

Facial Analytics for Emotional State Recognition

A DISSERTATION SUBMITTED TO THE CENTRE OF SIGNAL AND IMAGE PROCESSING, DEPARTMENT OF ELECTRONIC AND ELECTRICAL ENGINEERING AND THE COMMITTEE FOR POSTGRADUATE STUDIES OF THE UNIVERSITY OF STRATHCLYDE IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY

By

Konstantinos Papapazachariou

2017

Declaration

I declare that this Thesis embodies my own research work and that it is composed by myself. Where appropriate, I have made acknowledgments to the work of others.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by the University of Strathclyde Regulation 3.50. Due acknowledgment must always be made of the use of any material contained in, or derive from, this thesis.

Signed:

Date:

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof John J. Soraghan and Dr. Gaetano Di Caterina for their guidance, support and encouragement throughout my MPhil. They have always given me the right direction toward the exciting research areas in our field and the means to work independently with great freedom. Their unwavering enthusiasm for image processing kept me constantly engaged with my research. Without their mental and practical support to numerous personal concerns, I could not finish the work in this thesis.

I would like to thank my family, for their endless love, spiritual and material support. My parents were always believed in me and regardless of the distance, they were trying to encourage me. Without my family's constant support and understanding, it would not have been possible for me to achieve my educational goals.

I certainly be remiss to not mention and sincerely thank my many friends and colleagues in the Electronic & electrical engineering Centre for Signal & Image Processing (CeSIP) at the University of Strathclyde where it has been my pleasure and honor to study for the past year.

I dedicate this thesis to my parents, my brother and my girlfriend.

Abstract

For more than 75 years, social scientists study the human emotions. Whereas numerous theories developed about the provenance and number of basic emotions, most agreed that they could categorize into six categories: anger, disgust, fear, joy, sadness and surprise. To evaluate emotions, psychologists focused their research in facial expressions analysis. In recent years, the progress in digital technologies field has steered the researchers in psychology, computer science, linguistics, neuroscience, and related disciplines towards the usage of computer systems that analyze and detect the human emotions. Usually, these algorithms are referred in the literature as facial emotion recognition (FER) systems. In this thesis, two different approaches are described and evaluated in order to recognize the six basic emotions automatically from still images.

An effective face detection scheme, based on color techniques and the well-known Viola and Jones (VJ) algorithm is proposed for the face and facial characteristics localization within an image. A novel algorithm which exploits the eyes' centers coordinates, is applied on the image to align the detected face. In order to reduce the effects of illumination, homomorphic filtering is applied on the face area. Three regions (mouth, eyes and glabella) are localized and further processed for texture analysis.

Although many methods have been proposed in the literature to recognize the emotion from the human face, they are not designed to be able to handle partial occlusions and multiple faces. Therefore, a novel algorithm that extracts information through texture analysis, from each region of interest, is evaluated. Two popular techniques (histograms of oriented gradients and local binary patterns) are utilized to perform texture analysis in the abovementioned facial patches. By evaluating several combinations of their principal parameters and two classification techniques (support vector machine and linear discriminant analysis), three classifiers are proposed. These three models are enabled depending on the regions' availability. Although both classification approaches have shown impressive results, LDA proved to be slightly better especially regarding the amount of data management. Therefore, the final models, which utilized for comparison purpose, were trained using LDA classification.

Experiments using Cohn-Kanade plus (CK+) and Amsterdam Dynamic Facial Expression Set (ADFES) datasets demonstrate that the presented FER algorithm has surpassed other significant FER systems in terms of processing time and accuracy. The evaluation of the system involved three experiments: intra-testing experiment (train and test with the same dataset), train/test process between CK+ and ADFES and finally the development of a new database based on selfie-photos, which is tested on the pre-trained models. The last two experiments constitute a certain evidence that Emotion Recognition System (ERS) can operate under various pose and light circumstances.

List of Abbreviations

AAM:	Active appearance model
ADFES:	Amsterdam Dynamic Facial Expression Set
ANN:	Artificial neural network
ASM:	Active shape model
AU:	Action units
CK:	Cohn-Kanade
CK+:	Cohn-Kanade plus
CVPR:	Computer Vision and Pattern Recognition
DCT:	Discrete cosine transform
EG:	Eyes-glabella
ERS:	Emotion recognition system
FACS:	Facial action coding system
FER:	Facial expression recognition
FNN:	Feedforward neural network
HMM:	Hidden Markov models
HOG:	Histograms of oriented gradients
ID:	Iterative-Discriminative
LBP:	Local binary patterns
LDA:	Linear discriminant analysis
M:	Mouth
MAP:	Maximum a posteriori
MEG:	Mouth-eyes-glabella
MLP:	Multi-Layer Perceptron
NB:	Naive Bayes
NN:	Neural network
PCA:	Principal component analysis
RBF:	Radial Basis Function
RGB:	Red-green-blue
RNN:	Recurrent neural network
ROI:	Region of interest
SVM:	Support vector machines
VJ:	Viola and Jones algorithm
VLBP:	Volume local binary patterns

Table of Contents

Declaration.....	ii
Acknowledgment.....	iii
Abstract.....	iv
List of Abbreviations	v
Table of Contents	vi
List of Figures.....	viii
List of Tables	xii
Chapter 1	14
Introduction.....	14
1.1 Preface.....	14
1.2 Motivation of Our Research.....	15
1.3 Summary of Original Contributions.....	16
1.4 Organization of the Thesis	16
Chapter 2	18
Emotion Recognition in Social and Computer Science.....	18
2.1 Introduction	18
2.2 Social science	19
2.2.1 Sociology and psychology of emotion.....	19
2.2.2 Facial expression of emotion	21
2.2.3 Primary emotions	22
2.2.4 Facial action coding system	25
2.3 Face datasets for expression recognition.....	26
2.4 Face detection and facial features localization.....	28
2.4.1 Knowledge-based methods	28
2.4.2 Feature-invariant methods	30
2.4.3 Template matching methods	31
2.4.4 Appearance-based methods	32
2.5 Facial feature extraction	36
2.5.1 Geometric feature-based approaches	38
2.5.2 Appearance feature-based approaches	40
2.6 Facial expression classification and recognition.....	47
2.6.1 Template matching.....	48
2.6.2 Artificial Neural Networks	48
2.6.3 Naïve Bayes	49

