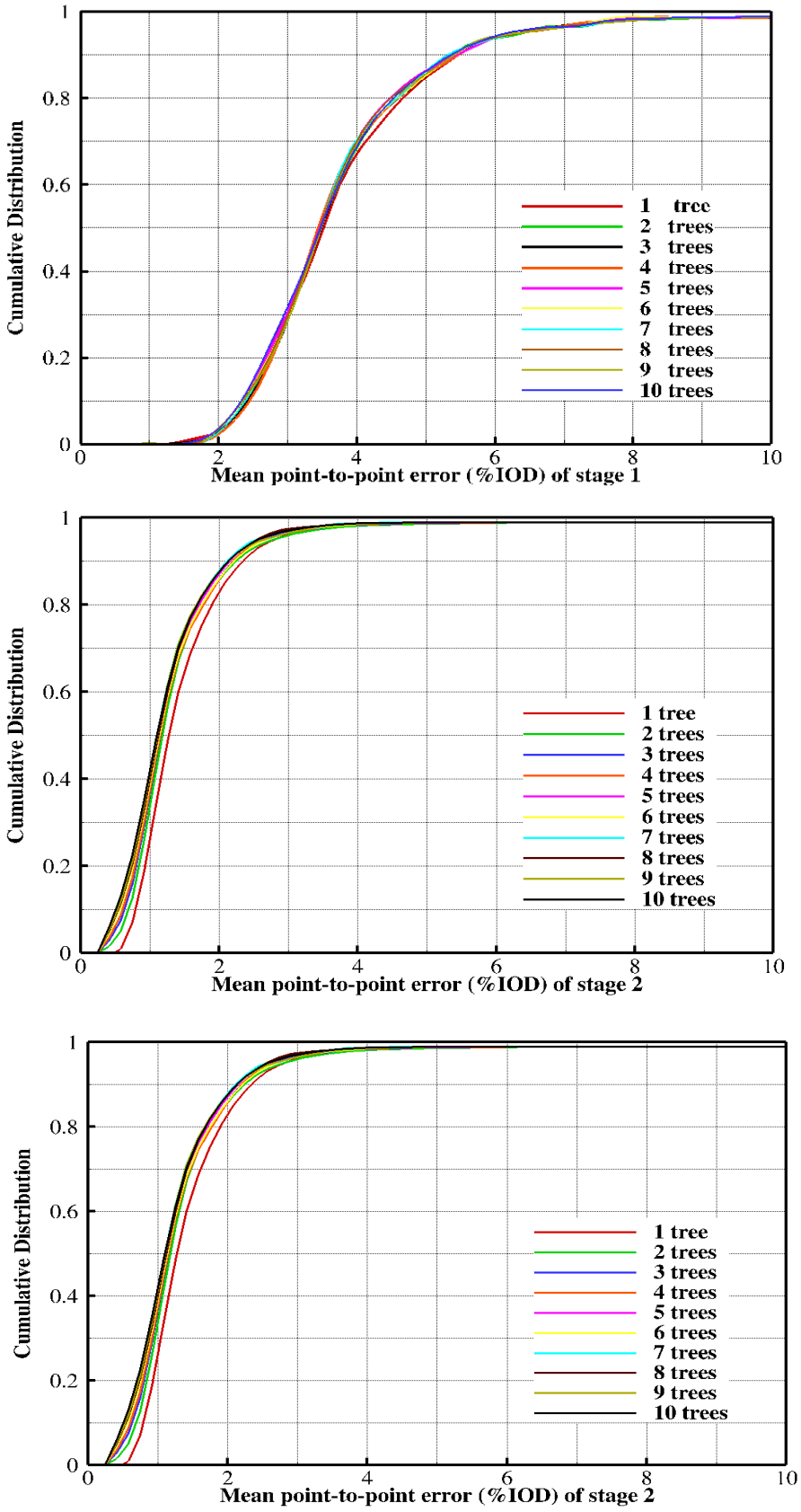


Fig. 5.8 Number of trees optimization results of three stages: (1-2) trees of stages one and two and 7 trees for stage three show sufficient performance. Stage-2 is initialized by stage-1's results. Stage-3 is initialized by stage-2's results. Performance is given as a point-to-point error as a percentage of the IOD.

Table 5.2 Optimal values for three stages RFRV-CLM parameters used in this thesis for facial expression localization (FEL)

Parameters	Stage-1	Stage-2	Stage-3
n_{trees}	1-2	1-2	7
n_{feat}	500	500	500
N_{min}	1	1	1
D_{max}	20	20	20
w_{frame}	30	60	120
w_{patch}	21	21	21
d_{max}	12	12	12
d_{search}	7	7	7

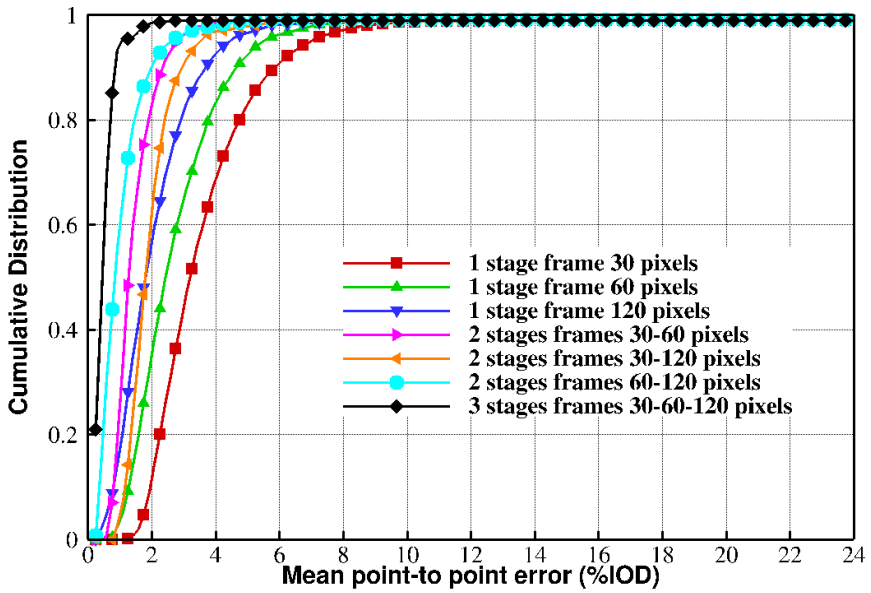


Fig. 5.9 Performance evaluation of the combination of using optimal stages parameters on FACES dataset.

5.5.1.2 Experiment 2 - Age Effect on Automatic Landmark Localization

Case Description: In the experiment described in Section 5.5.1.1, the proposed detector was optimized and evaluated with the presence of a large range of expression and age variabilities. The experiment in this section aims to investigate the effect that age patterns may have on the point localization performance. The aim here is not only to investigate the effect of the age pattern on the model performance but also to perform more validation for the optimal parameters of the proposed FEL.

The idea here is the presence of age features, which appear in an uncontrolled process under the effect of several internal and external factors such as health, lifestyle and weather, might hinder the process of placing the points around the target contour correctly and hence might lead to poor ability by the automatic detector to accurately detect the key points. Figure 5.10 highlights and compares the areas of the facial features where the consistent landmark placing across the object was impeded due to the deformation in the muscles of the face owing to the ageing effect which leads to a lack of locally distinctive structures. In the figure, the eyelid's appearance is different in young, middle, and old aged people in which a small part of the eye border with the middle age and a large part with the old age is hidden due to the sagging eyelids (see red circle in Figure 5.10). Further, the blue and black curves in the same figure illustrate the distortions in the mouth and chin areas due to the ageing.

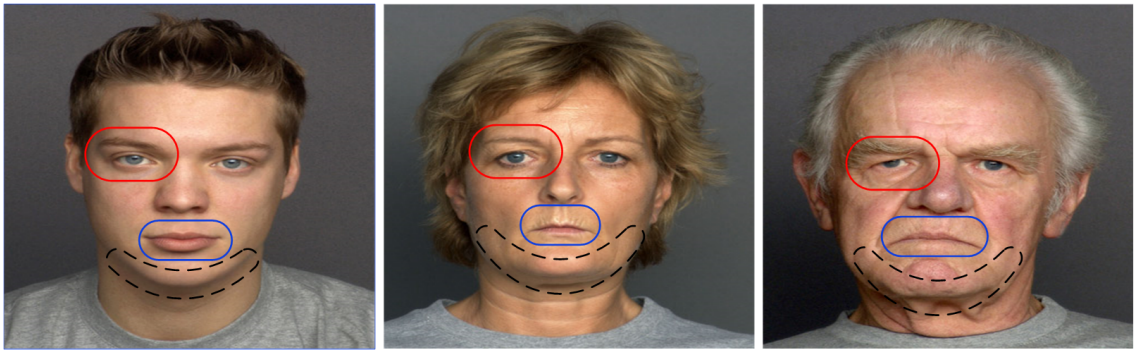


Fig. 5.10 Illustration of the effect of ageing patterns on the appearance of the contour of face components: young age (left), middle age (middle), and old age (right).

In this experiment, 2052 images of three age groups from the FACES dataset described in 3.2.1 are used. Subjects in each age group are split into training (50%) and testing (50%) sets. Three age-group specific RFRV-CLM-based FEL models were trained using the parameters from Table 5.2, one for each age group. The resulting models (young-group model, middle-group model, and old-group model) were then tested on each of the age group test sets. One age-agnostic RFRV-CLM-based FEL model was also built by combining the training sets from the three groups. The models were then tested by combining the test sets from the three groups.

Point-to-Point Results: Figure 5.11 (a), (b), and (c) summarises the results for each age-specific model. The results from the age-group-specific detectors show that for all models the errors increase as target age increases; each age-specific model works better on the younger test group than the older test group, suggesting that it is harder to locate features accurately on older faces than younger.

Figure 5.11 (d) compares the performance of a model trained on all ages (age-agnostic model) to the age-specific point detectors. It shows that the age-agnostic model works almost as well as the specific models for each age range, it is especially true for the middle

age data. This might be because the middle-age data contain some faces whose apparent age is younger than their chronological age, while some look older, so this age group has some of the characteristics of the younger and older age groups, making it somewhat similar to the agnostic group.

Furthermore, running the age-agnostic model on the old-age data set performs slightly better than the old-age model. This finding might be because not only does the age-agnostic model contain more shape information, enhancing its performance, but it might also be because some older faces' apparent ages are younger than their chronological age. Running the age-agnostic model on the young-age data set also performs slightly better than the young-age model. Again this might be because of a mismatch in the apparent ages.

We thus use the age-agnostic models for locating the points in subsequent experiments, as it avoids the requirement to know the target age. The age-agnostic model achieves an error below 3.4% (2.1 mm) on 99% of examples. Table 5.3 shows the statistics derived from Figure 5.11 of the fully automated system using a 3-stage RFRV-CLM .

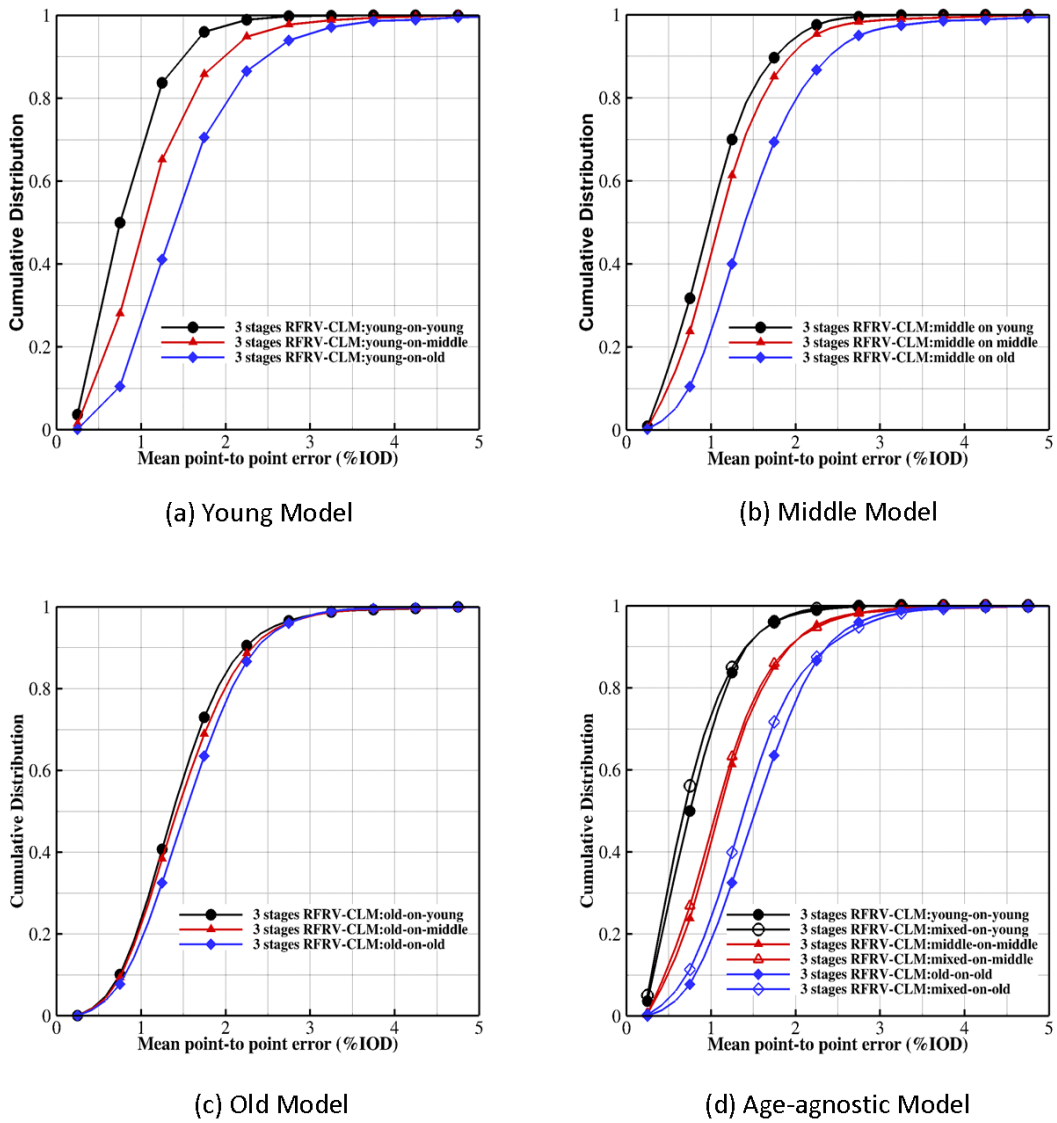


Fig. 5.11 CDFs of the mean point-to-point errors of the 76-point 3-stage RFRV-CLM age-group-specific detectors: (a) trained on the young age group data and tested on the young, middle, and old age groups data, (b) trained on the middle age group data and tested on the young, middle, and old age groups data, (c) trained on the old age group data and tested on all three age groups data, and (d) comparing age-group-specific detectors to the age-agnostic detector: **black lines** represent the young group error when tested with age-group-specific and age-agnostic models, **red lines** represent the middle group error when tested with age-group-specific and age-agnostic models, and **blue lines** represent the old group error when tested with age-group-specific and age-agnostic models.

Table 5.3 Statistics of the mean point-to-point errors between manual and automatic points detected using the FACES dataset derived from Figure 5.11.

Train	Test	Mean	Median	90%	95%	99%
Young	Young	1.45	0.836	1.58	1.82	2.58
Young	Middle	1.39	1.28	2.16	2.51	3.63
Young	Old	1.76	1.62	2.66	3.09	4.63
Middle	Young	1.3	1	1.69	2.29	2.79
Middle	Middle	1.44	1.35	1.97	2.47	3.52
Middle	Old	1.78	1.66	2.64	3	4.62
82 Old	Young	1.71	1.64	2.48	2.78	3.64
Old	Middle	1.76	1.68	2.56	2.88	3.64
Old	Old	1.82	1.77	2.62	2.9	3.54
Mixed	Young	1.18	0.763	1.43	2.1	2.62
Mixed	Middle	1.40	1.30	2.13	2.53	3.27
Mixed	Old	1.74	1.63	2.61	3.03	3.78
Mixed	Mixed	1.66	0.64	1.15	1.45	3.4

Overall Shape: Figure 5.12 shows the effect of varying the first four shape parameters of the age-group specific models and age-agnostic model with ± 3 standard deviations from the mean value, describing the global variations in the points and the difference across the age. The shape models have 53, 55, 56, and 59 modes for young, middle, old, and age-agnostic models respectively, which explain 98% of the variations in the landmarks' positions in the training set.

Quantitative Results: Figure 5.13 shows the quantitative results of the points' positions corresponding to the internal facial components (i.e. the eyes, eyebrows, nose, and mouth) and the face outline (i.e. the chin). These results show that: (i) The old data shows the maximum mean error of all points, considering the age-specific models. This might be because age-related structural changes in the face across the different age groups cause a difficulty in predicting the best position for each facial point. (ii) The chin points exhibit the maximum mean error on all of the groups. This result is likely due to both the poor local image texture of these points and the fact that the face outline is easily affected by variations in pose and age-related changes in structure. (iii) Very small mean errors less than 2% of the IOD are found around the eyes, nose, mouth and eyebrows. (iv) The age-agnostic model in some parts such as mouth, brows and chin of the face components performs as well as or better than some of the age-group-specific models.

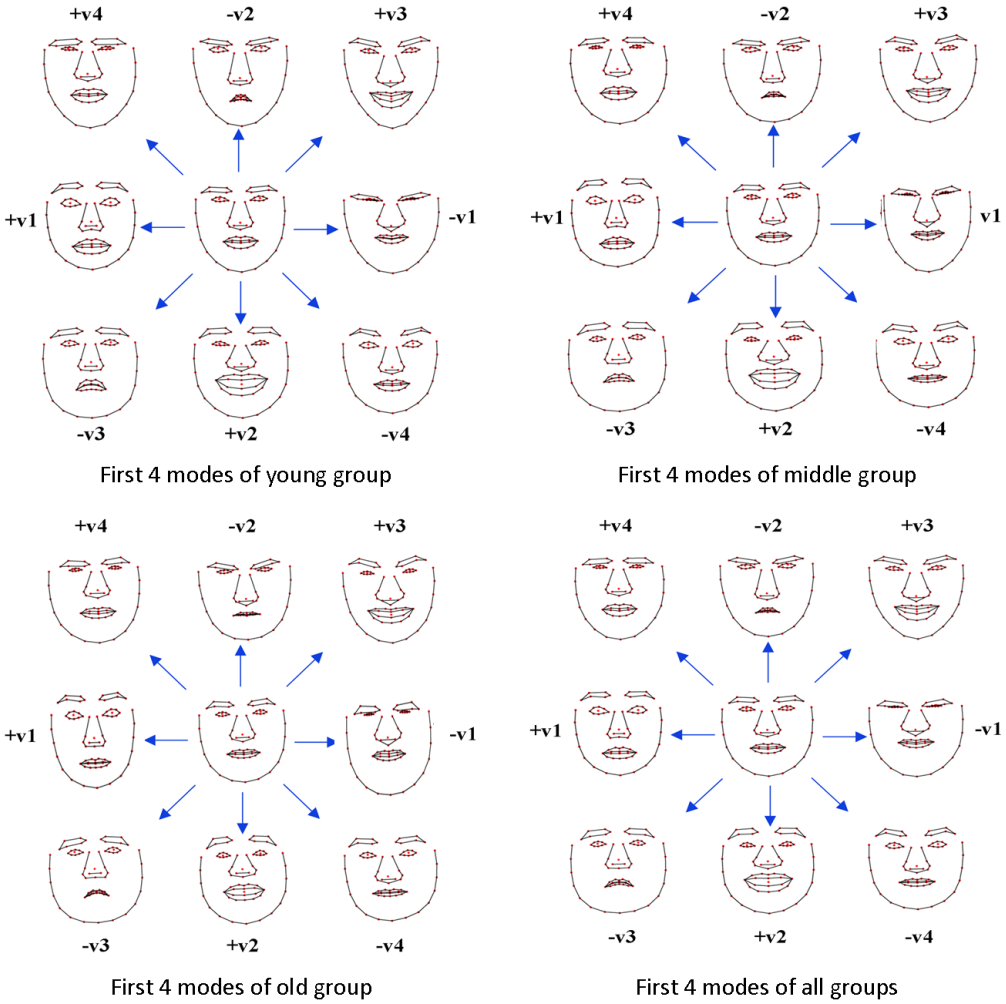


Fig. 5.12 Shape modes showing the shape changes across the age process.

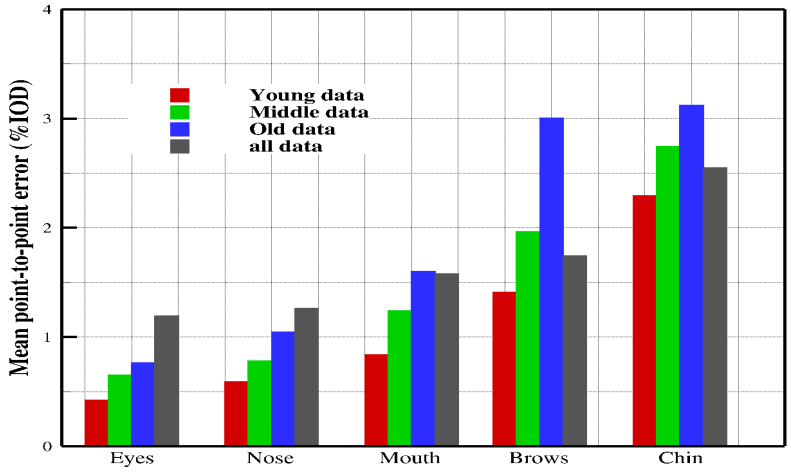


Fig. 5.13 Mean point-to-point errors between manual and automatic points of face components of FACES dataset.

5.5.1.3 Experiment 3 - Cross Data Evaluation (Transfer Learning)

Case Description: A key aspect of evaluating the performance of any learning model is to train and test the model using different datasets. The benefit here is to avoid over-fitting and to test the ability of the algorithm to generalize to new (unseen) data. In the previous experiments, the proposed FEL was trained using samples with a wide range of ages and expressions from the FACES dataset, and produced a 76-point model to use on the test data. The experiment in this section investigates the ability and generality of this model to adapt to new datasets: the LifeSpan and NEMO datasets of fake and spontaneous expressions. The LifeSpan and NEMO datasets are not labelled manually.

Results: The automatic localization results were very satisfactory on these datasets of posed and spontaneous expressions. Figure 5.23 shows examples of successfully and unsuccessfully automatic found points from samples of different ages from the LifeSpan and NEMO datasets.

5.5.1.4 Experiment 4 - Compound Emotions Effect on Landmark Location

Case Description: For further testing of the proposed facial landmark detector, in this experiment, the variation of 22-compound emotions is considered. This experiment is beneficial for more validation of the proposed detector along with the optimal parameters against a rich set of 6-basic and 15-non-basic expressions plus the neutral expression instead of only the 6-basic expressions in the previous experiments. The problem with the compound emotions is that most of the 22 emotion's shapes are partly similar since each emotion is generated from a combination of two basic emotions. For instance, Figure 5.14 shows the combination of happy and surprise expressions that generates a happily-surprise expression, in which the muscle's deformations are partly similar to both the happy and surprised expressions (see highlighted parts). These similarities lead to huge variations and it might hinder the process of the detector to accurately localize the fiducial points automatically and by consequence, the process of facial expression recognition will be hindered as well.

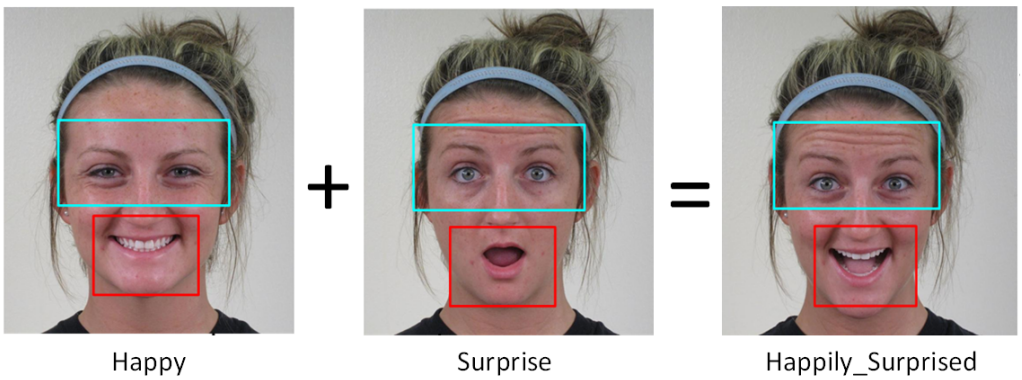


Fig. 5.14 Illustration of the similarities between the basic emotion and compound emotions.

In this experiment, 2,200 images from the compound emotion dataset annotated with 78-points were used and split into training (50%) and test (50%) sets. RFRV-CLMs using the parameter values from Table 5.2 were trained on the training set. The resulting model was tested on the test set and then the process was repeated with the test and training sets interchanged.

Point-to-Point Results: Figure 5.15 shows the CDF of the mean point-to-point error of different stages of the 78-point model computed after running different stages of the model from a range of perturbed positions on the set. The results show that the two stage model with 60-120 frame widths or the three stage model with 30-60-120 frame widths again gives the best accuracy of the RFRV-CLM fitting with a wide range of convergence across basic and non-basic expressions. The mean point-to-point error between the manual and automated points detected by the model is within 5.71% (3.6 mm) on 90% of the data, which gives a good indication of the robustness of the method when tested on a wide variation in the shape and appearance of 22 emotions. Figures 5.24, 5.25, and 5.26 show examples of successful and unsuccessful results of one person displaying 22 compound expressions. The failure was mostly around the mouth where the mouth features have disappeared owing in the extremely angry expression. Overall, the mouth corners are detected with less accuracy than other landmarks, as they are affected more under expression changes.

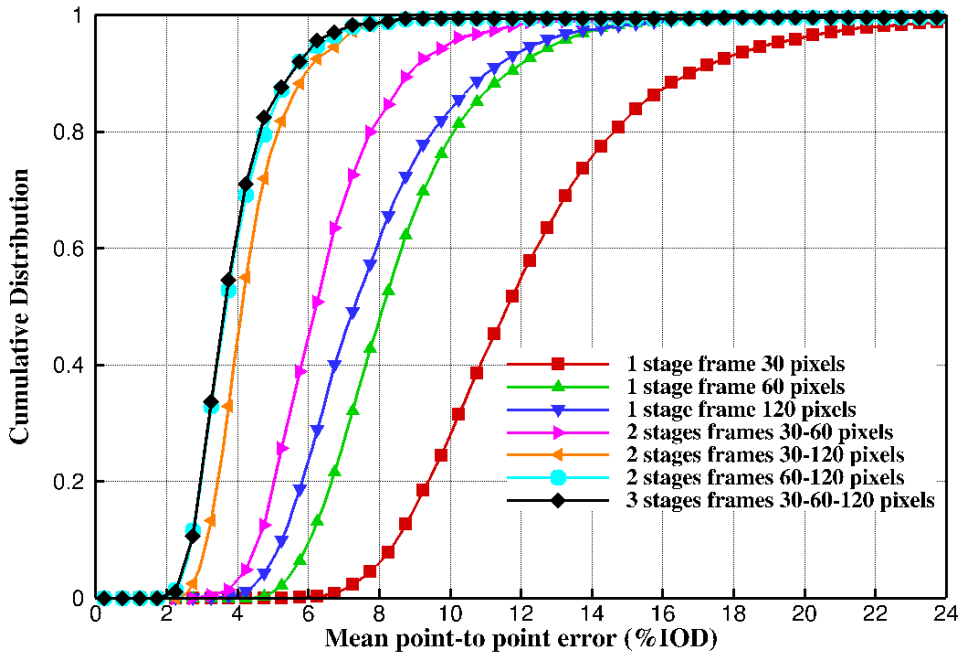


Fig. 5.15 Performance evaluation of the combination using optimal parameters using the compound emotions dataset.

Having proved the capability and the generality of the proposed model against 22-compound expressions, a further experiment was performed to investigate the effect of the compound expressions on the performance of facial features key points localization. We used 2,200 images: 700 images of the basic expressions set and 1,500 images of the compound expressions set. Each set is divided into (50%) training and (50%) testing (person independent), We trained three FEL models as follows: (i) trained on the basic expressions only and tested on the basic and compound data, (ii) trained on the compound expressions only and tested on the basic and compound expressions, and (iii) trained on the all the data and tested on basic and compound expressions.

The results from the three models are summarized in Figure 5.16 which shows the CDFs of the mean point-to-point error and Table 5.4 which shows the statistics derived from Figure 5.16. These results show that compound features have a significant effect on the point-localisation process and the errors are increased by including the compound emotions in the training set. For example, in the case of basic expression localization, the smallest error is obtained if the model is trained on basic data only with 3.32% (2.1 mm) mean error on 99% of the data and the error increase when including the compound data in the training set as shown in Table 5.4. In the case of compound expressions, despite smaller errors being obtained using the model trained on the compound data only with 7.12% (4.7 mm) mean error, we showed that using the model trained on the basic emotions only with 7.29% (4.6 mm) mean error is sufficient to generalize across 22 emotions. This might be because each compound expression is generated by combining two basic emotions. In summary, the results in experiment 4 indicated that training the FEL model using the basic expressions data only is sufficient to detect the facial features points of both basic and compound emotions.

Table 5.4 Statistics of the mean point-to-point errors between points detected manually and automatically in the compound emotions dataset.

Train	Test	Mean	Median	90%	95%	99%
Basic	Basic	**2.45**	**0.72**	**1.47**	**1.91**	**3.32**
Compound	Basic	2.48	0.72	1.41	1.85	4.57
Basic + Compound	Basic	2.71	0.91	2.08	2.54	4.29
Basic	Compound	3.94	3.67	5.44	5.97	7.29
Compound	Compound	**3.76**	**3.48**	**5.23**	**5.83**	**7.12**
Basic + Compound	Compound	4.17	3.94	5.47	6.37	7.79

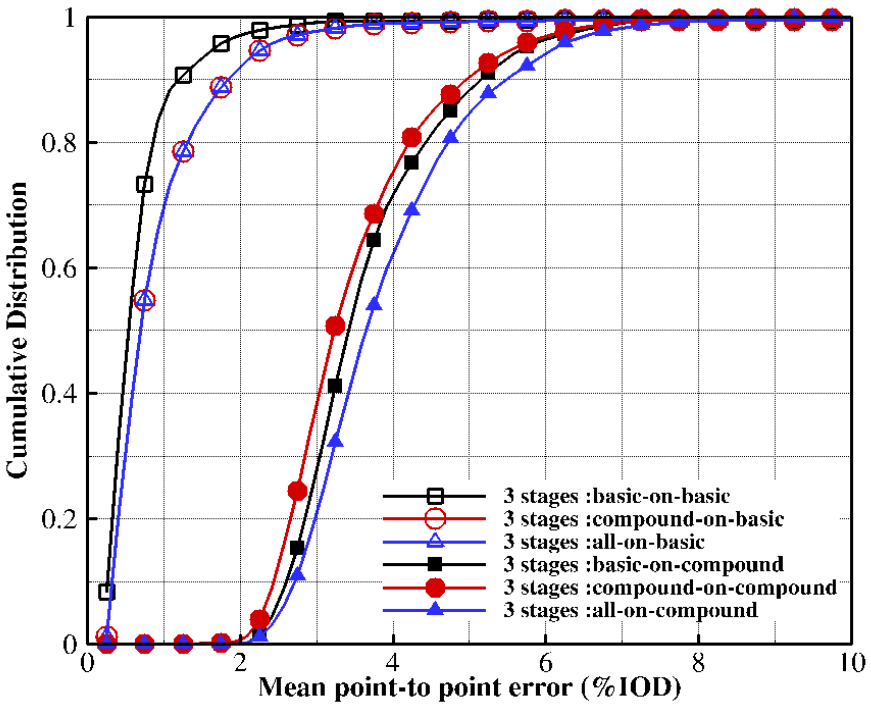


Fig. 5.16 CDFs of the mean point-to-point errors comparing the basic expressions model to the compound expressions model.

Overall Shape: The basic and compound shape models have 44 and 52 modes respectively, which explain 98% of the variation in the landmarks' positions in 700 images for the basic and 2200 images for the compound training set. Figure 5.17 shows the effect of varying the first four shape parameters with ± 3 standard deviations from the mean value of the compound and basic models, describing the global variations in the points and the difference between the basic and compound data.

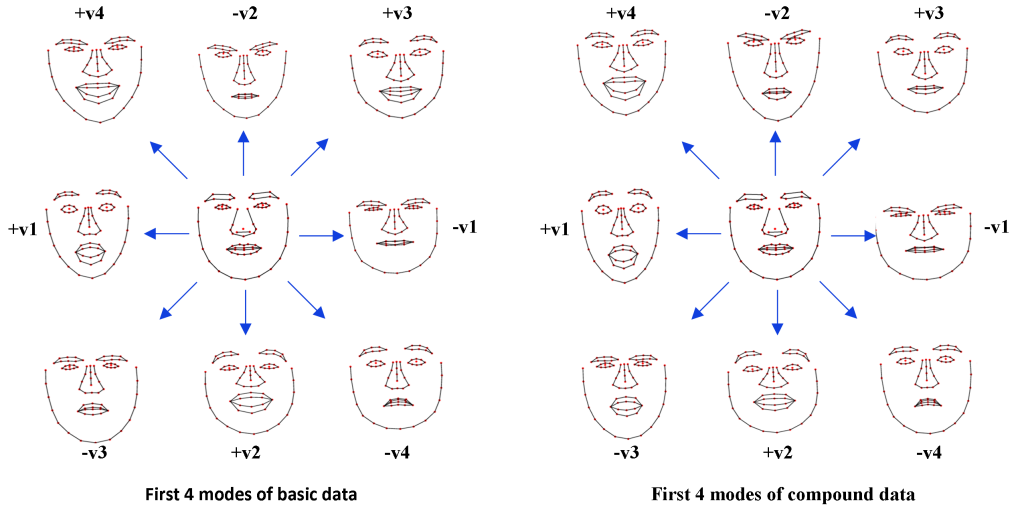


Fig. 5.17 Shape modes showing the difference in shape changes between basic and compound shape models.

Comparing to Other Methods: Table 5.5 compares our results on the compound emotions dataset to the results of alternative methods reported by Du et al. (2014) of using AAM (Cootes et al., 2001) AAM with RIK (Hamsici and Martinez, 2009), the Manifold approach (Rivera and Martinez, 2012), and the Pairwise optimization approach Du et al. (2014). These results show that RFRV-CLM outperforms the other techniques.

Table 5.5 Comparison among the results of the proposed FEL system and the results of the alternative methods tested on the compound emotions dataset.

Methods	Mean points-to-points error %
AAM (Cootes et al., 2001)	7.19
Manifold approach Du et al. (2014)	7.65
Active Appearance Model with RIK (Du et al., 2014)	6.34
Pairwise optimization approach (Du et al., 2014)	5.39
RFRV-CLM (Present work)	**4.60**

5.5.1.5 Experiment 5 - Expression's Intensity Effect on Landmark Location

Case Description: In the previous experiments, the proposed FFPD is optimized, tested, and validated against the variations of three age groups with six basic expressions (Experiments 1 and 2), across datasets (experiment 3), and 22 compound emotions (experiment 4). In this experiment, the variations of the intensities with which expressions are shown is considered. This consideration is beneficial for more validation for the presented FEL with the optimal parameters. The idea here is that the shape features for the expressions with different intensity are partly different. For instance, the contours of the open eye and mouth when showing surprise have different diameters as the intensity of the expression increases, as shown in Figure 5.18 Therefore, having an accurate and robust FEL to generalize across different intensities is necessary.

The experiment was performed using a set of 180 sequences from 62 subjects displaying 1-6 emotions of the CK+ dataset. Each sequence from the CK+ dataset began from a neutral expression and ended with the peak intensity of each expression, resulting in 3,164 images. Each frame of each sequence was annotated with 70 landmark points. RFRV-CLM using the optimal parameters described in table 5.2 was trained on half of the data, producing a 70-point model. The resulting model was tested on the test set (the other half of the data) and then the process was repeated with the test and training sets interchanged.

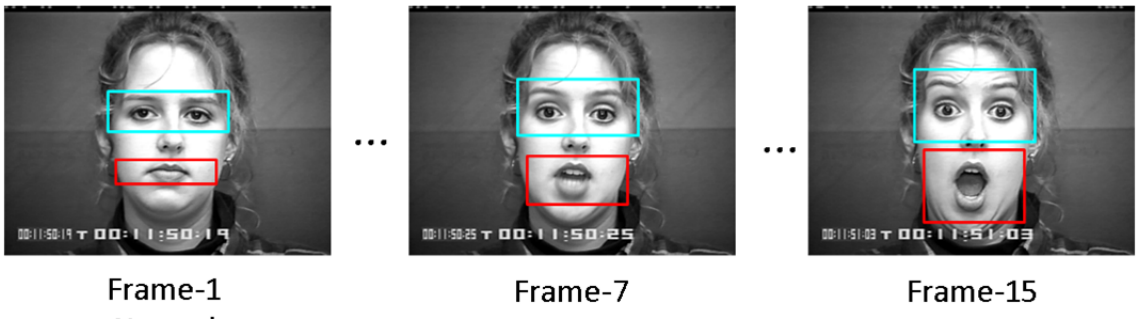


Fig. 5.18 Illustration of the changes in face components' shape of one person from CK+ dataset displaying the surprise expression with increasing intensities: ≈ 0% happy (left), ≈ 50% happy (middle), and ≈ 100% happy (right).

Point-to-Point Results: Figure 5.19 shows the CDF of the 70-point model computed after running different numbers of stages of the model from a range of perturbed positions on the set. Experiments and results, broadly speaking, show that applying three stages of 30-60-120 pixels frame widths of RFRV-CLM performs best. The results also show that the accuracy of the three stages of RFRV-CLM fitting has a wide range of convergence across different intensities of expressions. The mean point-to-point error between the manual and automated points for this data set was within 3.97% (2.5 mm) on 95% of the data, which means that the model range can successfully capture the facial-expression points at different intensities. Figures 5.27 and 5.28 show the performance on one video of a person showing the surprise emotion in increasing intensity.

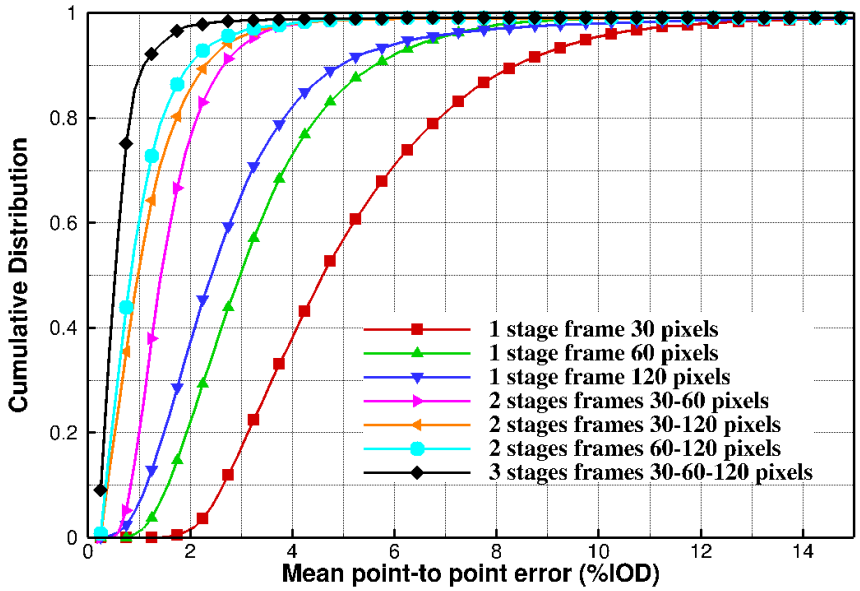


Fig. 5.19 Performance evaluation of combination using optimal stages parameters on CK+ dataset.

Overall Shape: The shape model has 31 modes, which explain 98% of the variation in the landmarks' positions. Figure 5.20 show the effect of varying the first four shape parameters within ±3 standard deviations from the mean value of the models.