2.6.4	Classification trees	50
2.6.5	Support vector machine	51
2.6.6	Linear discriminant analysis	53
2.7	Summary	54
Chapter 3		55
Emotion Recognition System (ERS).....		55
3.1	Introduction	55
3.2	Algorithm overview	56
3.3	Face and salient regions detection.....	58
3.3.1	Face detection based on skin color	58
3.3.2	Face detection based on VJ algorithm	59
3.3.3	Salient regions detection	62
3.4	ERS using HOG descriptor	66
3.4.1	HOG Descriptor’s parameters assessment.....	67
3.4.2	Classification parameters assessment	70
3.4.3	HOG training and testing using the CK+ database	71
3.4.4	Generalization of the algorithm	79
3.4.5	SelfieDat: A new database	80
3.4.6	A comparison with state-of-art	83
3.5	ERS using LBP descriptor.....	84
3.5.1	LBP Descriptor’s parameters assessment	85
3.5.2	LBP training and testing using the CK+ database	88
3.5.3	ADFES and SelfieDat	95
3.5.4	Comparison of the performance.....	97
3.6	Summary	98
Chapter 4		99
Conclusion and Future Work.....		99
4.1	Conclusions	99
4.2	Future Work	100
Appendix A: The SelfieDat database		102
References.....		104

List of Figures

Figure 2.1: Sociological Analysis of Emotion [13].	20
Figure 2.2: (a) Plutchik's emotion wheel , (b) Plutchik's cone-shaped model [44, 45].	23
Figure 2.3: Example of the six basic emotion expressions. (a) Anger; (b) Disgust; (c) fear; (d) Joy; (e) Sadness and (f) Surprise [52].	24
Figure 2.4: Russell's circumplex model of emotion. Their level of arousal is demonstrated on the vertical axis and their valence on the horizontal axis.	24
Figure 2.5: FER system pipeline.	28
Figure 2.6: An example of a rule, which applied in the first step. Uniform gray level presented at the top of the head (light shady area).	29
Figure 2.7: The eye-analogue segmentation [68]. (a) The eye-analogue pixels and segments, (b) The eye-analogue segments at the labeling stage	30
Figure 2.8: Features' marking process. (a) Labeled training image; (b) The landmark points and (c) Shape-free patch of the face [77].	32
Figure 2.9: Eigenfaces production process [84, 85]. The original 100 faces are "unfolded" into each of the vectors and then PCA applied to the whole array to produce the 36 eigenvectors.	33
Figure 2.10: Example of five Haar-like patterns used in Viola-Jones algorithm. The size and position of a pattern can change while its black and white rectangles keep their ratio the same. In (a) and (c) two-rectangle patterns are presented, in (b) and (c) three-rectangle are presented and (e) a four-rectangle.	34
Figure 2.11: The four array references [92]. The rectangle A's sum of pixels represented by the integral image at the point 1. At the point 2 the value is A+B, at the point 3 is A+C and at the point 4 is A+B+C+D. Thus, the sum of the pixels in D is $4 + 1 - (2 + 3)$.	35
Figure 2.12: Cascade Classifier Scheme [92]. Every sub-window pass through a series of classifiers. The first classifier rejects a large number of examples without limited processing. The next tiers discard more examples until the number of sub-windows sufficiently reduced. Further processing can be further stages of the cascade or a different detection system.	36
Figure 2.13: Fiducial facial points [99]. The letters A-M represent the points on the edges of the facial characteristics.	38
Figure 2.14: ASM Facial Features from Shbib et al. [100].	39

Figure 2.15: The linear appearance variation of an independent AAM [103]. The face in (a) represents the base (i.e. initial face) and (b-d) are represent a linear combination of m appearance images on the same pixel area.40

Figure 2.16: Complex Gabor filter ($f = \frac{\pi}{3}$, $\theta = 0$ and $c_1 = c_2 = \pi$) [108]. (a) The real part's graphic representation and (b) the imaginary part's graphic representation.....41

Figure 2.17: The simplest case of LBP in 3x3 pixels neighborhood [113]. (a) The initial image, the center pixel's value used as a threshold, (b) The thresholding image, (c) Weights of each pixel, (d) The resulted image after the LBP implementation.....42

Figure 2.18: Gradient Vector, Magnitude, and Angle.44

Figure 2.19: Example of the HOG descriptor of 8x8 cell size and 9 orientation bins operation on a 237x264 image. (a) HOG visualization, (b) Histogram of gradients.....45

Figure 2.20: Classification Algorithm47

Figure 2.21: Neural Networks classification for six basic emotions (AN: Anger, DI: Disgust, FE: Fear, HA: Happy, SA: Sad and SU: Surprise). (a) A full-face approach, (b) A feature-based approach.49

Figure 2.22: Classification Tree example. Three emotions and the neutral state are classified using a decision tree classifier with five nodes and 11 branches.51

Figure 2.23: Simple two-class classification example of Support Vector Machine. In this example an optimal linear discrimination function maximizes the margin between the data of the two separated classes. The optimal hyperplane explained by the function of the support vectors.52

Figure 2.24: Simple two-class classification example of Linear Discriminant Analysis. In this example, LDA is used to find an optimal linear model that best separates two classes.54

Figure 3.1: Schematic overview of the proposed framework.55

Figure 3.2: Block diagram of the ERS algorithm.57

Figure 3.3: Skin-color-based method applied successfully on a face image. (a) The original image, (b) Image after the skin filtering, (c) Gray-scaled image, (d) Original image with bounding box around the face region.59

Figure 3.4: Example of face detection algorithm..61

Figure 3.5: Face alignment example.....62

Figure 3.6: Homomorphic filtering.....64

Figure 3.7: Example of Homomorphic filtering. (a) The original face image. (b) The face image after the homomorphic filtering application..64

Figure 3.8: Salient regions in bounding boxes. (a) Face from the ADFES database, (b) Face from the CK+ database.....	65
Figure 3.9: Salient regions detection with VJ algorithm. (a) Two individuals from CK+ and ADFES databases with three emotions each: joy, anger and sadness, (b) Mouth region detection, (c) Left and right eyes (including the eyebrow) region detection and (d) Glabella region detection.....	66
Figure 3.10: Feature extraction using HOG descriptor (cell size: 8, orientation bins: 9) in three steps: Input image, HOG visualization and concatenated histogram of cells. (a) Joy emotion, (b) Surprise emotion and (c) Disgust emotion.	69
Figure 3.11: Schematic example of HOG extraction in the right eye.....	70
Figure 3.12: Mouth region. Average recognition rates evaluation for all the possible combinations of HOG descriptor's parameters with five PCA variance levels. Parameters are placed based on feature vector's size (min – max).	73
Figure 3.13: Mouth region's confusion matrix.	74
Figure 3.14: Eyes region. Average recognition rates evaluation for all the possible combinations of HOG descriptor's parameters with five PCA variance levels. Parameters are placed based on feature vector's size (min – max).	74
Figure 3.15: Eyes regions' confusion matrix.	75
Figure 3.16: Glabella region. Average recognition rates evaluation for all the possible combinations of HOG descriptor's parameters with five PCA variance levels. Parameters are placed based on feature vector's size (min – max).	75
Figure 3.17: Glabella region's confusion matrix.	76
Figure 3.18: EG-classifier confusion matrix.....	78
Figure 3.19: MEG-classifier confusion matrix.	78
Figure 3.20: Sample of the participants in selfie-experiment. (a) Anger, (b) Disgust, (c) Joy, (d) Sadness and (e) Surprise.	81
Figure 3.21: The disgust emotion (samples from SelfieDat). In the left image, the expression is more intensive (i.e. mouth is open and frowning) while in the right image the mouth is closed.	82
Figure 3.22: Feature extraction using LBP descriptor (cell size: 16, number of neighbors: 8) in two steps: Input image and concatenated histogram of cells. (a) Joy emotion, (b) Surprise emotion and (c) Disgust emotion. ..	87
Figure 3.23: Schematic example of LBP extraction in the right eye.	88