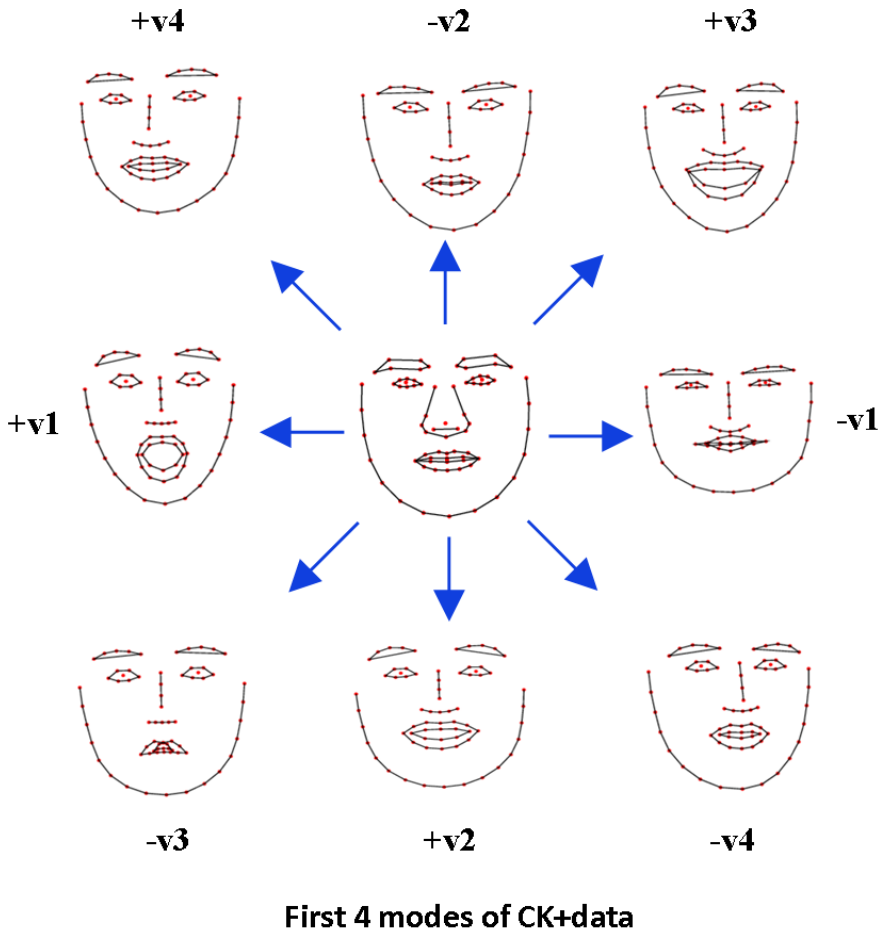


Fig. 5.20 Effect of varying each of the first four modes of CK+ data shape model parameters in turn between ±3 slandered deviation.

Comparison to other Methods: Table 5.6 compares our results to the results of alternative methods: using AAM (Cootes et al., 2001), AAM revisited (Matthews and Baker, 2004), Restricted Boltzmann Machines (RBM) using the Gaussian assumption (Wu et al., 2013), Restricted Boltzmann Machines (RBM) combined using Kernel Density Estimation (KDE) (Wu et al., 2013), and Gabor feature facial feature point detection: a comprehensive survey based boosted classifiers (Vukadinovic and Pantic, 2005) applied to the CK+ dataset. These results show that RFRV-CLM outperforms the other techniques.

Table 5.6 Experimental results of the proposed method compared to the alternative methods tested on the CK+ dataset.

Methods	Mean error %
AAM Cootes et al. (2001)	9.43
Active appearance models revisited(AAMs) (Matthews and Baker, 2004)	5.82
Gabor feature based boosted classifiers (Vukadinovic and Pantic, 2005)	7.00
RBM using Gaussian assumption (Wu et al., 2013)	4.83
RBM using KDE (Wu et al., 2013)	5.11
RFRV-CLM (Present work)	**3.19**

Summary: In summary, the empirical evaluation using the five different facial expression datasets demonstrated that a two stage model with 60-120 pixels frame width or a three stage model with 30-60-120 pixels frame width with the optimal parameters from Table 5.2 can give excellent performance for solving the problem of automatic facial expression feature detection across a wide variation in the age of the subject, 22 different emotions and the intensities of each expression.

Figure 5.21 summarizes the quantitative results of the internal facial components (i.e. eyes, brows, nose, mouth) and the face outline (external) (i.e. chin) on the three datasets used above. These results demonstrate that with the FACES dataset (red bars), the chin points exhibit the maximum mean error. This result is probably due to both the poor local image texture of these points and the fact that the face outline is easily affected by variations in pose and ageing. With the compound dataset (green bars), the mouth points exhibit the maximum mean error since they are subject to numerous variations of 22 different expressions. With the CK+ dataset, the mean error is approximately the same in all the face components since the capturing points started from the neutral state of the emotion and the intensity increased gradually in all the face points when the person showed expression.

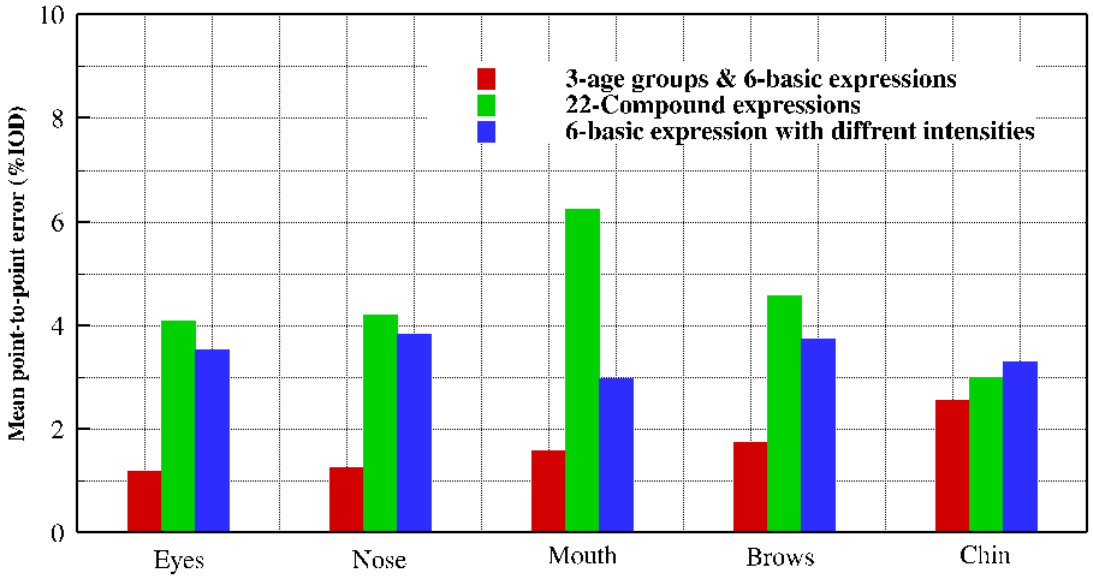


Fig. 5.21 Mean point-to-point errors between manual points and automatic points of the face components (eyes, nose, mouth, brows and chin) of FACES, compound and CK+ datasets.

The experiments (optimization and training) were performed using a 9,288 core cluster in The Manchester of University. For testing only a single processor core was used. The average time taken to annotate the 76 points across 1026 images was 157.5 millisecond for each image. Figures 5.22, 5.24 and 5.25, and 5.27 and 5.27 show examples of the manual annotation, response images and the automatic output points of five different facial expression data set of different ages, 22 expressions, and continuous intensity.

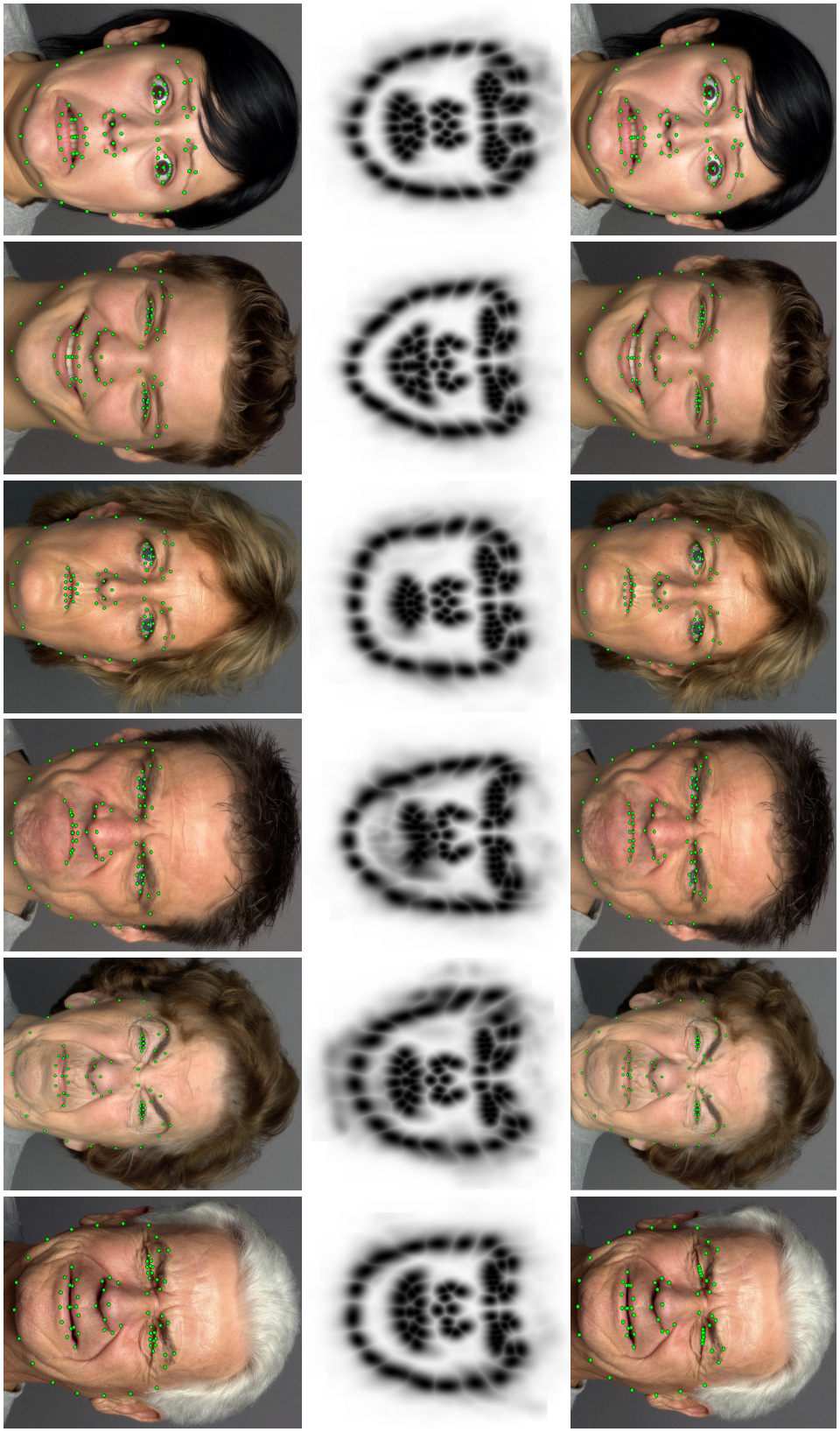


Fig. 5.22 Results of the proposed FEL on FACES dataset: manual points (first row), responses images (second row), and automatic points (third row).

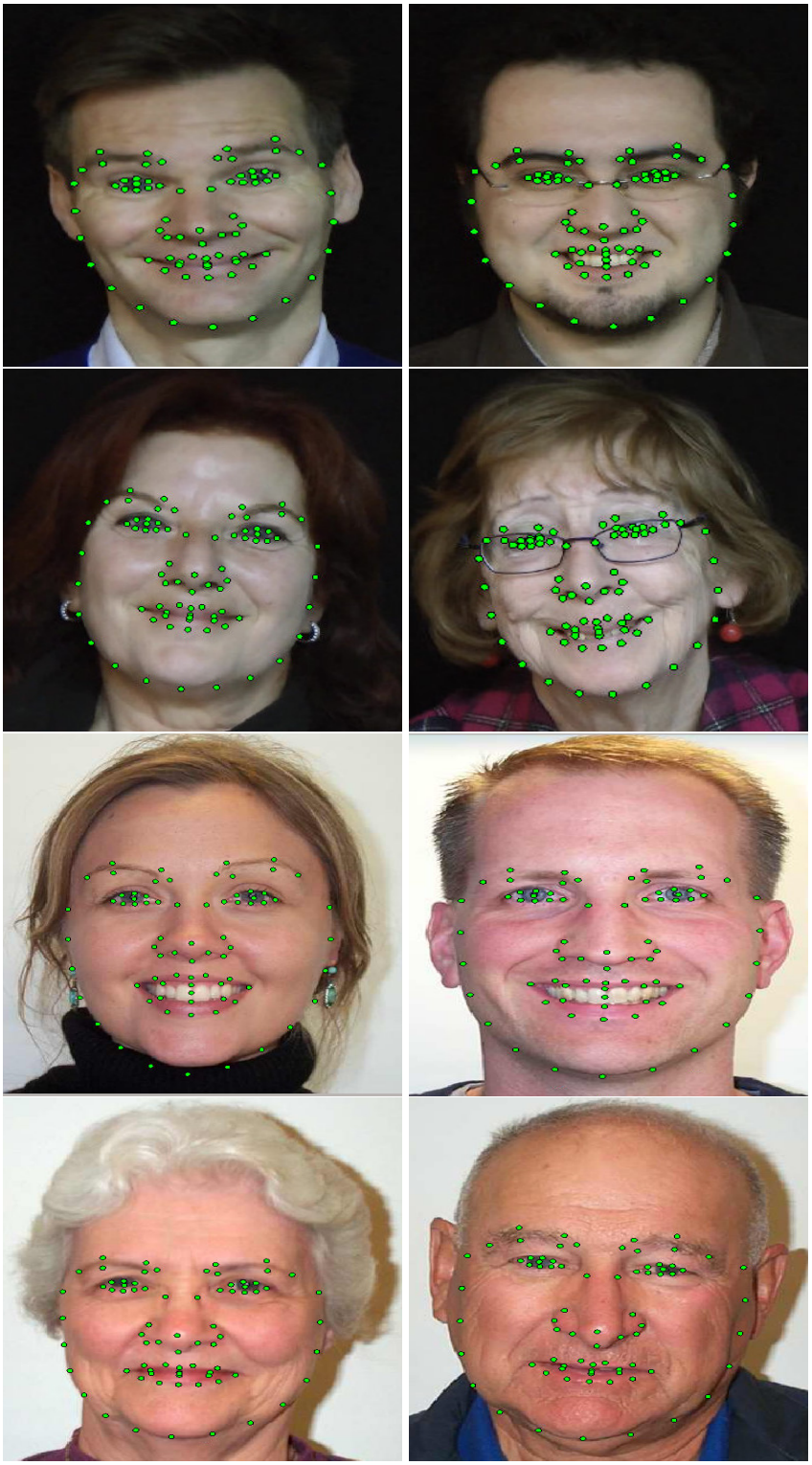


Fig. 5.23 Example results of the proposed FEL system trained using FACES dataset and tested on two ageing datasets: NEMO dataset of spontaneous expressions (top two rows) and LifeSpan dataset (bottom two rows).

Fig. 5.24 Example results of the proposed EFL detector on images of one subject from the compound emotion dataset: first row displays the manual points of the first 7 emotions. From left to right: happy, sad, fearful, angry, surprised, disgusted, happily surprised. Second row displays the response images. The third row shows the automatic points captured by the pre-trained model.

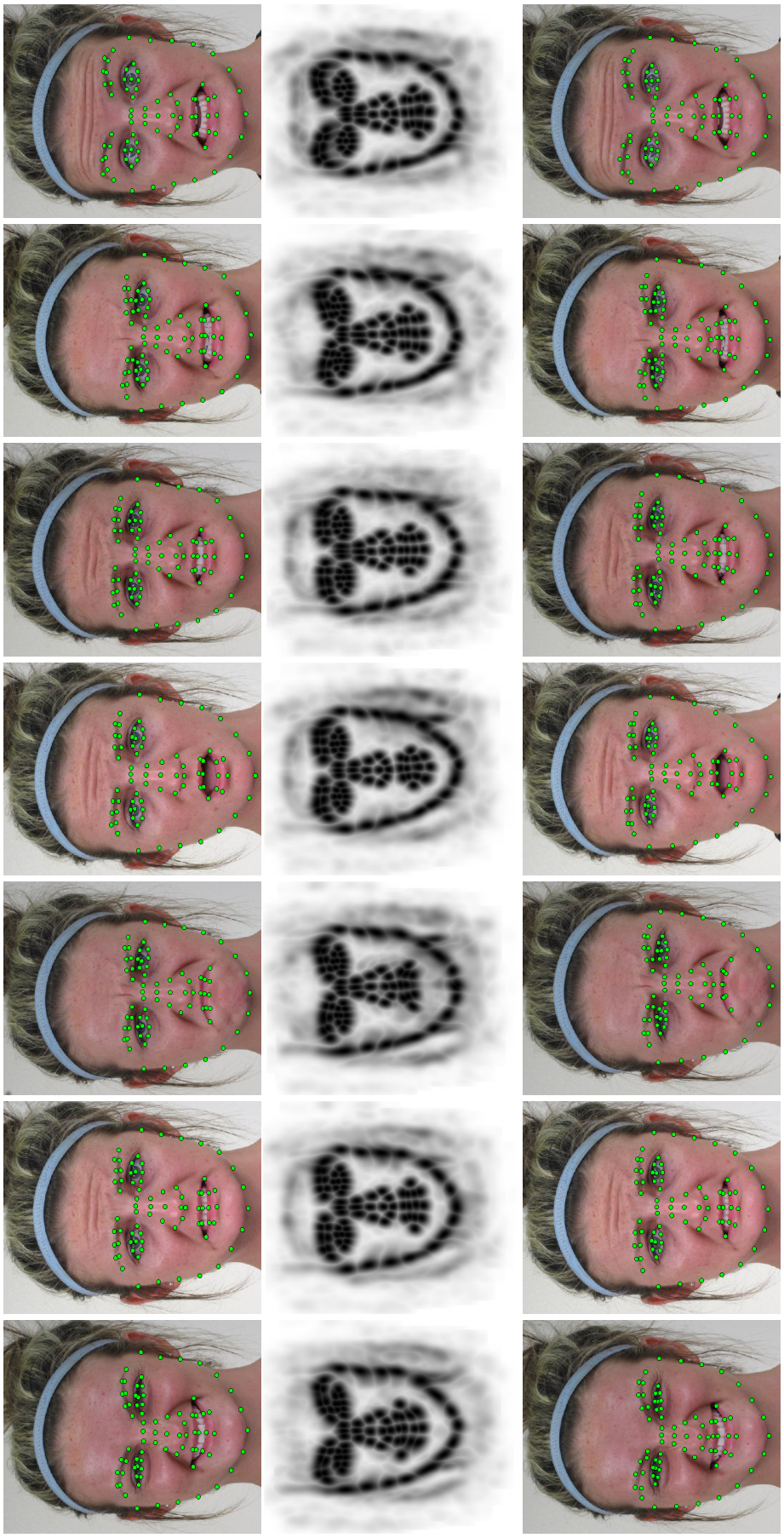


Fig. 5.25 First row displays the manual points for the second seven emotions of the same person from the previous figure: from left to right happily disgusted, sadly fearful, sadly angry, sadly surprised, sadly disgusted, fearfully angry, fearfully surprised respectively. Second row displays the response images for the emotions in the first row. The third row shows the automatic points captured by the pre trained model.

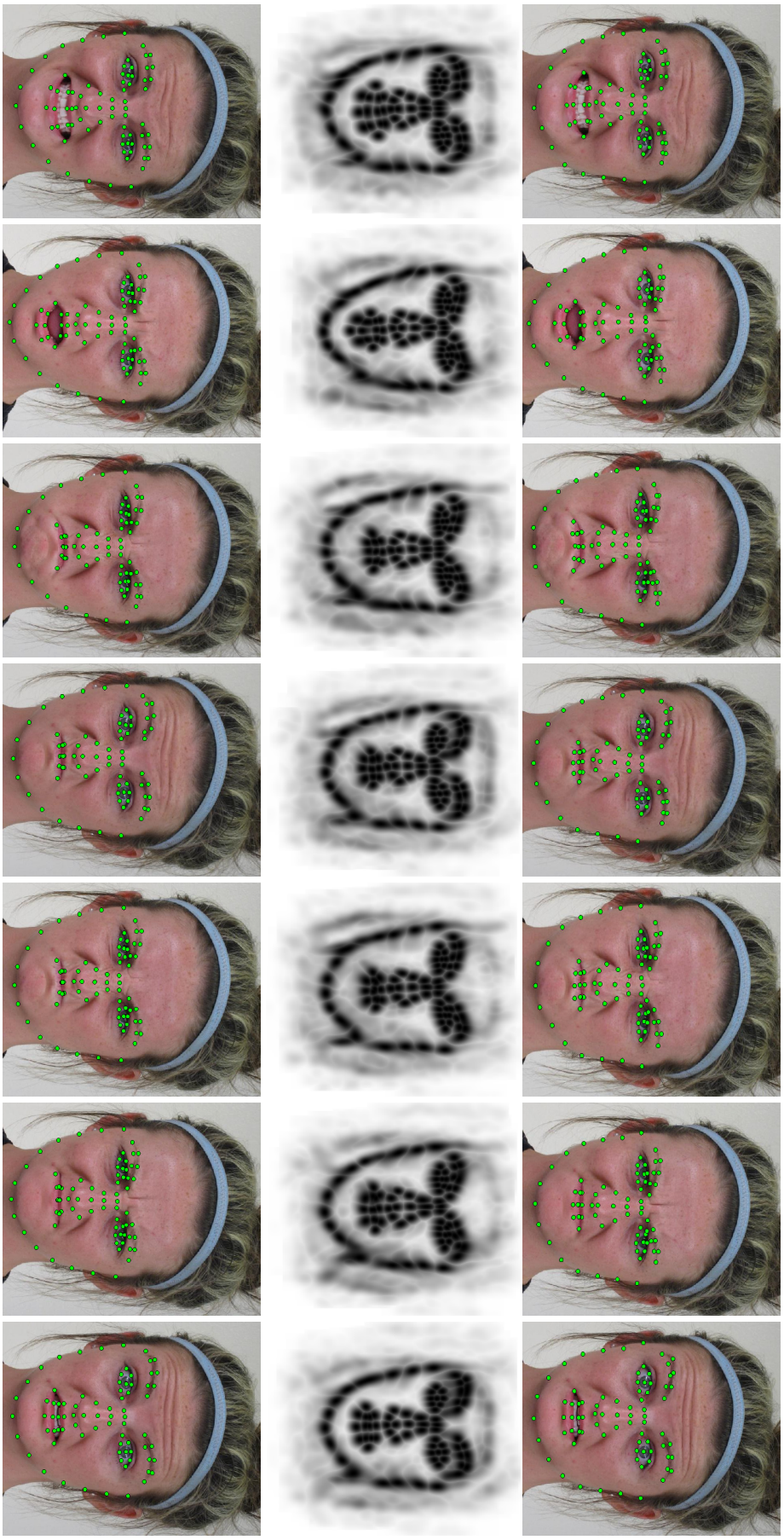


Fig. 5.26 First row displays the third seven emotions for the same person: from left to right fearfully disgusted, angrily surprised, angrily disgusted, disgustedly surprised, appalled, hatred, and awed. Second row displays the response images for the emotions. Third row shows the automatic points captured by the pre-trained model.