Figure 3.24: Mouth region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max). 90

Figure 3.25: Mouth region’s confusion matrix. 90

Figure 3.26: Eyes region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max). 91

Figure 3.27: Eyes regions’ confusion matrix. 91

Figure 3.28: Eyes region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max). 92

Figure 3.29: Glabella’s region’s confusion matrix. 92

Figure 3.30: EG-classifier confusion matrix. 94

Figure 3.31: MEG-classifier confusion matrix. 94

Figure A.1: The anger emotion in SelfieDat. 94

Figure A.2: The disgust emotion in SelfieDat. 94

Figure A.3: The fear emotion in SelfieDat. 942

Figure A.4: The joy emotion in SelfieDat. 942

Figure A.5: The sadness emotion in SelfieDat. 942

Figure A.6: The surprise emotion in SelfieDat. 943

List of Tables

Table 2.1: Specimen from FACS action units [53].	25
Table 2.2: Publicly available facial expression databases.	27
Table 2.3: Overview of significant methods that have been used in features extraction process. (Note: In the category column, G: Geometry, A: Appearance. The discrimination depends on which of these categories these methods were mostly used).	37
Table 2.4: Overview of significant classification methods were used in FER systems.	48
Table 3.1: The Face Detection algorithm of ERS	60
Table 3.2: Regions size – Emotion caused the maximum deformation.	65
Table 3.3: Number of feature for the mouth area based on cell size and orientation bins combinations.	68
Table 3.4: Number of feature for the eyes area (including both eyes) based on cell size and orientation bins combinations.	68
Table 3.5: Number of feature for the glabella area based on cell size and orientation bins combinations.	68
Table 3.6: Three models from glabella region are evaluated. Training was performed for 330 images of CK+ database. Testing was performed on one image of CK+ database. Train and test columns contain computational time in seconds.	71
Table 3.7: Number of images per emotion.	72
Table 3.8: Best accuracy of each region for LDA and SVM. C and B represent the cell size and orientation bins respectively.	73
Table 3.9: Best accuracy of each model using LDA and SVM.	77
Table 3.10: Average recognition accuracy for each emotion training CK+ and testing ADFES.	79
Table 3.11: Average recognition accuracy for each emotion training ADFES and testing CK+.	80
Table 3.12: Average recognition accuracy for each emotion training CK+ and testing SelfieDat.	82
Table 3.13: Average recognition accuracy for each emotion for M and EG classifiers.	83
Table 3.14: Comparison with the state-of-art methods for six basic emotions. Bold indicates the best results.	84
Table 3.15: Number of feature for the mouth area based on cell size and number of neighbors combinations.	86

Table 3.16: Number of feature for the eyes area based on cell size and number of neighbors combinations.86

Table 3.17: Number of feature for the glabella area based on cell size and number of neighbors combinations.86

Table 3.18: Best accuracy of each region for LDA and SVM. C and N represent the cell size and number of neighbors respectively.89

Table 3.19: Best accuracy of each model for LDA and SVM.93

Table 3.20: Average recognition accuracy for each emotion training CK+ and testing ADFES.95

Table 3.21: Average recognition accuracy for each emotion training ADFES and testing CK+.96

Table 3.22: Average recognition accuracy for each emotion. The results obtained using MEG-classifier model.....96

Table 3.23: Average recognition accuracy for each emotion. The results obtained using M-classifier and EG-classifier models. The second column shows the accuracy for M-classifier, while the third column presents the accuracy for EG-classifier.....97

Table 3.24: Comparison with the state-of-art methods for six basic emotions. Bold indicates the best results.....98

Chapter 1

Introduction

1.1 Preface

The human face provides a valuable and effective source of information about human behavior and maintains a manifestation of the affective state, cognitive activity, intention, personality, and psychopathology of a person. Moreover, the human face has always been appearing to be an active research subject in various scientific sectors (computer vision and graphics, pattern recognition and social science). Face recognition and facial expression analysis have been studied the past years extensively, resulting in remarkable application systems. Research findings in image processing and machine learning facilitate and lead recent studies into more accurate and robust results. However, several issues that triggered by the conditions imposed by many image analysis applications remain unsolved. For example, several emotion recognition systems require the appropriate light and pose conditions. They are still incapable of analyzing specific regions of the face and are insufficient to offer same time a solution for the emotion recognition under partial occlusion.

Emotion analysis is a timeless subject of study. Although scientists started studying different aspects of human emotions back in 19th century, the interest nowadays does not subside. There are still numerous open questions that should be given reliable answers. Despite the difference in social and computer science, the integration of knowledge from both these fields can improve the ability to understand the emotion and other non-verbal mechanisms. Many objective approaches have been proposed for the automatic emotion recognition in the past decade. They involved various techniques such as Gabor feature and active shape models (ASM), which in the most cases were time-consuming.

In this thesis, an automated, efficient and novel facial emotion recognition system is presented. The image is fed into the system and the possible faces are detected automatically. The face is aligned and homomorphic filtering is applied to reduce the effect of illumination within the image. The system identifies three regions of interest (ROIs) mouth, eyes and glabella (the area between the eyebrows). Two texture analysis techniques, Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP) are applied into the ROIs to provide regional and overall assessment for each of the six basic emotions. Classification performed based on two different methods, Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA).

Experimental results based on subjects from two well-known datasets demonstrate that the proposed methods can be effectively used for the assessments of facial emotion analysis. Furthermore, in order to prove the stability of the proposed system, a new dataset created with selfie images were taken under random circumstances, was tested. The results were impressive especially when tested with the classifiers were trained based on the HOG approach.

1.2 Motivation of Our Research

One of the most significant aspects of everyday life is communication, verbal or non-verbal. There are several non-verbal communication means including body posture, pitch, speed, tone and volume of voice, gestures and facial expressions, stance, and proximity to the listener, eye movements and contact, and dress and appearance. According to the social science [1-3], the facial motor expression component of emotion has the greatest impact on communication between two individuals. Expressions vary in terms of intensity and duration, and they can change rapidly during social interaction. Thus, they can provide powerful means in order to communicate ultimately without the use of linguistics.