Fig. 5.27 Example results of the RFRV-CLM detector on images of one subject from the Cohn-Kanade data set displaying the happy emotion in increasing intensity from neutral to peak happy: first row is the first seven frames, the second row displays the responses images for the frames in the first row, and the third row shows the automatic points captured by the automatic model.

Fig. 5.28 Example results of the RFRV-CLM detector on images of one subject from the Cohn-Kanade data set displaying the surprise emotion in continuous intensity from neutral to peak surprise: first row is the second seven frames for the subject in Figure 5.27, second row displays the responses images for the frames in the first row, and third row shows the automatic points captured by the automatic model.

5.5.2 Evaluation of Automatic Facial Expression Classification

In this section, we use RF classifiers to measure both the impact of the automatically found points' accuracy on the performance score of AFER and the influence of ageing, compound emotions and intensity on the accuracy of expression recognition. The accuracy of each classifier was tested by comparing the recognition rate of recognizing the expression using features of the automatic-based points against the recognition rate of recognizing expression using the features of the manual-based points. For expression classification, both the average of all classes with the standard error and per-class (confusion matrix) classification accuracy between the ground truth label and the predicted label are reported. To avoid over-fitting and identity bias issues, 10 fold cross validation (person independent experiments) are applied. For validation purpose of the shape-based AFER, simple and complex cases of facial expression recognition are considered. Every image is represented using the shape features, texture features, and appearance features. The shape feature is obtained by building a statistical shape model (SSM) which provides a linear model of the distribution of a set of points in an image. It represents the shape of the relevant object as a linear sum of modes representing the main variations of the shapes. The training data set consists of a set of images I labelled with N landmark points, x_l, where $l = 1,...,N$.

The shape of the face can be encoded as

$$x = (x_1,...,x_n,y_1,...,y_n)^T \tag{5.7}$$

A statistical **shape model** Cootes et al. (1995)) is trained by applying Principal Component Analysis (PCA) to the aligned facial shape vectors, creating a model with the form:

$$x = \hat{x} + Pb \tag{5.8}$$

where x represents the shape vector in the reference frame, $\hat{x}$ represents the mean shape; P is a matrix of the set of eigenvectors corresponding to the highest eigenvalues, which describe different modes of variation; b is a set of parameter values of the shape model.

To obtain the texture features, we build a **texture model** Cootes et al. (2001). For each example in the training set we warp the face into a reference frame defined by the mean shape, then sample at regular positions to obtain a vector of intensities g. We normalise each vector then apply PCA to obtain a texture model of the form

$$g = \hat{g} + P_g b_g \tag{5.9}$$

Where b_g is a vector of weights of the modes P. The texture of a new example can be encoded as the vector b_g which best fits such a model to the intensities from the sample.

To extract the appearance features in which the correlation between shape and texture features are learned, an appearance model Cootes et al. (2001) is built by applying another PCA to the concatenation of the shape b and the texture b_g parameters. The concatenation is performed in a weighted form to compensate for the difference in units:

$$b_a = \begin{pmatrix} W_b \\ b_g \end{pmatrix} \tag{5.10}$$

where W is a diagonal matrix of weights for each shape parameters. W is chosen to balance the total variation in shape and texture,

$$W = \left(\frac{TotalVar(texture)}{TotalVar(shape)} \right)^{\frac{1}{2}} I \tag{5.11}$$

Applying PCA on the concatenated vectors gives the model:

$$b_a = p_c c \tag{5.12}$$

where p_c are the eigenvectors and c is the resulting parameter vector.

The shape, b, texture b_g, and appearance c are used as feature vectors from which a random forest classifier is trained to distinguish among expressions. Figure 5.29 summarizes the experiments performed in this section.

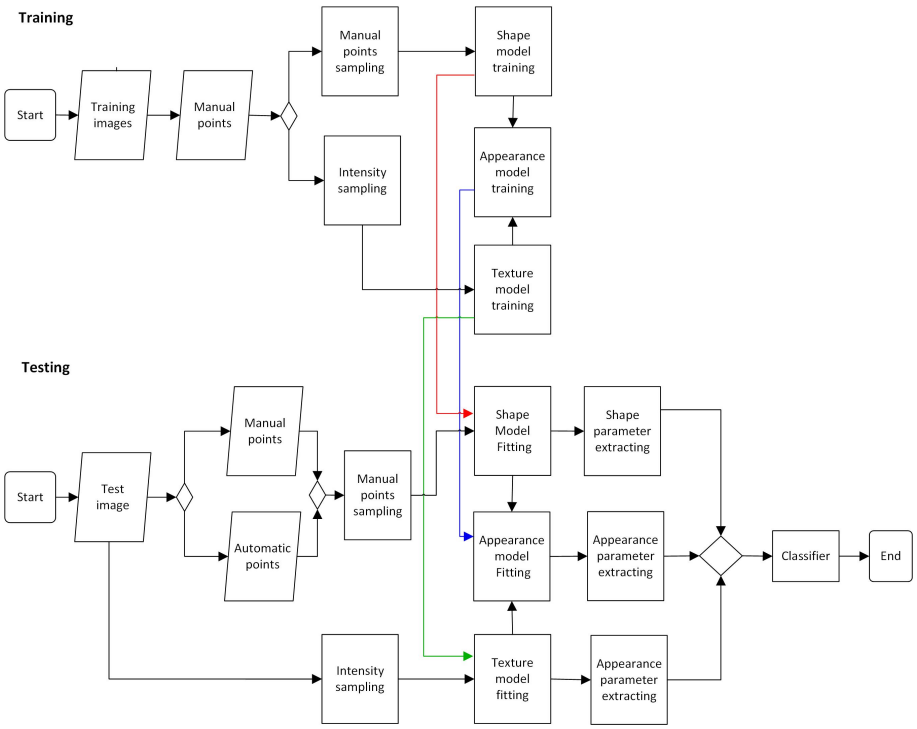


Fig. 5.29 Flow-chart giving an overview of the proposed experiments in this section. See main text for details.

5.5.2.1 Experiment 6 - Age Effect on Shape-Based AFER

Case Description: In this experiment, we use an RF classifier to study the influence of age on the performance of AFER and on the performance of the automatic points on the performance score of AFER. We trained shape, texture, and appearance models on half of the FACES data set using the manual points and then those models are used to extract the shape parameters b, the texture parameters b_g, or the appearance parameters c from the other half using both the manually located points and those from the automatic system and then swap for the other half. We then trained age-specific RFs to estimate the probability of each expression given a feature vector and age group (one for each age group data), and age-agnostic RF (one from all the data) to estimate the probability of each expression from all age groups.

Results The performance of the age-specific and age-agnostic classifiers, for each type of feature, on each age group is summarised in Table 5.7. This includes results where the landmark points were manually or automatically placed. This shows:

- Performance is best on the age-group for which the system was trained as described in the yellow, green, and blue cells for young, middle, and old groups respectively, and that performance degrades as the age difference increases (see the not highlighted cells of each age group.
- Performance on the older group (see blue cells) is worse than that on the young and middle groups (see yellow and green cells).
- Using the appearance features gives the best overall results (see bold fonts).
- When using the correct age-group classifier, there is only a small loss of performance when using features from automatically placed points, compared to those from the manually placed points (see the Man−Auto rows). Thus if we use the most appropriate age-specific classifier for each subject, we get improved performance, compared to using an age agnostic system.

Table 5.7 Expression classification results using manual and automated annotation for age-specific models and age-agnostic model. Shape, Tex, and App are the shape, texture, and appearance features respectively. Man−Auto is the difference between the manual-based and automatic-based performance.

			Test											
	Group	points	Young			Middle			Old			All		
			Shape	Tex	App	Shape	Tex	S&T	Shape	Tex	App	Shape	Tex	App
Train	Young	Manual	96.8	96.3	**97.7**	89.0	85.6	88.4	78.8	65.4	78.4	82.5	75.0	85.7
		Automatic	91.9	95.9	**96.5**	67.5	77.5	80.2	61.9	66.7	75.0	68.5	70.8	75.2
		Man−Auto	4.9	1.4	1.2	21.5	8.1	8.2	16.9	-1.3	3.4	14	4.2	10.5
	Middle	Manual	85.0	83.3	91.7	95.0	95.6	**97.6**	89.3	89.1	90.4	84.5	80.8	88.7
		Automatic	85.6	80.0	84.3	87.1	93.3	**94.7**	75.0	84.5	89.3	68.5	75.8	75.0
		Man−Auto	0.6	3.3	7.4	7.9	2.3	2.9	14.3	4.6	1.1	16	5.0	13.7
	Old	Manual	81.7	73.3	85.0	88.9	87.5	89.9	89.2	91.1	**93.8**	77.8	71.2	83.9
		Automatic	61.0	73.8	77.8	63.9	75.0	77.0	78.2	86.0	**88.5**	66.7	66.7	65.9
		Man−Auto	20.7	-0.5	7.2	25.0	12.5	12.9	11.0	5.1	5.3	11.1	4.5	18.0
	All	Manual	89.0	87.3	88.1	86.4	88.9	91.1	87.5	79.8	87.5	89.0	91.0	**93.4**
		Automatic	83.9	86.9	84.5	81.5	88.1	84.5	70.8	77.4	73.2	76.8	87.2	**90.1**
		Man−Auto	5.1	0.4	3.6	4.9	0.8	6.6	16.7	2.4	14.3	12.2	3.8	3.3

We further evaluate the effect of ageing on the system performance using a confusion matrix for automatic points and appearance features as described in Tables 5.8, 5.9, 5.10, and 5.11 for the young, middle and old groups, and age-agnostic classifiers in which the training and testing data have the same age limits. This shows that:

- performance for most expression categories is best on the young age group in which the system was trained with the expression variations only without the presence of the age pattern variations and that performance degrades as the age increases (see the diagonal elements).
- performance on the sad expression is worse than that on angry, disgusted, fear, happy, and neutral.
- performance on the sad expression is largely confused with the performance of other expressions and this confusion is increased as the age of the individual increases (see yellow, green, blue, and pink cells).

Table 5.8 Confusion matrix for expression classification of the young-group expressions classifier using appearance features.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	94.6	2.4	0.0	0.0	2.1	0.9
Disgust	1.5	97.0	0.0	0.0	0.0	1.5
Fear	0.0	0.0	97.0	0.3	0.3	2.4
Happy	0.0	0.0	0.0	100	0.0	0.0
Neutral	0.3	0.0	0.6	0.0	99.1	0.0
Sad	0.9	1.2	2.4	0.0	4.2	91.4

Table 5.9 Confusion matrix for expression classification of the middle-group expressions classifier using appearance features.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	94.0	4.8	0.0	0.0	0.3	0.9
Disgust	3.0	94.3	0.0	0.3	0.0	2.4
Fear	0.0	0.9	98.2	0.3	0.6	0.0
Happy	0.0	0.0	0.0	99.7	0.0	0.3
Neutral	0.0	0.0	0.9	0.0	97.9	1.2
Sad	3.6	2.4	2.1	0.3	7.4	84.2

Table 5.10 Confusion matrix for expression classification of the old-group expressions classifier using appearance features.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	88.1	8.0	0.0	0.3	1.5	2.1
Disgust	5.7	86.9	1.5	0.6	1.2	4.2
Fear	1.2	0.0	94.3	0.0	1.2	3.3
Happy	0.0	1.2	1.5	96.7	0.0	0.6
Neutral	0.0	0.3	0.3	0.6	96.7	2.1
Sad	8.3	8.0	5.4	0.9	9.2	68.2

Table 5.11 Confusion matrix for expression classification of the age-agnostic expressions classifier using appearance features.

Data	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	89.0	6.8	0.3	0.0	1.2	2.7
Disgust	5.4	90.5	1.2	0.3	0.9	1.8
Fear	0.9	0.0	95.2	0.3	0.9	2.7
Happy	0.0	0.6	2.1	96.7	0.0	0.6
Neutral	0.0	0.6	0.3	0.3	96.7	2.1
Sad	7.4	6.5	3.9	1.5	8.6	72.0

Table 5.12 shows a summary of the results of automatic expression classification rates presented in experiment 6. Going from the worst to the best performance of the classifiers, it can be seen that the overall worst performance is provided by the across age classifier. This is probably because training using one age group of the data cannot accommodate all the expected variation from the other groups. The second best performance is provided by the age-agnostic classifier. Finally, the best overall performance is provided by age-specific classifiers where the training and the testing data are from the same and limited age ranges. The results are then compared to the results of alternative approaches using the same datasets and with the results of using textures features presented in chapter 4. The comparison demonstrates that the best performance of within age group is obtained by Guo et al. (2013) with 97.85% followed by the age-agnostic model presented in this chapter with 93.4% and finally the across age group of the features presented in this chapter by 87.3%

Table 5.12 Comparison to previous work on FACES dataset.

	Within age group	Across Age Group	Age-agnostic
Guo et al. (2013)	97.85	64.04	88.80
Lou et al. (2018)	90.05	-	92.1
This thesis (Chapter 4)	93.1	73.69	90.1
Manual (Chapter 5)	96.4	87.3	93.4
Automatic (Chapter 5)	93.2	80.6	90.1

5.5.2.2 Experiment 7 - Compound Emotions Effect on Shape-Based AFER

Case Description: In this experiment, we use an RF classifier to measure the influence of 22 compound emotions on the performance of AFER and the performance of the automatic points on the performance score of AFER. We trained shape, texture, and appearance models on half of the compound data set using the manual points and then those models are used to extract the shape parameters b, the texture parameters b_g, or the appearance parameters c from the other half using both the manually located points and those from the automatic system and then swap for the other half. We then trained three expression classifiers: 6-basic, 15-non-basic, and 22-compound emotions classifiers to estimate the probability of each expression given a feature vector.

Expression Classification Results: The performances of the three classifiers are summarised in Table 5.13. This includes results where the landmark points were manually or automatically placed. This shows:

- Performance is best on the basic emotions, and degrades with non-basic and compound emotions due to the confusion in the expression classification because of the partial similarity between the basic and non-basic emotions as shown Tables 5.14, 5.15, and 5.16.
- The difference between the performance of manual-based points and automatic-based points is less when using texture features only, compared to that of using shape features only.
- We noticed that using shape feature alone led to better accuracy than using texture features only and best results are obtained by using the combination of shape and texture features. .

Table 5.13 Expression classification accuracy for manual and automated annotations of compound dataset. Shape, Tex, and App are the shape, texture, and appearance features respectively

Classifier	points	6-basic Emotions	15-non-basic Emotions	22-compound Emotions
Shape	Manual	93.2 ±0.2	60.9 ±0.1	56.7 ±4
Shape	Automatic	90.3 ±0.4	54.3 ±1.2	46.0 ±3.1
Tex	Manual	89.3 ±0.4	53.3 ±0.2	46.5 ±3.7
Tex	Automatic	88.9 ±0.6	51.9 ±1.6	44.1 ±5.8
App	Manual	91.6 ±0.3	56.5 ±1.5	51.8 ±4.6
App	Automatic	**92.3 ±0.5**	**59 ± 2.1**	**55.6 ±4.9**

Tables 5.14, 5.15, and 5.16 illustrate confusion matrices of basic, non-basic, and compound emotions classifiers respectively based on automatic initialisation using appearance features. These results show that, most of the expressions were confused owing to the similarities between AUs of basic and compound emotions. The results in Tables 5.15 and 5.16 indicate that most errors made by the compound emotion classifier are consistent with the similarity in AUs activation between the confused expressions. For instance, in Table 5.16, 34.0% (red cell) of sadly disgusted (row l) were confused with angrily surprised (column q) because three AUs are shared between them; 31.0% (yellow cell) of fearfully surprised (row o) were confused with awed (column v) and 33.7% (green cell) of angrily surprised (row q) were confused with sadly surprised (column l). Although some compound emotions share the same AU, their categories are distinct, which led to a very good recognition rate. For instance, many AUs of happily disgusted (row i) are the same as those of disgust (row g), but the confusion between them is approximately zero (i.e. 0.0%) (blue cell) of disgusted (row g). In summary, the results in Table 5.16 are consistent to some extent with the similarities between the AUs described in Table 2.3.

Table 5.14 Confusion matrix and accuracy of the seven basic emotions (a:neutral, b:happy, c:sad, d:fearful, e:angry, f:surprised, g:disgusted) using shape features.

	a	b	c	d	e	f	g
a	99.0	0.0	0.0	0.0	0.0	1.0	0.0
b	0.0	99.0	0.0	1.0	0.0	0.0	0.0
c	5.0	2.0	87.0	0.0	4.0	1.0	1.0
d	2.0	0.0	3.0	85.0	2.0	6.0	2.0
e	0.0	0.0	1.0	0.0	90.0	1.0	8.0
f	0.0	0.0	0.0	0.0	0.0	100.0	0.0
g	0.0	1.0	3.0	4.0	6.0	0.0	86.0
Mean	92.3						

Table 5.15 Confusion matrix and accuracy of fifteen non-basic emotions: (h:happily surprised, i:happily disgusted, j:sadly fearful, k:sadly angry, l:sadly surprised, m:sadly disgusted, n:fearfully angry, o:fearfully surprised, p:fearfully disgusted, q:angrily surprised, r:angrily disgusted, s:disgustedly surprised, t:appalled, u:hate, v:awed) using shape features.