Nowadays computer science research is focused on human-centered applications that attempt to describe numerous functions of the human body and soul. The task of emotion detection involves image and speech processing and pattern recognition. It refers to methods that aim to identify emotions in images, videos or speech that can be used for high-level analysis, such as, behavioral understanding. Human emotion recognition using image processing approaches, usually studies facial expressions' unique characteristics. In general facial expression analysis within a face image depending on the acquisition device (i.e. camera) and the techniques are chosen. In general, the well-known methodology includes texture or geometric methods, generic algorithms and machine learning. Nowadays, artificial intelligence (AI) research focuses on the promising field of Deep Learning. Deep learning is a class of machine learning that involves supervised and unsupervised algorithms for classification and pattern analysis respectively. This field has advanced significantly over the years due to works of great scientists like Andrew Ng, Geoff Hinton, Yann LeCun, Adam Gibson, and Andrej Karpathy [4-6].

The application of the challenging task of emotion recognition has attracted much attention in recent years. For example, emotion transferring to 3-D games or film avatars is used to create characters' behavior and automatic control their reactions. On the other hand, emotion recognition can have very specific uses every day, such as job interviews and crimes investigations. Usually, the listener is responsible for judging based on his experience, sometimes resulting to false appraisal. Therefore, automatic facial expression recognition (FER) could become a vital tool for emotions decoding. Due to its potential applications, such as social security, entertainment, medical care and human-computer interface, FER may assist scientists to obtain the meaningful information from the large-scale data and speed up the manual task of processing data on human affective behavior. However, problems such as the illumination variations, image resolution and pose affect computer based systems' reliability in FER.

1.3 Summary of Original Contributions

The main research contributions of this thesis are described below.

1. The first contribution is represented by a novel face detection method that combines two well-known algorithms [7, 8], decreasing the probability of a false or misdetection significantly. Many face detection algorithms have been proposed in the literature as reported in [Section 2.4](#). Viola and Jones algorithm has been utilized in the past for various detection purposes such as humans, faces or items and is known for its simplicity and processing time. However, in many cases due to light condition and occlusions is possible to be compromised. For this reason, the detection scheme of this project has been supported by color-based face detection approach, developing a robust algorithm that eliminates any false detections. Moreover, the face detection algorithm is designed to manage images that include more than one person. Thus, by evaluating multiple faces can be at same time, the final system is computationally and time efficient ([Chapter 3 – Section 3.3.1](#), [Chapter 3 – Section 3.3.2](#)).
2. The second contribution is represented by a novel emotion recognition algorithm that confronts the partial occlusion issues by evaluating only the available face regions. Recently different methods for automatic emotion recognition have been proposed [9-13] but none of them took into account possible occlusions might arise on the human face. For example, a person who is surprised and covers his mouth could cause lead to a false detection or worse to misdetection. The improved ERS algorithm can effectively utilize the available regions and evaluate the emotion based on three pre-trained classifiers: mouth-eyes-glabella (MEG), mouth (M) and eyes-glabella (EG) ([Chapter 3 – Section 3.3.3](#)).
3. The last contribution is represented by a detailed examination that performed for two texture descriptors, HOG and LBP. In order to determine what combination of parameters is the most effective in facial expression analysis, extensive experiments were carried out, demonstrating the pros and cons of those two techniques. The latter, can help other researchers to understand the capability of those methods, not only in facial expression analysis but also in other relative projects ([Chapter 3 – Section 3.4](#), [Chapter 3 – Section 3.5](#)).

1.4 Organization of the Thesis

This thesis is organized as follows.

Chapter 1 describes the objective and motivation for the research, along with a brief summary of the original contributions that are presented in this work.

Chapter 2 initially reviews the various theories developed regarding the human emotion from social science. In order to develop a robust automatic application, all the available sources theoretical and not, have to be evaluated. In emotion analysis, the contribution of social science is endless. Findings in fields of psychology and sociology are presented. Furthermore, the relationship of the emotion with the facial expressions and the well-known facial action coding system are discussed. Moreover, face databases that have been most used in the literature are reviewed. Then, by following all the required steps to develop an emotion recognition algorithm, facial image processing techniques are discussed. Firstly, face and facial feature detection methods that have been used in past research are reported and evaluated. Secondly, feature extraction approaches are analyzed, and finally, classification methods are reviewed.

Chapter 3 initially introduces a novel algorithm for multiple face detection in an image. Then, two texture analysis methods (HOG and LBP), as well as two classification methods (SVM and LDA), are applied to three different databases. The experimental process is described and results are compared with other state-of-the-art methods.

Overall conclusions for this thesis are reported in Chapter 4, along with suggestions for possible future work.

It is worth mentioned that since judging performance is difficult except in the context of a particular application, performance is generally characterized using more subjective/comparative terms such as “good”, “poor”, “excellent” “exceptional” and “adequate”. It is acknowledged that these have no objective status in themselves, but should help to give a sense of the comparative merits of different systems and parameters.

Chapter 2

Emotion Recognition in Social and Computer Science

2.1 Introduction

Human emotion has numerous social and physical aspects [14] including psychophysiology, neural correlates, development, perception, addiction, and depression. Facial gestures can convey pain, alertness, personality and interpersonal relations [15]. Facial expressions of emotion research, has its roots in social science, although several fields such as computer science and AI have shown an increased interest. Recent years have seen an expeditious development of algorithms, which automatically recognize the facial expressions of emotion. In computer science, there is a keen interest in applications that can analyze the facial expressions for various reasons such as marketing, human-computer interaction and fatigue monitoring. U.S. National Science Foundation [16] and IEEE community [17] have sponsored several conferences and workshops related to this highly-prospective research area. As it mention earlier in Section 1.2, nowadays there is a tendency to replace traditional image representation techniques with more efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction (i.e. Deep Learning). However, Deep Learning includes several known architectures and as a result, it is important to determine the nature of the machine-learning problem before it is applied. Considering the aims of this research, there are four reasons deep learning is unsuitable:

- 1) Large amount of data – Deep learning requires an extremely large amount of data for training purposes otherwise is unlikely to outperform other approaches.
- 2) Is computational expensive to train – Complex models can take weeks to be trained using several pc with very expensive Graphic Processing Units (GPUs).
- 3) Neural networks are not probabilistic – Unlike their statistic or Bayesian counterparts, NNs do not provide much information about the confidence of the classifier. Therefore is difficult to manage the cost of making errors and tune the classifier properly.
- 4) Deep learning is not a substitute for understanding the problem deeply – Essentially one of the most important aim of this research is to determine the correlation between the facial patches and their lower level features. By using traditional machine-learning techniques there is a great chance to select and create the best input features.

In this research, human emotions are studied from the computer science and image processing perspective, although social and cognitive science was a guide in every step. In order to create an automatic facial expression recognition system, the first step was to understand how humans decrypt facial expressions. For this reason, Section 2.2 provides an overview of emotion studies from the social and cognition science

literature. In Section 2.3 face databases characteristics briefly described and some of the most known databases are compared. In Section 2.4 face detection and feature localization techniques are briefly discussed. In Section 2.5, feature extraction approaches are analyzed based on geometry or appearance. Finally, in Section 2.6 the last part of an FER system, the classification of emotions is discussed. Significant classification methods are presented and their operation is described.

2.2 Social science

According to Turner [18], humans are highly emotional animals. During the species evolution, hominids and specifically *Homo sapiens* increased the demonstration of their feelings. Later on, societies were evolved depended on the emotional bonds among people. Thus, history has shown that in the name of love or hate, critical incidents were allowed to happen. For example, famous monuments were built for religious devotion and crimes were committed because of the human greediness or jealousy. In recent decades, scientists have had a keen interested in human emotions and various theories have developed. Initially, social science and specifically its branches of psychology, sociology, and anthropology were the first academic disciplines concerned with human emotions. Later on, scientists from various domains such as cognition, neuroscience and computer science became interested in the emotion analysis.

2.2.1 Sociology and psychology of emotion

Although numerous theories developed about the meaning of emotion, the topic is still elusive. In 1981, Klenginna and Klenginna [19] counted more than one hundred definitions. Several scientific fields contributed in emotions' practice, with sociology and psychology, were considered as the principal research areas.

The first sociologists in the 19th century with few exceptions were unsuccessful in describing human emotions, and their findings were lacking in detail. For instance, Marx [20], gave priority to conflicting emotions of alienation and anger while Durkheim [21] determined the negative emotions of egoism and anomie. In the 1970s modern sociology, recognize the multilateralism of human emotions [22-24]. There are five areas (Fig. 2.1) that constitute the backbone of the sociological study of emotions: the dramaturgical (culture), structural (social structure), symbolic-interactionist (cognitive appraisal) and ritual and exchange (interaction) perspectives on emotion [25]. Dramaturgical meaning investigates the cultural nature of emotion. According to Gordon [26] and Peterson [27], aspects such as ideologies, vocabularies, knowledge and feeling rules perform as a cognitive protocol in humans, lead them to conform their emotions depending on the situation they experience.

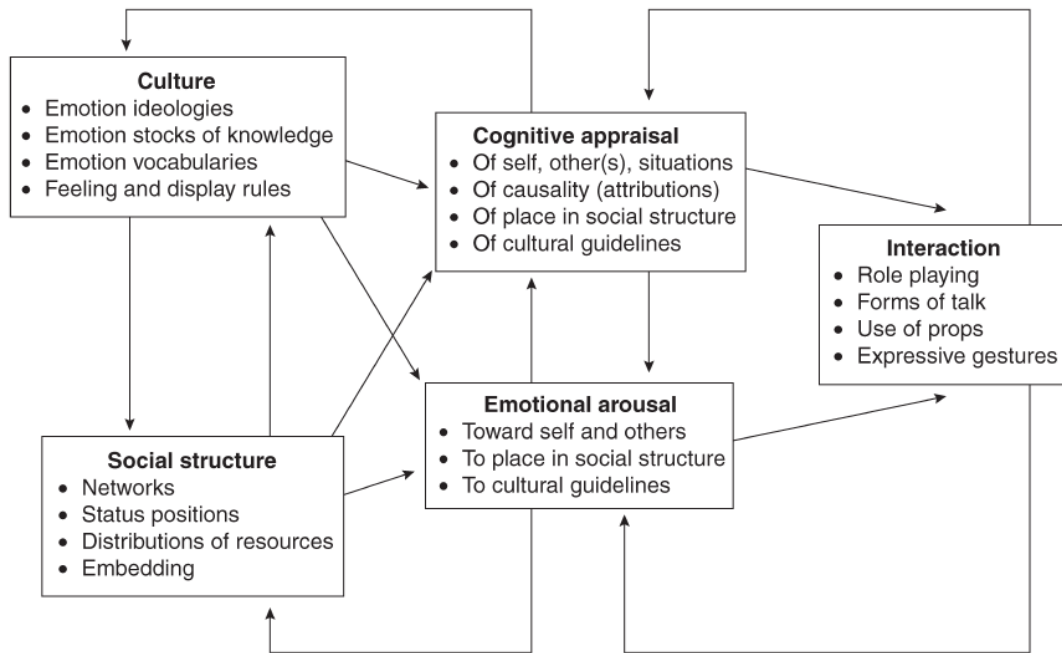


Figure 2.1: Sociological Analysis of Emotion [25].

The structural approach will be followed in every sociological examination of emotion. The properties that constitute the social structure always had a keen interest in sociology. In the microstructural analysis, the concepts of power and status variation among individuals are examined. Moreover, it determines the level of influence these characteristics have in peoples' decisions and how they recreate or alter the power and status order. Hence, the interaction between individuals exposes essential features of social structures and describes how human emotions are established within them. For example, Houser and Lovaglia [28] investigated emotion from the perspective of peoples' social-ranking. They demonstrated that higher-ranking individuals would feel satisfaction, happiness, and pride when their rank was identified by their actions. On the other hand, if they would experience status uncertainty, they will be sensitive to negative emotions such as fear or anxiety. In symbolic interactionist approaches focus on the person's essential being (self) and identities during the emotion occurrence.

Essentially using cognitive appraisals, individuals build their self-meanings and identity principles, collect information about who they are in situations, evaluate the level of agreement with their self-meanings and experience emotional arousals [29]. Ritual theories refer to the human need to participate in religious or solemn ceremonies. E. Durkheim [21] analyze the "collective effervescence" in Australian aboriginals society. Their repeated gatherings during the year produce a collective effervescence of positive emotions. Australian aboriginals like other religious or not groups related their positive emotions with symbols and materials, which they subsequently worshiped. Exchange approaches are based on the peoples' participation in trading. The action of buying and selling goods and services generate positive or negative emotions depending on the profit or loss. Exchanges can be classified into four categories [30]: "productive", which refers to a common model of individuals' contribution and almost same rewards; "negotiated", where people are making offers until they will agree on the value of services or goods; "reciprocal", where a person invests the resources for future rewards; and "generalized", where the resources cyclically provided from one

person to another. Molm [31] has shown that negotiated exchanges usually involve disagreements during the bargain and consequently kindle emotions of anger and irritation.

Psychology considers the features and patterns of phenomena that qualify as “emotional” and analyzes them regarding fundamental processes [32]. The analysis of an individual’s emotion results after a comprehensive collection of information on various aspects of his life. Environment and body mechanisms along with experiences, cognitive schemas, and behavioral skills could constitute a source of information for psychologists to explain the human emotion. For example, Russell’s [33] research focused on pleasure-pain and activation features while Ekman’s [34] explanation was based on nervous system and culture elements.

Many theorists have adopted, and others denied the concept of specificity of emotion. One of the first psychologists who dealt with emotion analysis was James [35]. In his approach, he did not accept the specific nature of emotion. In particular, he assumed that emotion stems from the cerebral cortex like all the other behavior mechanisms. However, Cannon [36] refuted it during his research on subcortical mechanisms. A conflict also raised by the unity nature of emotion, with DeDoux [37] in the recent past to support that several emotions do not engage common mechanisms.