	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
h	88.0	3.0	1.0	0.0	0.0	0.0	1.0	5.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
i	5.0	91.0	0.0	0.0	0.0	1.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
j	5.0	2.0	29.0	6.0	7.0	4.0	16.0	10.0	6.0	0.0	2.0	3.0	0.0	1.0	9.0
k)	0.0	0.0	1.0	81.0	1.0	1.0	2.0	0.0	0.0	0.0	4.0	0.0	2.0	7.0	1.0
l	0.0	0.0	3.0	0.0	64.0	1.0	0.0	1.0	1.0	23.0	0.0	1.0	0.0	2.0	4.0
m	0.0	2.0	7.0	9.0	0.0	32.0	11.0	2.0	7.0	1.0	9.0	8.0	6.0	6.0	0.0
n	0.0	6.0	14.0	3.0	4.0	6.0	52.0	1.0	6.0	4.0	0.0	0.0	0.0	3.0	1.0
o	13.0	0.0	6.0	0.0	1.0	0.0	3.0	47.0	8.0	2.0	0.0	1.0	0.0	0.0	19.0
p	8.0	9.0	8.0	0.0	0.0	3.0	13.0	6.0	44.0	2.0	0.0	2.0	0.0	1.0	4.0
q	0.0	0.0	0.0	1.0	27.0	0.0	3.0	0.0	1.0	65.0	0.0	0.0	0.0	2.0	1.0
r	0.0	0.0	1.0	9.0	0.0	5.0	5.0	0.0	0.0	0.0	58.0	4.0	11.0	7.0	0.0
s	0.0	1.0	2.0	2.0	3.0	3.0	0.0	1.0	2.0	0.0	2.0	71.0	4.0	2.0	7.0
t	0.0	1.0	0.0	4.0	0.0	5.0	4.0	0.0	0.0	3.0	9.0	4.0	48.0	22.0	0.0
u	0.0	0.0	1.0	10.0	1.0	3.0	2.0	0.0	0.0	2.0	7.0	0.0	21.0	53.0	0.0
v	3.0	0.0	2.0	0.0	10.0	0.0	4.0	13.0	1.0	1.0	0.0	4.0	0.0	0.0	62.0
Mean	59														

Table 5.16 Confusion matrix and accuracy of 22 emotions: (a:neutral, b:happy, c:sad, d:fearful, e:angry, f:surprised, g:disgusted, h:happily surprised, i:happily disgusted, j:sadly fearful, k:sadly angry, l:sadly surprised, m:sadly disgusted, n:fearfully angry, o:fearfully surprised, p:fearfully disgusted, q:angrily surprised, r:angrily disgusted, s:disgustedly surprised, t:appalled, u:hate, v:awed) using combined features.

classes	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
a	80.3	0.0	5.3	0.0	1.0	1.0	1.3	0.0	0.0	0.0	0.3	1.0	0.0	0.0	1.0	0.0	0.0	0.3	2.7	1.0	2.0	2.7
b	0.0	87.7	0.0	0.0	0.0	0.0	0.0	5.7	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
c	8.3	2.0	55.3	0.7	3.7	0.0	1.0	0.0	1.0	0.7	14.7	1.0	3.7	0.3	1.0	0.0	0.3	0.7	1.0	0.0	3.7	1.0
D	2.7	1.0	2.0	55.7	2.7	9.7	2.0	1.7	0.3	4.7	0.0	0.3	1.3	4.3	7.3	1.7	0.7	0.0	1.0	0.0	0.0	1.0
e	1.0	0.0	0.0	0.3	77.7	0.3	9.0	0.0	0.0	0.0	2.3	0.3	0.0	0.0	0.0	0.0	1.3	0.0	1.0	1.0	5.7	0.0
f	0.7	0.0	0.0	1.7	0.0	92.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	2.0
g	0.3	1.7	2.3	1.7	11.0	0.0	64.0	0.0	1.7	0.3	0.3	0.0	0.3	1.0	1.0	0.3	1.0	1.7	3.3	6.7	1.3	0.0
h	0.0	2.3	0.0	1.0	0.0	0.0	0.0	88.7	2.0	0.3	0.0	0.0	0.0	0.0	4.7	0.3	0.0	0.0	0.0	0.0	0.0	0.7
i	0.0	5.7	0.7	0.0	0.0	0.0	0.0	4.7	84.0	0.3	0.3	0.0	0.0	2.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0
j	2.0	1.7	5.7	5.0	1.3	0.0	0.0	3.0	0.0	16.7	2.3	4.7	5.7	20.3	10.3	8.3	0.7	0.7	1.3	0.0	1.3	9.0
k	1.0	0.0	15.7	0.0	2.7	0.0	0.0	0.0	0.0	0.0	60.7	0.3	1.3	0.0	0.0	0.0	1.0	8.7	0.0	1.7	6.3	0.7
l	0.0	0.0	0.0	1.0	0.0	10.7	0.0	0.0	0.0	0.3	0.0	50.0	0.3	0.0	0.3	0.0	34.0	0.7	0.0	0.0	2.0	0.7
m	0.7	1.3	8.0	0.3	1.0	0.0	4.3	0.0	0.7	10.3	7.0	0.0	19.7	9.7	0.7	10.7	0.0	6.3	7.7	7.3	4.3	0.0
n	1.0	2.0	1.0	1.3	1.3	0.0	1.7	0.7	4.3	18.0	1.7	2.0	5.0	42.3	0.7	9.0	4.7	0.7	0.0	0.3	1.7	0.7
o	0.0	0.7	0.0	3.7	0.0	2.0	0.0	7.7	0.0	6.3	0.0	1.3	0.0	2.7	37.7	6.0	1.0	0.0	0.0	0.0	0.0	31.0
p	1.0	5.3	0.7	1.0	0.3	0.0	0.0	4.7	6.0	9.3	0.0	0.0	1.7	16.3	11.3	36.0	1.7	0.7	2.3	0.0	0.3	1.3
q	0.0	0.0	0.3	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.3	33.7	1.0	0.7	0.0	0.0	58.7	0.0	0.0	0.0	2.3	0.0
r	1.7	0.0	1.7	0.0	2.7	0.0	6.3	0.0	0.0	0.0	13.3	0.3	2.7	2.0	0.0	0.0	0.0	40.7	0.7	20.3	7.7	0.0
s	3.3	2.0	4.0	0.7	0.0	2.7	5.3	0.0	0.0	0.3	0.0	0.0	2.3	0.0	3.0	1.7	0.0	0.3	63.7	1.0	2.0	7.7
t	0.3	0.0	1.7	0.0	3.3	0.0	4.3	0.0	0.3	0.0	2.3	0.0	3.3	2.0	0.3	0.3	1.7	9.0	1.7	37.3	32.0	0.0
u	2.0	0.0	1.3	0.0	13.0	0.0	1.7	0.0	0.0	0.0	6.0	1.0	1.0	1.0	0.0	0.0	1.3	3.7	0.0	35.3	32.7	0.0
v	3.0	0.0	1.0	0.7	0.0	5.7	0.0	2.7	0.0	5.0	0.0	7.3	0.0	1.3	26.7	0.3	1.3	0.0	2.7	0.0	0.0	42.3
Mean	55.6																					

5.5.2.3 Experiment 8 - Expression's Intensity Effect on Shape-Based AFER

In this experiment, we trained shape, texture, and appearance models using the entire CK+ dataset and the corresponding automatic points. These models are then used to extract the shape, texture, and appearance features from every frame in each video. We then assessed the performance of those features in distinguishing among emotions. The process of labelling training data of different expressions with different intensities is a time-consuming, expensive task prone to mistakes, possibly leading to unreliable labels. Here, we have only visualised the first three components for every frame in order to investigate the initial performance and the ability of those features in separating between the emotions in order to consider approaches that do not require manual emotion labelling, such as building an unsupervised system.

Figure 5.30 shows 2D and 3D description models of shape, texture, and appearance features based on automatic initialisation extracted from six videos of one person. Each video shows one of the following expressions: angry, disgusted, fear, happy, sad or surprise. Each expression starts from a neutral expression and ends with an expression at high intensity. In the figure, each path of points represents a video of multiple frames and each point represents a frame. These results show that shape features (top row) give the best performance of a smooth path of each video with a very clear separation among the six expressions with a good separation among the frames in the same video (same path). This outcome demonstrates that extracting the features in this way gives encouraging results for automatic expression classification based on intensity. Consequently, based on the results in Figure 5.30 and given that only a simple visualisation gave a clear separation among the expressions, we believe that building an unsupervised system to recognize the expression based on its intensity and based on the extracted features is a promising option. Therefore, it is possible for an unsupervised system to perform recognition functions of a similar standard to a supervised facial-expression-recognition system.

Figure 5.30 shows 2D and 3D description models of shape texture, and appearance features based on automatic initialisation extracted from many persons (all the data). These results show again the best results are achieved using the shape features and including the texture features might not improve the performance. Although the frames in the middle are overlapped due the neutral frames, the results are encouraging and warrant more investigation.

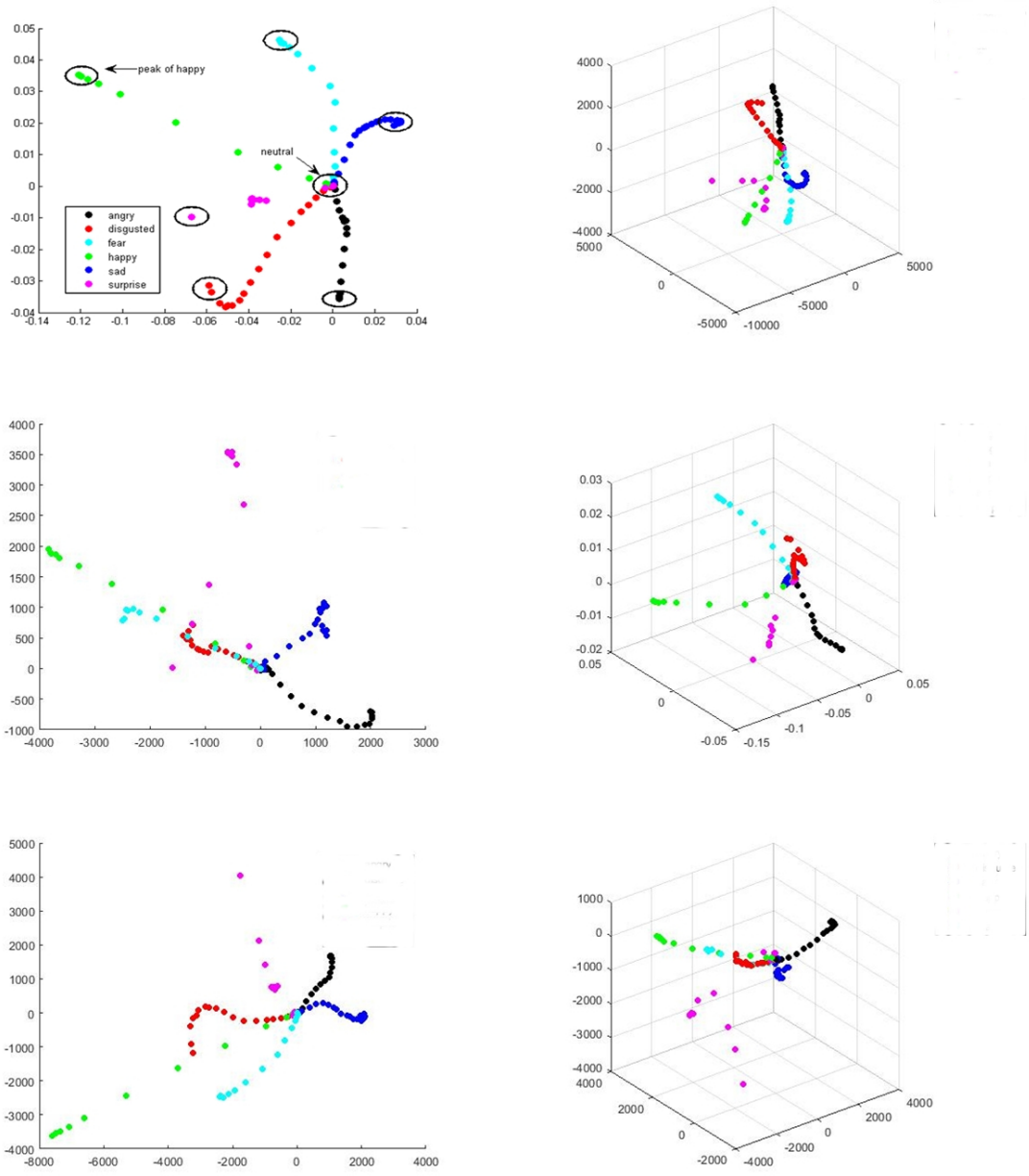


Fig. 5.30 Illustration of a 2D (left) and 3D(right) description models using shape (top row), texture (middle row), and appearance (bottom row) features of one person showing six expressions of different intensities. Peak intensities are circled.

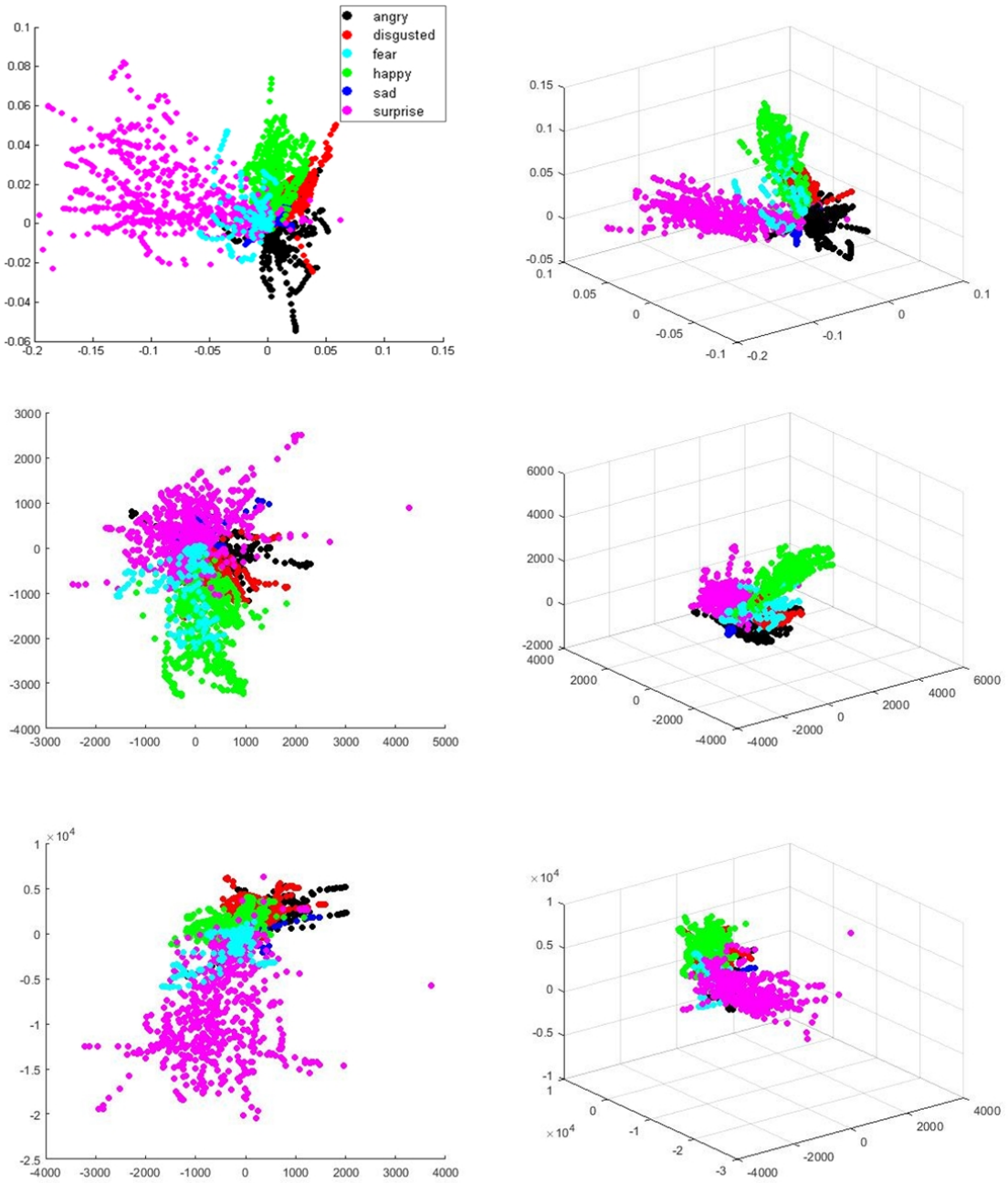


Fig. 5.31 Illustration of a 2D (left) and 3D (right) description models using shape (top row), texture (middle row), and appearance (bottom row) features of all CK+ dataset showing six expressions of different intensities.

5.6 Discussion

In this chapter, shape features were used to localize face and expressions, build a shape-based AFER system, and measure the influence of ageing and compound emotions, and intensity of expression on both the facial expression localization and recognition.

For the facial expressions localization, an automatic FEL system using the RFRV-CLM framework was developed to localize the facial expressions features. A parameter sensitivity analysis of the RFRV-CLM parameters was performed to find the optimal parameters that best solve the problem of facial expressions detection. We noticed that varying the parameters had little effect on the performance except for the frame width which had a significant effect. We found that three stages gave optimal results. The sensitivity and generality of the proposed FEL and the optimal parameters under several factors of facial expressions: age (young to old), expression (6-basic and 22-compound), intensity (neutral to peak), fake expression, and spontaneous expressions were measured. The results showed that the proposed localizer achieved good performance on five different facial expression datasets and the mean error of points' localization using RFRV-CLM was 3.4% of the IOD (2.1 mm) on 99% of samples, in 22 expressions it was 5.71% of the IOD (3.6 mm) on 90% of the data, and in continuous intensities it was 3.97% (2.5 mm) on 95% of the data, outperforming the mean error of alternative methods tested on the same datasets. This empirical evaluation showed that the proposed method generalizes well across a wide variety of face appearances, suggesting that it works sufficiently well to initialize further processing such feature extraction followed by age, expression, and intensity modelling for automatic facial expressions recognition.

For facial expression classification, a shape-based, texture-based, and appearance-based AFER was developed using the points from the proposed detector. In case of recognizing the expressions in a large range of age, the results again (as in the previous chapter) indicated that recognizing the expression in a small range of ages outperforms the performance of using the large range of ages and the combination of shape and texture features presented in this chapter outperformed the texture features in the previous chapter. In the case of classifying 22-compound emotions, the results demonstrated that the shape-based AFER slightly outperformed the texture-based AFER. Despite these good results of using shape features presented in this chapter, more investigations are required to investigate the possibilities that can be used for compound emotions modelling. In the case of expression recognition of different intensities, the results suggested that using the shape features extracted in this chapter is very promising for expression recognition of continuous intensities.

Chapter 6

Development and Comprehensive Evaluation of an Age-Based AFER System

6.1 Introduction

The results of the texture-based AFER system and the shape-based AFER described in Chapters 4 and 5 respectively indicated that modelling (training and testing) facial expressions using a limited range of ages has achieved better results than using a wide range of ages in AFER.

Motivated by this observation, this chapter examines the development of an age-specific AFER system that explicitly estimates age group and expression from the face image within a single framework. The underlying idea is that separating the age factors in the process of AFER modelling can reduce the confusion in the expression classification by the automatic classifiers when applied to different age groups. The second objective of this chapter is to study the influence of apparent age on the performance of AFER. The underlying hypothesis is that we do not know the real age of the person so we use the visual appearance to estimate age. And since some people might look younger or older than their real age it might be better to classify the expressions based on the apparent age. In summary, the argument of this chapter is that using the age group and in particular the apparent age instead of real age as prior knowledge for the expression classification task can help to improve the performance of expression classification.

The rest of the chapter is organized as follows. Section 6.2 introduces our motivation for combining classifiers. Section 6.3 describes the proposed method, and its framework, formula and stages. Section 6.4 represents a series of experiments to evaluate the proposed method. Finally, a summary of the findings is given in Section 6.5.

6.2 Motivations

The main goal of any AFER system is to achieve the best possible accuracy for facial expression classification. This aim traditionally led to the development of different facial expressions classification systems, in which a single classifier is used to recognize among several expressions. The final solution is to choose the classifier with the best performance. However, it has been observed in multi single-classifier systems that one classifier's misclassification might be correctly classified by one of the others Glodek et al. (2011); Kittler et al. (1998); Lanitis et al. (2004); Woźniak et al. (2014). This observation suggests that multiple-classifier systems potentially offer complementary information about the problem at hand which could be harnessed to improve the overall performance. Moreover, experiments in previous chapters have also shown that the performance of single-classifier facial expressions systems was best on the age range for which the system was trained and that performance decreases as the age difference between the training set and the testing set is increased. This observation suggests that using different age-specific facial expression recognition systems (one trained for each age group range) potentially offered complementary information. These observations motivated the idea of combining classifiers. The underlying assumption of combining classifiers is that the decision regarding the expression is made by combining the opinions of multiple individual facial expression classifiers in order to obtain a consensus decision. There are various classifier combination schemes depending on the problem at hand and it has been experimentally demonstrated that some of them outperform the single classifier scheme (Woźniak et al., 2014).

Therefore, in this thesis an age-specific-based AFER system is proposed that explicitly estimates age group and expression from the face image in a single framework. The proposed system consists of an age group estimator to estimate the age group of a subject. We then recognize the expression using the weighted average rule of the output of a set of age-group-specific expressions classifiers (one trained for each age group). The underlying assumption is that an age-specific classifier with expression variation only will outperform one expression classifier trained using a wide variation of both age and expression patterns because the overlap between age and expression features might lead to poor performance in the AFER system (see the sensitivity of automatic expression modelling to age features in Chapter 4.4.2.1 and Chapter 5.5.2.1). The other underlying reason is that when a single model cannot properly fit the data, the ensemble can make multiple estimates that each make errors in different ways then vote or average their predictions, cancelling out the errors of individuals by the committee decision.

6.3 Methods

The goal is to decrease the negative effect of ageing features, which can resemble emotions, for better expression recognition. Many systems in the literature e.g. (Martinez and Valstar, 2016; Sariyanidi et al., 2015) have taken the training and testing data (input space) from a limited range of ages. Experiments reported in this thesis showed that the accuracy of those systems is reduced if the age difference between the test image and the training set is big where the system is trained on a limited range of ages and tested using a different range of ages (see the not highlighted cells in Table 5.7). The experiments also found that extending the input space to span a wider range of ages for both the training and the testing will not improve performance since this will begin to include older subjects whose aged features appear in a similar manner to some expressions and the overlap between them will lead to poor performance of the expression classifier (see pink cells in Table 5.7). Therefore, recognizing the expression in a narrow age range should decrease the variation in age features and hence the performance of expression classifiers should be improved (see yellow, green, and blue cells in Table 5.7).