As mentioned at the beginning of this section there are various definitions about human emotion. Some of them are based on feelings especially that of pleasure or pain [32], which in their turn belong to a broader domain, that of “appraisal”. According to the Appraisal Theory, emotion is believed to be extracted from the conscious evaluation of events [38, 39]. In other words, emotion occurs from the explanation of actions, sometimes without psychological arousal [40]. Particularly important and commonly accepted was Scherer’s [41] interpretation, who considered not emotion alone but streams of processes that correspond to specific components. He defined emotion as “an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism”.

The distinction of emotions is another topic that preoccupied the scientific community for years. Initially, a verbal classification has been used, with some emotion labels to elucidate basic human feelings such as joy or anger and others labels to represent combinations of them [42]. In particular, feelings could be demonstrated through actions and bodily expressions [34, 42, 43], motivational states [44] and various combinations of psychological reactions [35]. It is common body movements to co-occur with specific emotions. The last is a frequent phenomenon in facial expression of emotion.

2.2.2 Facial expression of emotion

Facial expressions are some of the most common non-verbal ways for people to express their feelings. Argyle [45] distinguished the face from other bodily signals for the amount of information it conveys. Facial expression research initiated at the end of 19th century by Darwin [46]. In his work in 1872, he included both humans and animals, while he was able to use a global spectrum of data gathered from many different places in the world. The purpose of Darwin’s research was to demonstrate that there is a set of facial expressions humans use on a universal basis and his methods had

a significant impact on later studies. Despite the fact that almost 100 years passed after Darwin's work, the facial expression of emotion became a trend in scientific society in the 20th century, with many studies to corroborate or refute it.

Although several theorists viewed facial expression either as a communication signal or emotional arousal, Ekman, [47] considered false and misled the above dichotomy. According to his theory, both communication and emotion, are directly related and occur simultaneously during the facial expression production. Later on, he defined [48] that facial expressions, constitute a visible form of expressing feelings and cultural notions which in their turn contain the human emotions. Thus, people can develop "emotional and social familiarity". Humans develop their ability to identify and imitate expressions from a very young age [49], and this is helpful to establish social interaction and effective interpersonal relationships. However, not everyone has the same capability to recognize facial expressions. Masten [50] demonstrated that maltreated children have better and faster understanding of facial emotions and this is the result of their abuse. Another important factor in recognition of emotions is the intercultural adjustment. Yoo [51] determined that international students in the United States have difficulties in identifying specific emotions.

In a broader manner, facial expression can be considered as a result of emotional arousal or as consequence of the facial muscles activity. These concepts were explained extensively based on two fundamental approaches: measuring judgments regarding the message and measuring sign vehicles that transmit the message [52]. Message-judgement approaches study emotion as a cause of a specific stimulus. In other words, the observer has to make assumptions about the reasons the facial expression produced (i.e. emotion, mood or attitude). Based on message-judgements approach, Ekman [1, 34] and Izard [53] concluded that there are six basic emotions (joy, surprise, anger, fear, disgust, sadness) and every other emotion is produced from various combinations of them. The sign-based approach, focus on the facial surface's activity. For example, the change of the shape of facial characteristics and head pose during the affective state. Based on this approach P. Ekman established the well-known in behavior research, Facial Action Coding System (FACS) [54] which is explained in greater detailed in Section 2.2.4.

2.2.3 Primary emotions

During the human evolution, emotions nature has changed in neuroanatomical terms. Essentially humans were able to mix primary emotions to produce new ones. In his research, Plutchik [55] suggested a groundbreaking geometrical view of how emotions are combined to construct new ones. He believed that there are eight primary emotions which acting as opposite pairs: anger vs. fear; anticipation vs. surprise; joy vs. sadness and trust vs. disgust. According to his theory, the primary emotions are similar to the primary colors and the mix of them can produce other types of emotions. Therefore, he introduced an outline (see Fig. 2.2) within it, emotions vary regarding position and color intense. The 2-D model, which is divided into eight zones representing the basic

emotions, focus on showing the opposites. On the other hand, the 3-D model represents better the graduation of the emotion's intensity.

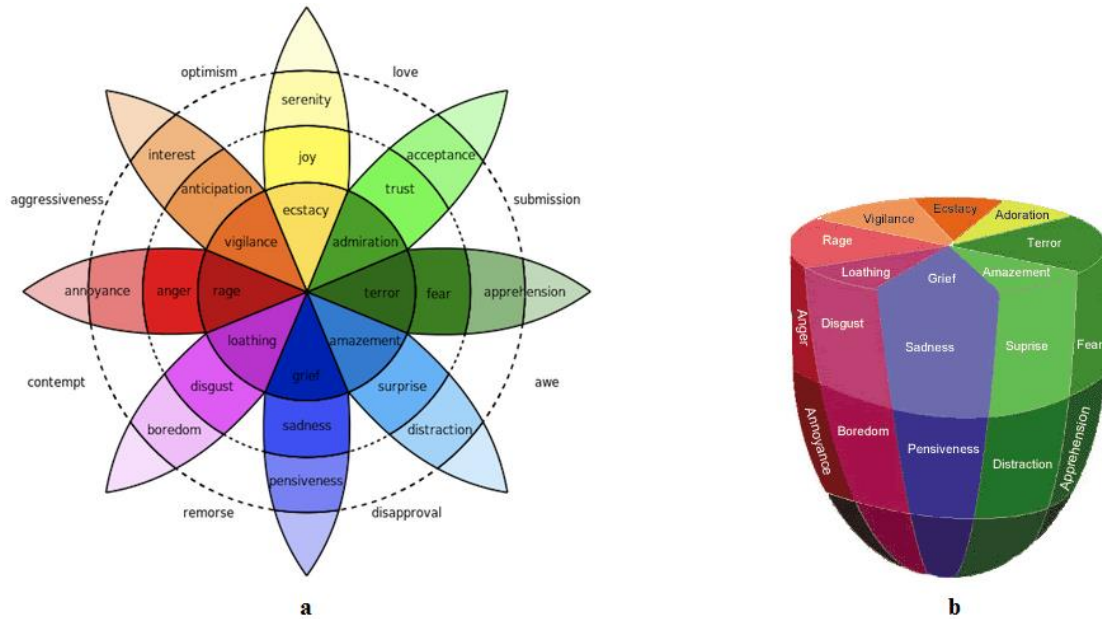


Figure 2.2: (a) Plutchik's emotion wheel , (b) Plutchik's cone-shaped model [56, 57].

Based on the message-judgement approach described in Section 2.2.2, Ekman [58] developed his theory where he suggested that there are six basic emotions interrelated with six facial expressions. In particular, these emotions were anger, disgust, fear, joy, sadness and surprise (see Fig. 2.3 for the depiction of these expressions). Due to their common use in every culture across the world, these emotions were characterized “universal” first by Darwin [46] and later on by Ekman [58], continuer of Darwin's research. Most notably because of its coherence, Ekman's work was adapted for many emotion detection and recognition projects in computer science.

However many researchers, especially from the cognitive science field, expressed their disagreement about the discrimination of emotions [3, 59]. They advocated that emotions should not have separated meanings since they are consistently interrelated [60, 61]. Particularly interesting was Russell [62] explanation who designed a system which has two axes to represent the arousal and valence. Arousal measured in the vertical axis where the endpoints represent the emotion activity from mild to intense. On the other hand, valence is unfolded on the horizontal axis measuring the degree of happiness from unpleasant to pleasant. Depending on those two characteristics, emotions are presented into a 2-Dimensional disc (see Fig. 2.4). Although Russell's model has described human emotions in depth, it never mentioned how the facial expressions are related to them. Thus, it is not applicable to an automatic recognition system.



Figure 2.3: Example of the six basic emotion expressions. (a) Anger; (b) Disgust; (c) fear; (d) Joy; (e) Sadness and (f) Surprise [63].

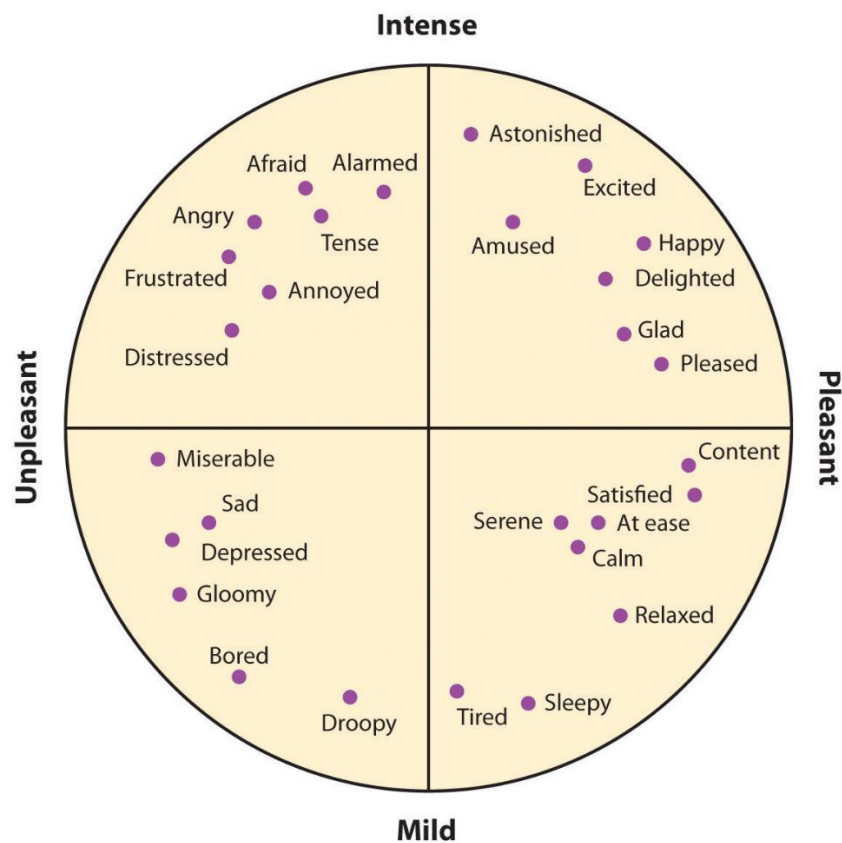


Figure 2.4: Russell's circumplex model of emotion. Their level of arousal is demonstrated on the vertical axis and their valence on the horizontal axis.

2.2.4 Facial action coding system

In 1970s Ekman and Friesen [54] developed the Facial Action Coding System which characterized as one of the most useful tools in facial expression analysis [14]. Essentially separates the face into “action units” depending on the stimulation of muscles. Action units (AUs) could be related to more than one muscle.). The elements that made FACS globally recognized are its connection with face anatomy and its precise measurements in facial motion. There are 44 different action units along with a few types of head and eye positions and movements. Therefore, facial expressions can be explained via a system of fundamental elements or AUs. It might require one or more AUs to describe a specific facial expression. For instance, a pretended Pan-Am smile can be described only by the zygomatic major movement (AU 12) while the Duchenne smile requires two facial muscles, the zygomatic major and the inferior part of orbicularis oculi be activated (AU 12, AU 6).

The classification of FACS [64] was a tough task which required much effort to be completed. In particular, the observer had to measure the gray level variation manually during the expressions alternation in images and in some cases to record the electrical activity of facial muscles. Initial FACS edition includes 30 AUs. However, after new discoveries, the last edition has 44 AUs and other “action descriptors”. The latter are referred to facial activity that could not be described with anatomical principles. The combination of AUs can produce more than 7000 expressions [65]. In Table 2.1, a part of the AUs library is presented.

Regarding accuracy, AUs are accompanied by their intensity level. Initially, there were three levels (X, Y, and Z) but in 2002, they updated to five levels (A-E).







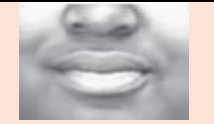
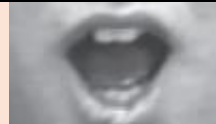
Face Action Units			
AU 5	AU 6	AU 9	AU 4
			
Upper Lid Raiser	Cheek Raiser	Nose Wrinkler	Brow Lowerer
AU 17	AU 23	AU 12	AU 27
			
Chin Raiser	Lip Tightener	Lip Corner Puller	Mouth Stretch

Table 2.1: Specimen from FACS action units [64].

2.3 Face datasets for expression recognition

For an effective facial expression recognition algorithm development, one or more well-labeled databases of sufficient size are required. The variety of the datasets in this kind of research is one of the most significant aspects to build a robust classification model. The most databases are created by asking the participants to display several expressions of emotion. Thus, a characteristic that distinguishes the various databases is if the expressions are performed intentionally or spontaneously. For instance, CK, CK+, and ADFES databases were created based on deliberate expressions while DISFA [68] is built by using spontaneous expressions. Subject's identity includes general features such as sex, age, ethnicity and personal features such as skin texture, hairstyle and color, the shape of the face and other. As a result, strong datasets are distinguished by their ability to include as many of these aspects as possible. Another significant characteristic is the resolution of images. Depending on the method employed, especially in regards to the feature extraction part of the algorithm, an image can be effective in low resolutions, which in turn could be beneficial to the system's computational rate. Six of the most used datasets in literature are presented in table 2.2.

Cohn-Kanade (CK and CK+) [66] databases, based on FACS model (see Section 2.2.4) preferred by most of the FER projects. In these datasets, 100 university students who they ranged in age from 18 to 30 years were participated. In addition, they are characterized by the diversity of the subjects, sixty-five percent are female, 15 percent are African-American, and three percent are Asian or Latino. Grayscale image sequences (640 by 480 pixel) from neutral to target emotion are used. Also, the fact that they can be employed in either static or frame-based system offers flexibility in their use.