The proposed method in this chapter aims to learn the relationship between age and expression and how using it may lead to better recognition of the expression. Our model consists of three main stages: (i) an automatic facial key point detector and feature extractor, (ii) an age group estimator, and (iii) an ensemble of age-group-specific expression recognizers. In the first stage, an automatic facial key points detector is learned from the training data to be used to extract the features regarding age and expression. Since both age group estimation and facial expression classification tasks started from the same face image and they are entirely related, the same features are used by both tasks. In the second stage, an age group estimator is learned using the extracted features to estimate the age group of the testing image. We compute the probability that the test image came from each of the age group classes (a description of the classes is given below) to provide a priori information regarding age to the third stage. In the third stage, a set of age-group-specific expression classifiers are trained and used to classify the expression type. The expression classifiers give the probability of the face being in each of the expression classes, in each of the age groups. The probabilities are then weighted by the age-group probability to give the final output.

The essential idea is to take advantage of the hypothesis that the expressions can be more accurately recognized by using an age-appropriate classifier. The benefit of this method is increasing the ability and the accuracy of the system of facial expression recognition by using all the representations of the age patterns jointly to make the final decision. Fig. 6.1 describes the proposed system in this chapter.

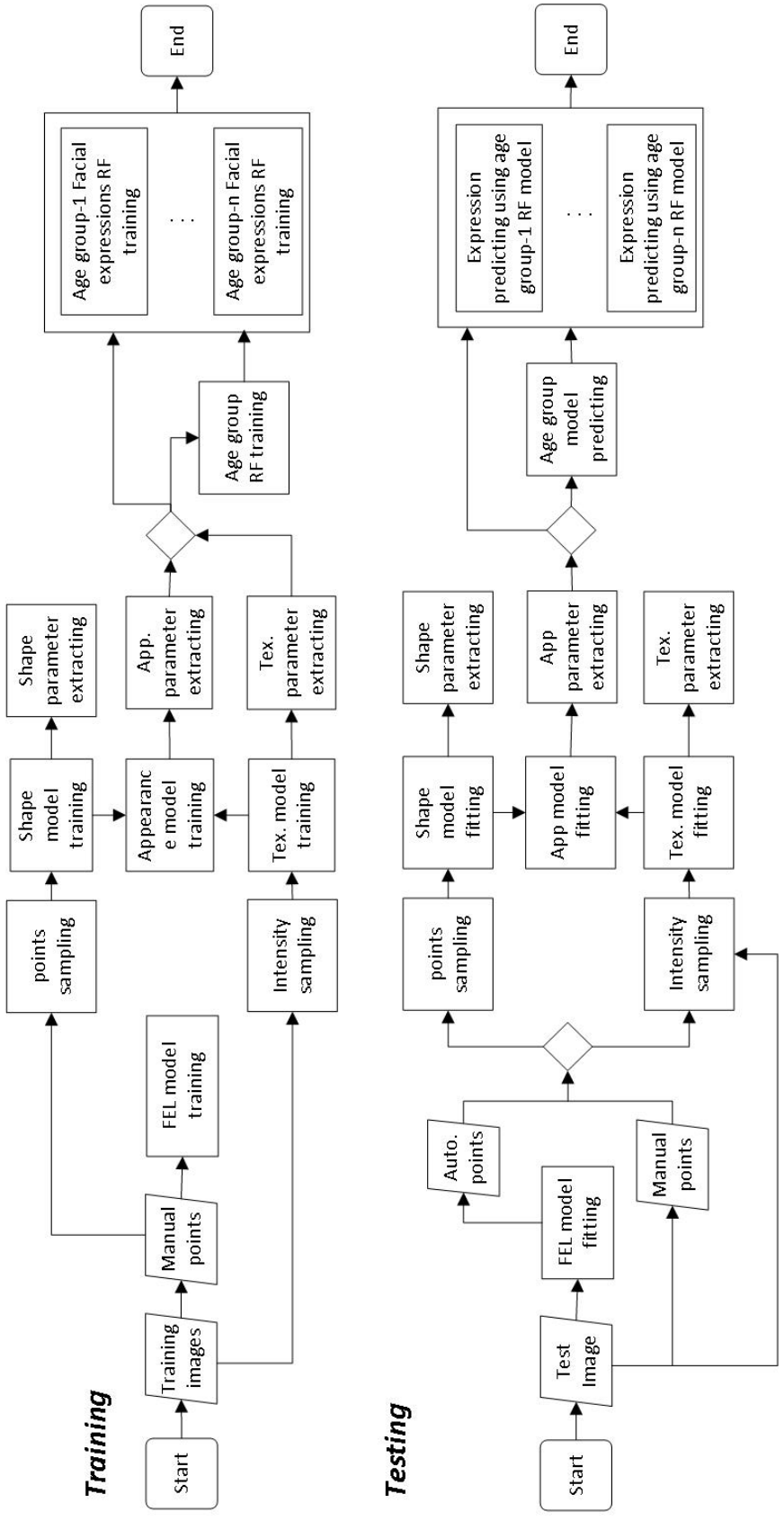


Fig. 6.1 System overview showing our three stage system. The first stage is the automatic point detector and feature extractor giving both age and expression information to send to the age group estimator (stage 2) to estimate the age group and to the age-group-specific expression classifiers (stage 3) to estimate the expression category.

6.3.1 Model Formula

Suppose we have N labelled training images and for each image we have $I_i = (v_i, a_i, e_i)$ where v_i is the feature vector for image I_i, a_i is the age group label for image I_i (in our experiments $a \in \{1,2,3\}$) for young, middle, and old age groups respectively, and e_i is the expression class labels for image I_i (i.e. $e \in \{1,2,3,4,5,6\}$) for angry, disgust, fear, happy, neutral and sad respectively.

Once the feature vector v is extracted, one age group estimator $f_a(v)$ and three age-group-specific expression classifiers $f_e(v)$ are trained for predicting the subject's age group a and expression e using equations 6.1 and 6.2 respectively as follows.

$$f_a(v) = \underset{a}{\operatorname{argmax}}\, p(a|v) \tag{6.1}$$

Where $p(a|v)$ represents the probability that v is in age class a and the winning class is the one with the largest a.

$$f_e(v) = \underset{e}{\operatorname{argmax}}\, p(e|v, f_a(v)) \tag{6.2}$$

where $p(e|v, f_a(v))$ is the confidence that v has expression e when estimated using the classifier for age group $f_a(v)$ (see Equation 6.1). The winning class is the one with the largest e.

The number of age-group-specific expression classifiers depends on the number of age groups in the training data. For the experiments reported in this thesis, we have trained three age-specific expression classifiers: young, middle, and old age groups. The age ranges of each class are defined below.

Once the age group and age-group-specific expression classifiers are learned, they can be used in one of two schemas to recognize the expression. The schema differ in how the age and expression classifiers are used, in either a hard or a soft schema.

In the hard-schema, the estimated age group a is used to select one age-specific expression recognizer as in Eq. 6.3.

$$e = \begin{cases} \operatorname{argmax}_e\, p(e|v,1), & \text{if } f_a(v) = 1 \\ \operatorname{argmax}_e\, p(e|v,2), & \text{if } f_a(v) = 2 \\ \operatorname{argmax}_e\, p(e|v,3), & \text{if } f_a(v) = 3 \end{cases} \tag{6.3}$$

Where $p(e|v,1)$, $p(e|v,2)$ and $p(e|v,3)$ represent the probability that v has expression e when estimated using classifier for young, middle and old age groups respectively. The winning class is the one with the largest e.

On the other hand, using a soft-schema, the output of expression type e will make full use of the confidence values of age estimator and all the age-group-specific expressions classifiers as in Eq. 6.4.

$$f_e(v) = \underset{e}{\operatorname{argmax}}\, p(e|v) \tag{6.4}$$

where:

$$p(e|v) = \frac{\sum_a p(a|v)p(e|v,a)}{\sum p(a|v)} \tag{6.5}$$

where $p(a|v)$ is the probability that v belongs to age group a from Equation 6.1 and $p(e|v,a)$ is the probability that v has expression e using the classifier of age group a. The winning class is the one with the largest e.

The advantage of using the weighted combination of all the classifiers is that more information about the classification can be obtained and subjects whose apparent age puts them in the wrong chronological age group will be dealt with more effectively (see results in Table 6.6). Details of each stage are described below.

6.3.2 Stage One - Facial Expression Localization and Feature Extraction

In the previous chapter, we described a methodology for facial expression point localization using a RFRV-CLM framework in a multi stage fashion, and facial expression feature extraction including the shape b,the texture b_g, and the appearance c features. This methodology is the basis of age-group-based AFER system described in this chapter.

6.3.3 Stage Two - Age Group Estimation

From stage one each image is represented as a feature vector v_i including shape b, texture b_g, or appearance c with its corresponding age group a and expression category e.

In this stage, the data from the three age and expression datasets described in Section 3.2.1 are combined with the corresponding manual and automatic points to create a data set which contains subjects with ages ranging from 8 to 94 years. Fig. 6.2 shows the age distribution from the three datasets. Then the data is divided into three groups of age; Young (18-39), Middle (40-69), Old (70-94). Table 6.1 shows the lower and upper age limits of the three age groups - 1:young, 2:middle, and 3:old. Subjects with ages from 8 to 17 were not included because there were too few samples. The objective of this stage is to estimate the age group a. Therefore, we train a random forest classifier Breiman (2001) described in Section 3.5 to estimate the probability that an example with feature vector v belongs to age group a, $p(a|v)$. The most probable age group is given using Equation 6.1.

Table 6.1 Lower and upper age limits in years of three age groups

Group	1	2	3
Lower	18	40	70
Upper	39	69	94

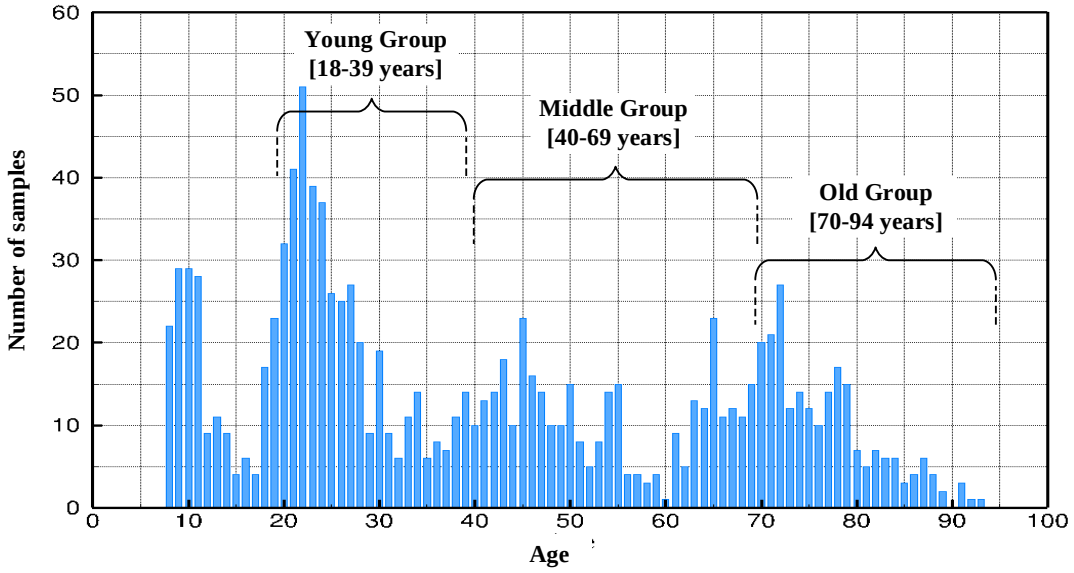


Fig. 6.2 Age distribution of three age and expressions datasets combined: FACES, Lifespan and NEMO

6.3.4 Stage Three - Expression Classification

The aim of this stage is expression recognition. For each age group, a, a separate random forest classifier is trained to estimate the probability of each expression, $p(e|v,a)$ using Equation 6.2.

One approach to classifying a new image (of unknown age) is to estimate the age group using Equation 6.1 then use the appropriate expression classifier for that age using Equation 6.3.

An alternative is to weight by the probability using Equations 6.4 and 6.5. The essential idea here is to first classify the input test image into a specific age group, and then the expression recognition is performed in the classified group. For the experiments reported in this thesis, three age-specific expression classifiers are trained for young, middle, and old groups respectively.

The advantage of using the weighted combination of all the classifiers is that more information about the classification can be obtained and subjects whose apparent age puts them in the wrong chronological age group will be dealt with more effectively (see results in Table 6.7).

Once the three stage models are built and combined, the fully automated age-specific expression classification system will start on the test image using the proposed facial expression keypoint detector (stage 1) to extract the age and expression features. The extracted features are then sent to the age group estimator (stage 2) to estimate the probability of each age group, followed by the use of the corresponding expression classifiers (stage 3) for each specific age group indicated by the results of age estimator (stage 2) to recognize the expression based on the combinations of the results of all classifiers.

6.4 Experimental Evaluation

In this section a series of experiment is presented to evaluate the proposed age-group-based AFER system.

Database: The experiments are performed using FACES, LifeSpan, and NEMO datasets described in 3.2.1.

Evaluation Metric: For age group estimation and expression classification, both the average of all classes with the standard error and per-class (confusion matrix) classification accuracy between the ground truth label and the predicted label are reported. To avoid overfitting and identity bias issues, 10-fold cross validation (person independent) experiments are applied. In all the experiments below the random forest classifier described in Section 3.5 was used for age group estimation and expression classification.

6.4.1 Experiment 1 - Age Effect on AFER

Case Description: In the previous chapter, we studied the influence of age on the performance of automatic facial expression classification. We performed experiments with both the manually located points and those from the automatic system. In each case we used feature vectors which were either the shape parameters b, the texture parameters b_g or the appearance dc. We trained age-specific random forests to estimate the probability of each expression given a feature vector and age group, $p(e|v,a)$, and age-agnostic RFs to estimate the probability of each expression for all age groups, $p(e|v)$. the results are summarized in Table 6.2, in which we indicated that when using the most appropriate age-specific classifier for each subject, we get improved performance, compared to using an age agnostic system.

Table 6.2 Expression classification results using manual and automated annotation for age-specific models and age-agnostic model.

Feature	Young		Middle		Old		all	
	Manual	Automatic	Manual	Automatic	Manual	Automatic	Manual	Automatic
Shape	96.8±0.2	91.9±0.6	95.0±0.7	87.1±1.3	89.2±0.6	78.2±2.2	89.0 ±0.5	76.8 1 ±1.6
Tex	96.3±0.7	95.9±0.9	95.6±0.1	93.3±1.2	91.1±0.3	86.0±1.8	91.0 ±0.3	87.2 ±1.3
App	97.7±0.7	96.5±0.2	97.6±0.2	94.7±0.6	93.8±0.4	88.5±0.9	93.4 ±0.6	90.1 ±0.8
Average	Manual	96.3		Automatic	93.2			

6.4.2 Experiment 2 - Real Age Effect on AFER

Case Description: The results in experiment 1 suggested that the best performance was achieved using age-specific classifiers if the age of the subject is known. If this isn't known, then it must be estimated from the image. We thus trained a Random Forest to estimate the age-group given different feature types. The benefit from the age-group estimator is to use its output as prior knowledge to the age-specific expressions classifiers by combining them in a single framework. The idea is that the age group estimator aids by predicting the age group of the test image in the first place and we use this as prior knowledge in the second stage to select the most age-appropriate expressions classifier (see Eq. 6.3). The aim here is to achieve an identical performance to the performance of using age-specific classifiers only.

Results: The performance of the age group classifier is summarised in Table 6.3 that gives the percentages of the data assigned to the correct age groups. It shows that using appearance features gives the best overall performance, and that manual annotations lead to slightly more accurate results with 86.6% compared to the automatic annotations with 84% correct assignments.

Table 6.3 Accuracy of age group estimation using real age. Shape, Tex, and App are the shape, texture, and appearance features respectively

Feature	Manual Points	Automatic Points
Shape	81.3±0.6	76.0±0.2
Tex	86.1±0.6	82.6±0.2
App	**86.6±0.1**	**84.0±1.1**

Table 6.4 shows the performance results of the expression classification when combining the age-group estimator with age-specific expression classifiers in a single model that we called hard-level-based or real-age-based schema (see Equation 6.3). These results show that although the accuracy of age group estimation is relatively low (84%), its combination with age-specific expression classifiers helped to achieve comparable performance,

with 95.2% manually and 93.8% automatically shown in Table 6.4, to the performance of using age-group specific models with 96.3% manually and 93.2% automatically (see Table 6.4) .

Table 6.4 Expression classification results using hard-level (real-age-based) schema. Shape, Tex, and App is the shape, texture, and appearance features respectively

Feature	Young		Middle		Old	
	Manual	Automatic	Manual	Automatic	Manual	Automatic
Shape	95.0±0.2	91.2±0.6	94.1±0.7	85.1±1.3	89.8±0.6	76.4±2.2
Tex	94.9±0.7	95.6±0.9	92.6±0.1	92.9±1.2	88.1±0.3	85.5±1.8
App	96.9±0.7	96.7±0.2	96.1±0.2	94.7±0.6	92.7±0.4	90.0±0.9
Average	Manual	95.2		Automatic	93.8	

Table 6.5 shows the confusion matrix for the age-group classifier using automatic points of appearance features. These results reveal that 6.8% (see yellow cell) of the young group and 9.6% (see blue cell) of the old group were classified in the middle age group and 12.3% and 16.2% (see green cells) of the middle group were classified in the young and old age-groups respectively. This big confusion among age groups might be because there are people who look younger or older than their real age, indicating that taking the apparent age into account might enhance the performance of the AFER system.

Table 6.5 Confusion matrix for real age group classifier using appearance features using automatic points.

Group	Young	Middle	Old
Young	91.2	6.8	2.0
Middle	12.3	71.5	16.2
Old	1.1	9.6	89.3
Mean	84.0±1.1		

6.4.3 Experiment 3 - Apparent Age Effect on AFER

Case Description: The results of experiment 2 suggested that it might be better to analyse the effect of apparent age on the performance of AFER. As a result, in Equations 6.4 and 6.5 we propose to weight and combine the probability of all age group estimators with the all age-group-specific expression classifiers in order to have full use of all the age and expressions representations. The idea is to fully utilize the representations not only from one expression classifier such as the last calculated results in Table 6.4, but from all the

age and expressions classifiers in what can call a soft-level-based or apparent-age-based schema.

Results: Table 6.6 shows the recognition accuracies of the apparent-age-based schema. In this table, the performance of the age group estimator and age-group-specific expression classifiers are combined in which the expression is estimated using all expression classifiers in proportion to their age weights from the age estimator. These results demonstrate that the weighted combinations scheme helped to improve the performance by 2.5% for the manually annotated data and 1.9% for the automatically annotated data compared to the real-age-based schema in the previous experiment.

Table 6.6 Expression classification accuracies using soft-level (apparent-age-based) schema. Shape, Tex, and App are the shape, texture, and appearance features respectively

Feature	Young		Middle		Old	
	Manual	Automatic	Manual	Automatic	Manual	Automatic
Shape	97.9±0.3	92.9±0.4	97.3±0.5	89.5±0.3	93.0±0.5	79.4±1.5
Tex	97.4±0.5	96.9±0.1	96.8±1.1	95.2±0.5	92.8±0.6	88.5±0.1
App	98.6±0.1	98.0±0.2	98.8±0.3	95.3±0.4	95.9±0.5	93.80±0.2
Average	Manual	97.8		Automatic	95.7	

These results motivated us to label the dataset with apparent age label. After that each image in the data base is labelled manually by the apparent age group label by five assessors. We also used the age group estimator to label each image by the apparent age. Those labels are used to redivide the dataset into three age groups by moving some people among the groups. We then re-trained the age group estimator and the age-specific facial expression classifiers using the new groups and the apparent age group label in order to use them again as in Equations 6.4 and 6.5. We train the age group and expression classifiers several times (one-based on 'seach assessor labels, and one-based the age group classifier labels).

Table 6.7 shows the summary results of expression recognition using the schemas of combining age estimator and age-specific expressions classifiers as follows.

- Yellow cells show the overall mean performance of expression classification when three age facial expression classier are trained and tested using the same age range (each classifier is trained and tested on the same age range). In this case the age is known and selected manually by the user.
- Green cells show the overall mean performance when using the age classifier to automatically select the age-specific expression classifier (what we called hard decision). In this case one of the expression classifiers will be used based on the output from the age group estimator.

- Blue cells show the overall mean performance using a probabilistic approach (using equation 6.4), avoiding a hard decision. In this case all the expression classifiers will be used, each based on its weight from the age group estimator and the final outputs will be combined and normalized.
- Gray cells show the overall mean performance of the soft approach when we train the age group estimator and the expression classifiers on age groups defined by the apparent age of the individuals (as estimated by five assessors), rather than their real age.
- Pink cells show the overall mean performance of the soft approach when we train the age group estimator and the expression classifiers on age groups defined by the apparent age of the individuals (as estimated using the age group classifier), rather than their real age. This shows that such a "soft" approach leads to better results, which are actually better than those achieved when the true age is known (yellow cells).

In summary, the results in Table 6.7 demonstrate that the models are sensitive to the age of the individual, and that expression can be better estimated if we know the subject's approximate age and can use an appropriate age-specific classifier. These results demonstrate also that is better to use apparent age in selecting the most appropriate expression classifier than chronological age.

6.4.4 Comparing to other Methods

Table 6.8 compares our results to the very recent results presented in Lou et al. (2018) and Yang et al. (2018) who used Local Binary pattern (LBP) feature and convolution neural network (CNN) respectively. These results show that the mean results of three datasets of our automated system outperform the results of the methods presented in those two papers. The reason that our results are better than the results of the method presented in Lou et al. (2018) might be because in our method the texture features are combined with shape features (compact representation), while in Lou et al. (2018) they used the texture features only. Moreover, the reason that our results are better than the results of Yang et al. (2018) using DL might be because we train the model using representative datasets of both age and expressions deformations, while in Yang et al. (2018), the authors used the pre-trained model on the MORPH dataset with 55,134 images of age only.

Table 6.7 Summary results of four schemas of combining age estimator with age-group-specific expression classifiers described in this chapter, tested on the FACES data set.

Group	points	Method				
		Age Specific Known Real Age	Age Specific Equation 6.3 Estimated Real Age	Weighted Equation 6.4 Estimated Real Age	Weighted Equation 6.4 Estimated Apparent Age by 5 assessors	Weighted Equation 6.4 Estimated Apparent Age by age group classifier
Young	Manual	97.7 ±0.4	96.9 ±0.7	98.6 ±0.1	98.2 ±0.4	98.8 ±0.2
	Automatic	96.5 ±0.4	96.7 ±0.2	98.0 ±0.2	97.6 ±0.6	98.3 ±0.1
Middle	Manual	97.6 ±0.4	96.1 ±0.2	98.8 ±0.3	98.4 ±0.7	98.5 ±0.4
	Automatic	94.7 ±1.1	94.7 ±0.6	95.3 ±0.4	95.8 ±0.6	96.0 ±0.2
Old	Manual	93.8 ±1.5	92.7 ±0.4	95.9 ±0.5	96.0 ±1.0	96.3 ±0.4
	Automatic	88.5 ±2.1	90.0 ±0.9	93.8 ±0.2	93.9 ±1.1	93.7 ±0.3
Mean	Manual	96.4 ±0.8	95.2 ±0.4	97.8 ±0.3	97.5 ±0.7	**97.9 ±0.3**
	Automatic	93.2 ±1.2	93.8 ±0.6	95.7 ±0.3	95.8 ±0.7	**96.0 ±0.4**

Table 6.8 Expression recognition accuracies of FACES, Lifespan and NEMO datasets

		Lou et al. (2018)	Lou et al. (2018)	Yang et al. (2018)	[Present work]	[Present work]
Data	Emotion	Independent Learn % of age and expression	Joint Learn % of age and expression	Joint Estimation of age and expression	Age-Group Specific % of known age	Weighted Combination % of age and expression
FACES	Neutral	91.2	95.9	-	97.9	98.5
	Happy	99.4	98.8	-	98.8	99.1
	Anger	84.8	88.3	-	92.2	95.3
	Disgust	89.4	92.9	-	92.7	95.4
	Fear	92.4	94.1	-	96.5	97.1
	Sadness	83.1	83.0	-	81.3	89.0
	Average	90.1	92.2	95.1	93.2	96.0
Lifespan	Neutral	97.4	96.2	-	98.7	99.2
	Happy	85.9	88.1	-	85.3	91.6
	Average	91.7	92.2	96.3	92	95.4
NEMO	Neutral	97.9	98.1	-	95.1	100
	Happy	97.5	97.9	-	94.8	99.1
	Average	97.7	98.0	-	95.0	99.6
Mean		93.2	94.1	95.7	93.4	**97.0**
Error		6.8	5.9	4.3	6.6	**3.0**

6.4.5 Computation Complexity

Results in Table 6.9 show the computation complexity of the current work of combining the age and expressions' classifiers when using both the estimated age expression classifier (hard level) and when using the weighted expression classifiers (soft-level) based on both the real and apparent age.

The time was recorded in milliseconds (ms) and was compared to the time of the method in Lou et al. (2018) which recorded the time in seconds (for a simpler comparison it was converted to ms). The results demonstrated that in addition to lower test time of our method with 0.16 ms per image without feature extraction computation performed on a Dell Intel(R) Core(TM) i5-2400 CPU 3.10 GHz and 16GB memory than the method in Lou et al. (2018) with 10 ms performed on a machine with 2 Intel(R) Xeon(R) CPU X5570 2.93 GHz and 64 GB memory, the computation time of the soft-level schema of using all the age and expressions classifiers is comparable to the computation time of hard-level schema of using one expression classifier select by the age estimator.

Table 6.9 Computational complexity (in millisecond) of the proposed method.

Method	Test time(ms)
Age-agnostic	0.06
Age-specific (known real age)	0.14
Age-specific (estimated real age)	0.08
Weighted (estimated real age)	0.15
Weighted (estimated Apparent age)	0.16
Independent Learn Lou et al. (2018)	7.00
Joint Learn Lou et al. (2018)	10.00

6.4.6 Example Results

Figure 6.3 shows some examples of results from our system including automated facial expression points localization, automated age group estimation, and automated facial expression recognition on three different age and expressions datasets. These results show that, although the system makes some mistakes in the age group estimation, the expression is recognized correctly due to the weighted combination schema of different expressions classifiers. For instance, the second image in the first row, in spite of the age group being predicated wrongly as middle age, the expression is recognized correctly as angry.

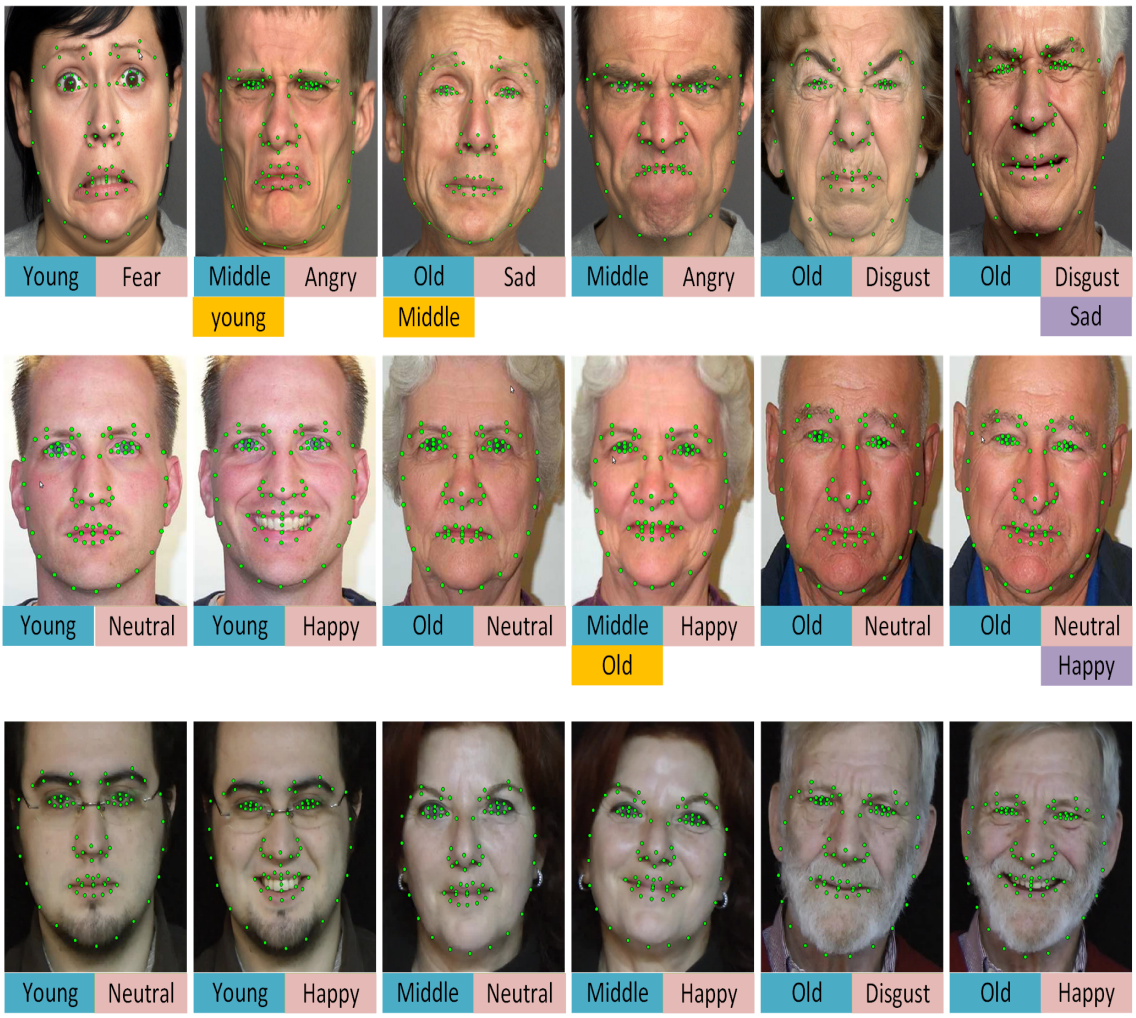


Fig. 6.3 Example results of our system including point's localization, age group estimation, and expression category recognition on three different age and expression datasets: FACES data (first row), LifeSpan data (second row), and NEMO data (third row). The predicted age group and expression are in blue and pink background respectively. The ground truth is in the yellow and purple background for age and expression respectively if the system makes a mistake in them.

6.5 Discussion

This chapter has presented an age-specific-based AFER system which used the age group information as prior knowledge to obtain better results of expression classification. The proposed system can be used for automatic facial feature point localization, automatic age group estimation, and automatic facial expressions recognition and deals with individuals from a wide range of ages. It showed that using age-specific expression classifiers gives better results than an age-agnostic classifier, and that integrating information over all age estimates leads to the best overall performance.

The performance of the proposed system was evaluated on three age and expression datasets to investigate the sensitivity of automatic expression classification to age features

(real age and apparent age). The results on all the datasets are promising, and they demonstrate the potential of our approach in real life applications. The face processing procedures described are fully automatic; errors for the classification experiments may be caused either by failure in locating landmarks accurately, or by the failure of the classification algorithm. We showed that the system can recognize the expression automatically based on the apparent age in which the weighted combinations of age and expression classifiers can be used to achieve better expression recognition. Finally, our results showed that the fully automatic system based on automatic points identifies 97.0% of expressions correctly, almost as effectively as the system based on manual points (97.9%).

Chapter 7

Conclusion

7.1 Thesis Summary and Conclusion

This thesis has covered the topic of AFER across a large range of ages (including older people), expressions (compound emotions), and intensities (ranging from a neutral expression to the apex of the target expression). A comprehensive study has been undertaken to investigate and measure the effect of those problems on the performance score of AFER, along with an examination and validation of the sensitivity and ability of the face modelling methods in order to explore whether they are sufficient in extracting meaningful face features to use in distinguishing among expressions under a wide range of face deformations related to the problems under study.

The study started by using texture measurement methods to examine the ability of the existing face descriptor methods in describing the face image with a wide range of face deformations regarding the problems under study and to analyse the impact of those problems on the face texture patterns and hence on the accuracy of texture-based AFER. In this study, as a first contribution, we showed that by using BRIEF (Calonder et al., 2012), we can develop a new face descriptor model that able to describe the face image and can generalize to new data sets. The BRIEF descriptor is able to generate the discriminative features globally from the image with explicit shape features. However, when BRIEF is used to generate features from an image with no explicit shape such as a face, BRIEF is unable to generate discriminative features. We thus proposed to use BRIEF locally to ensure that each pixel in the image is evaluated locally to capture the local shape surrounding it. Empirical and comprehensive evaluation on three different facial expression datasets demonstrated that this model gave satisfactory performance compared to other local face descriptor techniques including LBP (Ojala et al., 2002) and QLZM (Sariyanidi et al., 2013). The experimental results also showed that there is insensitivity to the choice of BRIEF's representation to the patch size and the descriptor length. The results show as well that the BRIEF-based method outperformed LBP in recognition rate and the

size of the representation. For example, in facial expression recognition, the BRIEF-based face descriptor achieved a classification rate of 96.0% with a histogram of integers ranged in [0,63], while the performance using the LBP method was 82.4% with a histogram of integers ranged in [0,255] on the same dataset. We also obtained comparable performance to the BOW and QLZM but with smaller standard deviation from the mean recognition rate suggesting that the proposed model is more stable.

Moreover, a comparative evaluation among three AFER systems including BRIEF-based, LBP-based, and QLZM-based AFER showed that using a large range of ages and compound emotions has a significant effect on the face texture pattern and on the performance of AFER. The results indicated that BRIEF achieved satisfactory accuracy on a range of facial expression recognition datasets with different characteristics. In the case of the large range of ages with 6-basic expressions, BRIEF outperformed the LBP and QLZM with the same configuration. These results also showed that the classifier trained on a limited range of ages has less confusion among the expressions than the classifier trained on a large range of ages. In the case of classifying 22-compound emotions, we showed that the texture features presented in this thesis are not sufficient to discriminate among them and more investigations are required to reduce the confusion between the basic and the non-basic expressions.

We then extended our study by using shape measurement methods to examine the sensitivity of the existing shape localization methods in detecting facial features points with a wide range of face deformations regarding the problems under study and then to analyse the impact of those problems on the face shape pattern and hence on the accuracy of shape-based AFER. In this study as a second contribution, we showed that by using RFRV-CLM framework (Cootes et al., 2012; Lindner et al., 2015), we can develop a fully automated FEL system that is able to detect the facial key points in a multiple-stage (coarse-to-fine) framework and can generalize accurately to new data sets with a wide range of facial appearance variations. The study showed also that despite the effect of the previously mentioned problems on the face shape pattern, the proposed facial expression localization based on RFRV-CLM achieved good performance against that effect. Empirical and comprehensive evaluation on five different facial expression datasets of simple (6-basic with a limited range of ages) and complex (6-basic with a wider range of ages, 22-compound emotions, and different intensities for the same expression) cases of facial expression demonstrated that this model gave excellent agreement with ground truth data and outperformed the results of other FFPD evaluated on the same datasets.

The results of the developed FEL system were then used in this thesis to initialise an automatic point-based AFER. We presented a comparative study among shape-based, texture based, and appearance-based AFER to assess the performance score of the automatic points-based AFER compared to the manual points-based AFER, along with a measurement of the effects of the problems under study on the overall outcomes of AFER.

This comparative study showed that the ageing, compound emotions, and the intensity patterns have a significant effect on the face shape pattern and on the performance of shape-based AFER. Although, the age and compound emotions patterns effect the shape and the texture features, the appearance-based AFER is less sensitive to that, and again the results showed that training and testing using a limited range of ages outperformed the results of training and testing using a large range of ages. The empirical evaluation showed that the developed FEL using RFRV-CLM generalizes well across a wide variety of face appearances with very few errors between the automatic and manual points, suggesting that it works sufficiently well to initialize further processing such as feature extraction. In case of facial expression with continuous intensity, the shape features only or its combination with texture feature (appearance) gave the best separation among videos of different emotions and among the frames of the same expression with a smooth path and they are encouraging to develop an affective learning algorithm as shown in Figures 5.30 and 5.31.

In summary, our study demonstrated that age-dependent facial expression recognition outperformed age-independent facial expression recognition and using appearance initialized by coarse-to-fine RFRV-CLM outperformed the other features described in this thesis. Figure 7.1 summarizes the work has been done in the proposed study regarding the performance of the texture-based and shape-based AFER on three datasets regarding the problems under consideration.

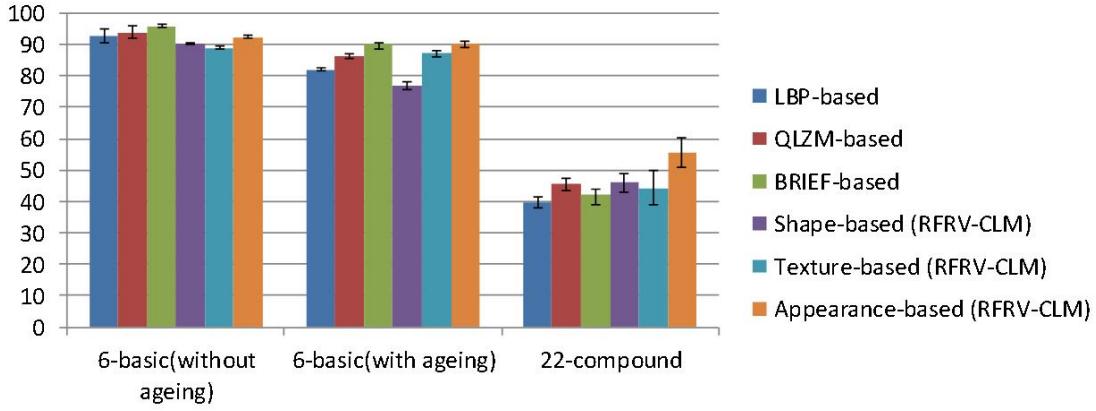


Fig. 7.1 Summary of the work that has been done in the reported study.

Our final contribution built on the first and the second contributions, and it is the development of a system for automated facial expression recognition (in terms of basic emotion categories) that is informed by the age of the people in the images. To this end, we segment the subjects into several age groups and use that information to augment learning of the facial expression classifiers. As the first step, the locations of fiducial facial points are extracted using the RFRV-CLM detector. These points are then used to build shape, texture, and appearance models in order to extract the shape, texture, and appearance features for input into the age classifiers, followed by age-specific facial

expression classifiers. The results on three data sets of age and facial expression coded in terms of emotion categories are presented. Experiments are done thoroughly, showing the benefits of the age segmentation. Furthermore, the proposed approach outperforms the age-agnostic classifiers, or is otherwise comparable, to the results found for alternative models recently applied to the problem. In this system, we showed that using the age information as prior knowledge to the facial expression recognition helped to reduce the confusions between the expression due to the age and expressions similarities. We then realized that using the apparent age as prior knowledge to the expressions classification helped to improve the performance more than that of using real age.

7.2 Thesis Limitations and Future Works

Although this work has been made a significant number of investigations to cover the objectives outlined in Section 1.5, it represents and to some extent is the start of a development and validation process that requires further investigations in the future, some of which are suggested below.

- One of the primary limitations of the proposed BRIEF-based face descriptor is that it is used to describe the face image of facial expression from static images. Since recognizing the expression from a video sequence is more reliable and easier due to the availability of more information about the expression, the BRIEF face descriptor can be extended in the future to a dynamic texture descriptor from three orthogonal planes. The benefit from this extension is that we can describe and extract the dynamic texture feature of the expression as in the cases of facial expression with continuous intensity. That extension can be similar to the extension of LBP to LBP-TOP (Zhao and Pietikainen, 2007) and LPQ to LPQ-TOP (Jiang et al., 2014, 2011).

- Moreover, due to the good results of the BRIEF face descriptor, including shorter descriptor and higher recognition rates, we believe that the proposed face descriptor could be applicable to several other face analysis tasks as well, especially real-time or mobile applications where memory and computational power is limited.

- Another future direction that can be drawn from this thesis is that owing to the excellent performance of FEL using the RFRV-CLM described in Chapter 5, the proposed FEL system can be a useful tool for building real-world systems and could be applied to other face applications that require high registration accuracies such as face detection, face recognition, and face synthesis, which could be an exciting area to research because the process of manually annotating images is a labour intensive process which can often contain erroneous annotations especially with

large datasets. A clear example is that the proposed FEL system and the extracted features based on it successfully captured the facial-expression points at different intensities as highlighted in Sections 5.5.1.5 and 5.5.2.3, it would be interesting to explore the use of that model in modelling the expression's intensity such as building an unsupervised model since the frames in the videos are not labelled. Another example is that in the case of the compound emotions, the proposed FEL system has successfully captured the facial-expression points of 22-compound emotions as highlighted in Section 5.5.1.4, it would be interesting to explore the use of that model in modelling the compound expressions such as modelling the similarities between the basic and non-basic expression or using the proposed model to capture the AUs of 22 expressions.

- One of the limitations of the age-specific AFER system described in Chapter 6 is that the number of age groups is restricted to a range of 1 or 3 groups. To obtain a general benefit from the idea of segment the data into age groups in the performance of expression classification and age estimation tasks, a joint optimization of the age segmentation and expression classification might be more useful, rather than using heuristics for grouping the age of the subjects.

- Furthermore, in this thesis, the impact of human ageing on the performance of expression recognition was evaluated using three age and expressions datasets described in 3.2. In these datasets, for each subject, there are images or videos for several expressions but at one age only of the subject. Further work can be done to improve age and expression datasets. For example, since the age-related structure progresses slowly and gradually across the age of the person, the creation of a facial expression dataset that covers a wide range of ages and expressions for each subject might be more believable when studying the effect of the human ageing on the performance of age estimation and facial expression recognition. In other words, collecting a new age and expressions dataset that contains at least one image or video for each subject at several ages and several expressions would be promising data for both age estimation and expression classification tasks. More datasets or expanding previous sets would be a simple improvement that can help move the research forward quicker.

- Moreover, the other limitation of the age-specific AFER system described in Chapter 6 is that the number of the assessors who labelled the ageing data with apparent age groups is restricted to five persons. To obtain a general benefit from the idea of using the apparent age in the performance of expression classification and age estimation tasks, more work is required to label the data by more assessors.

- Finally, despite the good performance that was obtained for landmark localization, face representation, age estimation, and expression classification using small training sample size, it would be interesting to investigate the use of other face representation methods such as DL which have showed remarkable performance in several computer vision and image processing researches on the small size of the data.

7.3 List of Publications

- N. Algaraawi, and T. Morris. Facial expression recognition with Binary Robust Independent Elementary Features (BRIEF). In preparation for submission.
- N. Algaraawi, and T. Morris. Comparative study a mong LBP, QLZM, and BRIEF features under Large Range of Age, Intensity, and Expressions Variations. In preparation for submission.
- N. Algaraawi, T. Morris, and T.F. Cootes . Facial Expression Localization under Large Range of Age, Intensity, and Expressions Variations using Coarse-to-Fine RFRV-CLM. IEEE Transactions on Affective Computing. In preparation for submission.
- N. Algaraawi, T. Morris, and T.F. Cootes. Fully Automated Age-Specific Expression Classification using Real and Apparent Age. IEEE Transactions on Affective Computing. In preparation for submission.

References

Adolphs, R. and Tranel, D. (2004). Impaired judgments of sadness but not happiness following bilateral amygdala damage. *Journal of cognitive neuroscience*, 16(3):453–462.

Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face recognition with local binary patterns. *Computer vision-eccv 2004*, pages 469–481.

Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041.

Allaert, B., Mennesson, J., Bilasco, I. M., and Djeraba, C. (2018). Impact of the face registration techniques on facial expressions recognition. *Signal Processing: Image Communication*, 61:44–53.

Alpaydin, E. (2010). Introduction to machine learning. [sl].

Alshamsi, H., Meng, H., and Li, M. (2016). Real time facial expression recognition app development on mobile phones. In *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on*, pages 1750–1755. IEEE.

Anderson, R., Stenger, B., Wan, V., and Cipolla, R. (2013). Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3382–3389.

Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (2016). Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 96–104.

Asthana, A., Lucey, S., and Goecke, R. (2011). Regression based automatic face annotation for deformable model building. *Pattern Recognition*, 44(10-11):2598–2613.

Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE.

Baker, S. and Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE.

Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y., and Martinez, A. M. (2017). Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*.

Breiman, L. (1984). Classification and regression trees.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breuer, R. and Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*.

Bromiley, P., Adams, J., and Cootes, T. (2015a). Localisation of vertebrae on dxa images using constrained local models with random forest regression voting. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 159–171. Springer.

Bromiley, P. A., Adams, J. E., and Cootes, T. F. (2015b). Automatic localisation of vertebrae in dxa images using random forest regression voting. In *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, pages 38–51. Springer.

Bromiley, P. A., Kariki, E. P., Adams, J. E., and Cootes, T. F. (2016). Fully automatic localisation of vertebrae in ct images using random forest regression voting. In *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, pages 51–63. Springer.

Bulat, A. and Tzimiropoulos, G. (2016). Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., and Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, 3(2):81–118.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., and Fua, P. (2012). Brief: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer.

Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

Caruana, R., Karampatziakis, N., and Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103. ACM.

Çeliktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13.

Chai, Z., Sun, Z., Tan, T., and Mendez-Vazquez, H. (2013). Local salient patterns—a novel local descriptor for face recognition. In *Biometrics (ICB), 2013 International Conference on*, pages 1–6. IEEE.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Chang, Y., Hu, C., Feris, R., and Turk, M. (2006). Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

Cootes, T. F., Ionita, M. C., Lindner, C., and Sauer, P. (2012). Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

Crammer, K. and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292.

Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., and Siddiqui, K. (2013). Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical image analysis*, 17(8):1293–1303.

Criminisi, A., Shotton, J., and Bucciarelli, S. (2009). Decision forests with long-range spatial context for organ localization in ct volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–80.

Criminisi, A., Shotton, J., Konukoglu, E., et al. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Criminisi, A., Shotton, J., Robertson, D., and Konukoglu, E. (2010). Regression forests for efficient anatomy detection and localization in ct studies. In *International MICCAI Workshop on Medical Computer Vision*, pages 106–117. Springer.

Cristinacce, D. and Cootes, T. (2008). Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067.

Cristinacce, D. and Cootes, T. F. (2006). Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3.

Cruz, A., Bhanu, B., and Yang, S. (2011). A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *Affective computing and intelligent interaction*, pages 341–350. Springer.

Dahmane, M. and Meunier, J. (2011). Continuous emotion recognition using gabor energy filters. In *Affective computing and intelligent interaction*, pages 351–358. Springer.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Dang, L. T., Cooper, E. W., and Kamei, K. (2014). Development of facial expression recognition for training video customer service representatives. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 1297–1303. IEEE.

Dapogny, A., Bailly, K., and Dubuisson, S. (2015). Pairwise conditional random forests for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3783–3791.

Dibeklioğlu, H., Salah, A. A., and Gevers, T. (2012). Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer.

Dibeklioğlu, H., Salah, A. A., and Gevers, T. (2015). Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279–294.

Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 36–47. International Society for Optics and Photonics.

Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.

Ebner, N. C. and Johnson, M. K. (2009). Young and older emotional faces: are there age group differences in expression identification and memory? *Emotion*, 9(3):329.

Ebner, N. C. and Johnson, M. K. (2010). Age-group differences in interference from young and older emotional faces. *Cognition and Emotion*, 24(7):1095–1116.

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM.

Ekman, P. (2002). Facial action coding system (facs). *A human face*.

Escalera, S., Fabian, J., Pardo, P., Baró, X., Gonzalez, J., Escalante, H. J., Misevic, D., Steiner, U., and Guyon, I. (2015). Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9.

Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., Tzimiropoulos, G., Corneou, C., Oliu, M., Ali Bagheri, M., et al. (2016). Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

Fan, H. and Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35.

Fanelli, G., Dantone, M., and Van Gool, L. (2013). Real time 3d face alignment with random forests-based active appearance models. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE international conference and workshops on*, pages 1–8. IEEE.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275.

Galvez-Lopez, D. and Tardos, J. D. (2011). Real-time loop detection with bags of binary words. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 51–58. IEEE.

Gálvez-López, D. and Tardos, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197.

Gao, X., Su, Y., Li, X., and Tao, D. (2010). A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(2):145–158.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al. (2011). Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer.

Graves, A., Mayer, C., Wimmer, M., Schmidhuber, J., and Radig, B. (2008). Facial expression recognition with recurrent neural networks. In *Proceedings of the International Workshop on Cognition for Technical Systems*.

Gunes, H. and Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):64–84.

Guo, G., Guo, R., and Li, X. (2013). Facial expression recognition influenced by human aging. *IEEE Transactions on Affective Computing*, 4(3):291–298.

Guo, G. and Wang, X. (2012). A study on human age estimation under facial expression changes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2547–2553. IEEE.

Hamsici, O. C. and Martinez, A. M. (2009). Active appearance models with rotation invariant kernels. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1003–1009. IEEE.

Hansen, M. F., Fagertun, J., Larsen, R., and Informatic, D. (2011). Elastic appearance models. In *BMVC*, pages 1–12.

Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2278–2288. IEEE.

Heinly, J., Dunn, E., and Frahm, J.-M. (2012). Comparative evaluation of binary features. In *Computer Vision–ECCV 2012*, pages 759–773. Springer.

Hess, U., Adams Jr, R. B., Simard, A., Stevenson, M. T., and Kleck, R. E. (2012). Smiling and sad wrinkles: Age-related changes in the face and the perception of emotions and intentions. *Journal of Experimental Social Psychology*, 48(6):1377–1380.

Hess, U., Blairy, S., and Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257.

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., and Traue, H. C. (2010). Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. *Acta psychologica*, 135(3):278–283.

Houstis, O. and Kiliaridis, S. (2009). Gender and age differences in facial expressions. *The European Journal of Orthodontics*, 31(5):459–466.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huang, C., Ding, X., and Fang, C. (2012). Pose robust face tracking by combining view-based aams and temporal filters. *Computer Vision and Image Understanding*, 116(7):777–792.

Huang, K.-C., Huang, S.-Y., and Kuo, Y.-H. (2010). Emotion recognition based on a novel triangular facial feature extraction method. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–6. IEEE.

Huo, Z., Yang, X., Xing, C., Zhou, Y., Hou, P., Lv, J., and Geng, X. (2016). Deep age distribution learning for apparent age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24.

Jain, D. K., Zhang, Z., and Huang, K. (2017). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognition Letters*.

Jia, J., Xu, Y., Zhang, S., and Xue, X. (2016). The facial expression recognition method of random forest based on improved pca extracting feature. In *Signal Processing, Communications and Computing (ICSPCC), 2016 IEEE International Conference on*, pages 1–5. IEEE.

Jiang, B., Valstar, M. F., Martinez, B., and Pantic, M. (2014). A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. Cybernetics*, 44(2):161–174.

Jiang, B., Valstar, M. F., and Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE.

Jin, X. and Tan, X. (2017). Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding*, 162:1–22.

Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991.

Kaulard, K., Cunningham, D. W., Bülthoff, H. H., and Wallraven, C. (2012). The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PloS one*, 7(3):e32321.

Kazemi, V. and Josephine, S. (2014). One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society.

Khotanzad, A. and Hong, Y. H. (1990). Rotation invariant image recognition using features selected via a systematic method. *Pattern recognition*, 23(10):1089–1101.

Kim, D. H., Baddar, W., Jang, J., and Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*.

Kim, M. and Pavlovic, V. (2010). Structured output ordinal regression for dynamic facial emotion intensity prediction. In *European conference on computer vision*, pages 649–662. Springer.

Kinoshita, K., Konishi, Y., Kawade, M., and Murase, H. (2012). Facial model fitting based on perturbation learning and it's evaluation on challenging real-world diversities images. In *European Conference on Computer Vision*, pages 153–162. Springer.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.

Koelstra, S., Pantic, M., and Patras, I. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954.

Kumari, J., Rajesh, R., and Pooja, K. (2015). Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491.

Lanitis, A., Draganova, C., and Christodoulou, C. (2004). Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, pages 2169–2178. IEEE.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Lee, D., Park, H., and Yoo, C. D. (2015). Face alignment using cascade gaussian process regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4212.

Lepetit, V. and Fua, P. (2006). Keypoint recognition using randomized trees. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1465–1479.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE.

Li, X., Xiaopeng, H., Moilanen, A., Huang, X., Pfister, T., Zhao, G., and Pietikainen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*.

Li, Y., Wang, S., Zhao, Y., and Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, 22(7):2559–2573.

Lien, J. J.-J., Kanade, T., Cohn, J. F., and Li, C.-C. (2000). Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146.

Lindner, C., Bromiley, P. A., Ionita, M. C., and Cootes, T. F. (2015). Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874.

Lindner, C. and Cootes, T. (2015). Fully automatic cephalometric evaluation using random forest regression-voting. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI) 2015–Grand Challenges in Dental X-ray Image Analysis–Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image*. Citeseer.

Lindner, C., Thiagarajah, S., Wilkinson, J., Consortium, T., Wallis, G., and Cootes, T. (2013). Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE transactions on medical imaging*, 32(8):1462–1472.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE.

Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 730–734. IEEE.

Liu, X. (2009). Discriminative face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1941.

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., and Chen, X. (2015). Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24.

Lou, Z., Alnajar, F., Alvarez, J. M., Hu, N., and Gevers, T. (2018). Expression-invariant age estimation using structured learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):365–375.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew, S., and Matthews, I. (2012). Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205.

Lucey, S., Ashraf, A. B., and Cohn, J. F. (2007). Investigating spontaneous facial action recognition through aam representations of the face. In *Face recognition*. InTech.

Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, 21(12):1357–1362.

Majumder, A., Behera, L., and Subramanian, V. K. (2013). Facial expression recognition with regional features using local binary patterns. In *International Conference on Computer Analysis of Images and Patterns*, pages 556–563. Springer.

Martinez, B. and Valstar, M. F. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in Face Detection and Facial Image Analysis*, pages 63–100. Springer.

Martins, P., Batista, J., and Caseiro, R. (2010). Face alignment through 2.5 d active appearance models. In *BMVC*, pages 1–12.

Martins, P., Caseiro, R., and Batista, J. (2013). Generative face alignment through 2.5 d active appearance models. *Computer Vision and Image Understanding*, 117(3):250–268.

Mary, R. and Jayakumar, T. (2016). A review on how human aging influences facial expression recognition (fer). In *Innovations in Bio-Inspired Computing and Applications*, pages 313–322. Springer.

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International journal of computer vision*, 60(2):135–164.

McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., and Picard, R. (2013). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888.

Mehrabian, A. (2008). Communication without words. *Communication theory*, pages 193–200.

Minear, M. and Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633.

Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

Mohammad, S., Morris, D., and Thacker, N. (2013). Texture analysis for the segmentation of optic disc in retinal images. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 4265–4270. IEEE.

Mohammad, S. and Morris, T. (2017). Binary robust independent elementary feature features for texture segmentation. *Advanced Science Letters*, 23(6):5178–5182.

Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

Motley, M. T. and Camden, C. T. (1988). Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Communication (includes Communication Reports)*, 52(1):1–22.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Ojansivu, V. and Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer.

Ono, A. (2003). Face recognition with zernike moments. *Systems and Computers in Japan*, 34(10):26–35.

Osman, O. F. and Yap, M. H. (2018). Computational intelligence in automatic face age estimation: A survey. *IEEE Transactions on Emerging Topics in Computational Intelligence*, (99):1–15.

Owusu, E., Zhan, Y., and Mao, Q. R. (2014). A neural-adaboost based facial expression recognition system. *Expert Systems with Applications*, 41(7):3383–3390.

Pantic, M., Pavlovic, V., and Rudovic, O. (2012). Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641. IEEE.

Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445.

Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, page 5. IEEE.

Picard, R. W. et al. (1995). Affective computing.

Prkachin, K. M. and Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274.

Pu, X., Fan, K., Chen, X., Ji, L., and Zhou, Z. (2015). Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing*, 168:1173–1180.

Quinlan, R. J. (1993). C4. 5: Programs for machine learning.

Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692.

Rivera, S. and Martinez, A. M. (2012). Learning deformable shape manifolds. *Pattern recognition*, 45(4):1792–1801.

Rogez, G., Rihan, J., Ramalingam, S., Orrite, C., and Torr, P. H. (2008). Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Rothe, R., Timofte, R., and Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15.

Rothe, R., Timofte, R., and Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157.

Rotshtein, P., Richardson, M. P., Winston, J. S., Kiebel, S. J., Vuilleumier, P., Eimer, M., Driver, J., and Dolan, R. J. (2010). Amygdala damage affects event-related potentials for fearful faces at specific time windows. *Human brain mapping*, 31(7):1089–1105.

Rowley, H. A., Baluja, S., and Kanade, T. (1996). Neural network-based face detection. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 203–208. IEEE.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE.

Sandbach, G., Zafeiriou, S., and Pantic, M. (2013). Markov random field structures for facial action unit intensity estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 738–745.

Saragih, J. and Göcke, R. (2009). Learning aam fitting through simulation. *Pattern Recognition*, 42(11):2628–2636.

Saragih, J. and Goecke, R. (2006). Iterative error bound minimisation for aam alignment. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 1196–1195. IEEE.

Saragih, J. and Goecke, R. (2007). A nonlinear discriminative approach to aam fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

Saragih, J. M., Lucey, S., and Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215.

Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133.

Sariyanidi, E., Gunes, H., Gökmen, M., and Cavallaro, A. (2013). Local zernike moment representation for facial affect recognition. In *BMVC*.

Schroff, F., Criminisi, A., and Zisserman, A. (2008). Object class segmentation using random forests. In *BMVC*, pages 1–10.

Shan, C., Gong, S., and McOwan, P. W. (2006). Dynamic facial expression recognition using a bayesian temporal manifold model. In *BMVC*, pages 297–306.

Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304. Ieee.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al. (2013). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840.

Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Sikka, K., Wu, T., Susskind, J., and Bartlett, M. (2012). Exploring bag of words architectures in the facial expression domain. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 250–259. Springer.

Singh, C., Mittal, N., and Walia, E. (2011). Face recognition using zernike and complex zernike moment features. *Pattern Recognition and Image Analysis*, 21(1):71–81.

Sumathi, C., Santhanam, T., and Mahadevi, M. (2012). Automatic facial expression analysis a survey. *International Journal of Computer Science and Engineering Survey*, 3(6):47.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*. MIT press.

Suwa, M. (1978). A preliminary note on pattern recognition of human emotional expression. In *Proc. of The 4th International Joint Conference on Pattern Recognition*, pages 408–410.

Tong, Y., Chen, J., and Ji, Q. (2008). A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):258–273.

Tong, Y., Chen, J., and Ji, Q. (2010). A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):258–273.

Tong, Y., Liao, W., and Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence*, 29(10).

Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869.

Tresadern, P. A., Ionita, M. C., and Cootes, T. F. (2012). Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision*, 96(3):280–289.

Tresadern, P. A., Sauer, P., and Cootes, T. F. (2010). Additive update predictors in active appearance models. In *BMVC*, volume 2, page 4. Citeseer.

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., and Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187.

Tuzel, O., Marks, T. K., and Tambe, S. (2016). Robust face alignment using a mixture of invariant experts. In *European Conference on Computer Vision*, pages 825–841. Springer.

Tzimiropoulos, G., Alabort-i Medina, J., Zafeiriou, S., and Pantic, M. (2012). Generic active appearance models revisited. In *Asian Conference on Computer Vision*, pages 650–663. Springer.

Tzimiropoulos, G. and Pantic, M. (2013). Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE international conference on computer vision*, pages 593–600.

Uřičař, M., Timofte, R., Rothe, R., Matas, J., et al. (2016). Structured output svm prediction of apparent age, gender and smile from deep features. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognision Workshop (CVPRW 2016)*, pages 730–738. IEEE.

Uricár, M., Timofte, R., Rothe, R., Matas, J., and Van Gool, L. (2016). Structured output svm prediction of apparent age, gender and smile from deep features. In *Proceedings CVPRW 2016*, pages 25–33.

Valstar, M., Martinez, B., Binefa, X., and Pantic, M. (2010). Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., and Scherer, K. (2011). The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE.

Valstar, M. F. and Pantic, M. (2012). Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Vukadinovic, D. and Pantic, M. (2005). Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1692–1698. IEEE.

Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., and Movellan, J. (2007). Drowsy driver detection through facial movement analysis. In *International Workshop on Human-Computer Interaction*, pages 6–18. Springer.

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., and Pantic, M. (2017). Deep structured learning for facial action unit intensity estimation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5709–5718. IEEE.

Wang, N., Gao, X., Tao, D., Yang, H., and Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65.

Wang, S., Wu, S., Gao, Z., and Ji, Q. (2016). Facial expression recognition through modeling age-related spatial patterns. *Multimedia Tools and Applications*, 75(7):3937–3954.

Whitehill, J., Bartlett, M., and Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE.

Woźniak, M., Graña, M., and Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17.

Wu, Y., Wang, Z., and Ji, Q. (2013). Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3452–3459. IEEE.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., and Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pages 57–72. Springer.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.

Yang, H. and Patras, I. (2013). Sieving regression forest votes for facial feature detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1936–1943.

Yang, H.-F., Lin, B.-Y., Chang, K.-Y., and Chen, C.-S. (2018). Joint estimation of age and expression by combining scattering and convolutional networks. *ACM TRANSACTIONS ON MULTIMEDIA COMPUTING COMMUNICATIONS AND APPLICATIONS*, 14(1).

Yang, P., Liu, Q., and Metaxas, D. N. (2009). Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1018–1025. Ieee.

Yang, S. and Bhanu, B. (2011). Facial expression recognition using emotion avatar image. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 866–871. IEEE.

Yu, Z. and Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM.

Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1 – 24.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928.

Zhao, K., Chu, W.-S., and Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399.

Zhong, Lind Liu, Q., Yang, P., Liu, B., Huang, J., and Metaxas, D. N. (2012). Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., and Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 386–391. IEEE.

Zhu, Y., Li, Y., Mu, G., and Guo, G. (2015). A study on apparent age estimation. In *Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on*, pages 267–273. IEEE.