Amsterdam Dynamic Facial Expression Set (ADFES) [63] is an elaborate set of 648-filmed emotional expressions. Instead of static pictures, it consists natural facial expressions, which are dynamic events that unfold in a particular fashion over time. It also contains nine emotions: the six basic emotions, along with contempt, pride and embarrassment. Expressions are performed by 22 models (10 female, 12 male). The subjects are North-European and Mediterranean. In addition, the ADFES dataset utilizes active head turning to clarify the directedness of the expressions.

Japanese Female Facial Expressions (JAFFE) database [67], includes only images of female subjects from Japan. The database contains 213 images of seven facial expressions (six basic facial expressions and one neutral) posed by 10 Japanese female models. Furthermore, images have been rated on six emotion adjectives by 60 Japanese subjects.

MMI database [72] consists 2894 sessions (44% are female) from students and research staff members aged from 19 to 62. Their ethnic backgrounds are European, Asian, or South American. In this database, 213 sequences are labeled with six basic expressions, in which 205 sequences are with frontal face.

Belfast Naturalistic Database [73] have become distinguished in the field of datasets because it combines both audio and visual material of people who express spontaneous emotions in television programs. The database contains 298 audiovisual

clips from 125 speakers (94 female and 31 male). The number of emotions for each individual is at maximum two, having a neutral and an emotional state.

Denver Intensity of Spontaneous Facial Action Database (DISFA) is a non-posed facial expression dataset. It contains videos from 27 adult subjects (12 females and 15 males) from various ethnic groups. It is built based on the facial action coding system (FACS) and includes 66 facial landmark points of each image in the database.

In face recognition as well as in feature extraction the head pose is an important aspect, which affects the performance of the system. Thus, face orientation constitutes one more important attribute. In this thesis, to confront the face alignment problem, an algorithm based on geometrical function is employed (see Section 3.3.1) for more information). In most of the databases, human faces are aligned and frontal posed. Even though, some databases include side poses [69, 70].

In most of the algorithms that are designed for facial expression recognition, the scene illumination variation is an additional issue. Some researchers proposed various methods to solve this problem. Only several databases [71] were developed with respect to the lighting issue. In this thesis, the features extraction approach has been used, encountered the illumination variation (see Section 3.4 and Section 3.5 for more information).

Database	Camera View	No of Subjects	Imaging	Images Resolution	Type
Cohn-Kanade CK+ Dataset [66]	Frontal	123	Video Frames	640* 490	Posed and Spontaneous
Amsterdam Dynamic Facial Expression Set ADFES [63]	Frontal	22	Static	720x576	Posed
Japanese Female Facial Expressions JAFFE [67]	Frontal	10	Static	256* 256	Posed
MMI Database [72]	Frontal and 90°	101	Static	720* 576	Posed and Spontaneous
Belfast Naturalistic Database [73]	Frontal	125	Video Frames	Mixed	Spontaneous
Denver Intensity of Spontaneous Facial Action Database DISFA+ [68]	Frontal	27	Video Frames	1024*768	Spontaneous

Table 2.2: Publicly available facial expression databases.

According to the requirements discussed in the beginning of this section, CK+ was chosen as the main dataset for the experiments that performed in this research. Furthermore, the fact that CK+ is employed in numerous similar projects in the past allows an extended comparison. ADFES database selected for internal comparison since the main model was built based on the CK+ dataset. ADFES quality of the images was almost excellent. Therefore, a good prediction rate would justify the ability of a

system that was trained based on low-quality images to perform well. Finally, a new experimental dataset (SelfieDat) with real life's conditions was formed and tested.

2.4 Face detection and facial features localization

Facial expression recognition, from generic images, requires the design of an algorithm, which combines three fundamental functions (see Fig. 2.5): Face Detection, Feature Extraction, and Data Classification. Initially, the face is detected and located within the image. Then, extraction of features that effectively represent the facial expressions is performed. In the last step, facial expression recognition is accomplished.

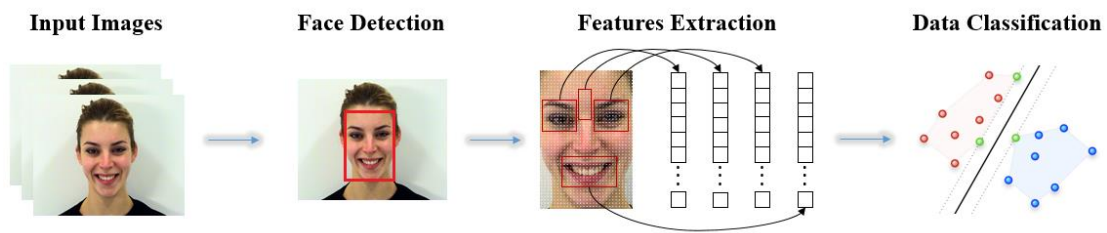


Figure 2.5: FER system pipeline.

In this research at the face detection phase a the region of interest defined. Then it subdivided into four different segments and subsequently data analysis and classification was performed. Segmentation of the human face is related to the shape, motion, color, texture and spatial arrangement of the face [74]. The purpose of facial feature localization is to identify and locate characteristics such as mouth, eyes, eyebrows and ears. There are many methods developed for detection in still images, and they can be grouped into four categories: Knowledge-based, Feature-based, Template-based and Appearance-based.

2.4.1 Knowledge-based methods

Knowledge-based methods rules obtained from the knowledge of human face structure. Thus, a normal face has two eyes symmetric to each other, a nose and a mouth. Facial features location, as well as the distances between them, can be very informative. Based on this information, fixed rules can be embedded in an algorithm, which performs the face detection and facial features extraction. It is common at the end of the process to verify the result in order to decrease the probability of a wrong decision.

Yang and Huang [75] in their work used knowledge-based methods to detect human faces in black and white pictures. The system they designed had three levels: image scanning for possible faces based on a set of rules (see Fig. 2.6); an 8x8 cells window apply a second set of rules over each candidate in order to evaluate them; a

final evaluation is performed and if facial characteristics of eyes and mouth were detected then the face is verified. Successful detection achieved in the 50 of 60 in total images. Later on inspired from their work, Kotropoulos and Pitas [76], introduced a rule-based localization method. Essentially, using a vertical and horizontal projection of the image searched for facial features. In order to acquire the left and right borders of the face, they identified sudden variations in the horizontal depiction of the image. Subsequently, by detecting the local minima in the vertical projection of the image, they determined the locations of the lips, nose, and eyes. Their algorithm applied on 37 frontal-posed face images, and the detection score was 86.5 percent. Mekami's [77] research, extended the above rules to acquire more information such as the inclination of the face.

The knowledge-based methods characterized by their low computational cost, and their performance is high only when they applied to images in the refined background. However, the translation of human knowledge into image processing rules makes this approach challenging. In particular, a very rigorous set of standards could drop the detection rate of the algorithm while a flexible set of standards could present many false positives. Furthermore, due to some rules, it would be costly and difficult to apply this method to non-frontal pose faces.

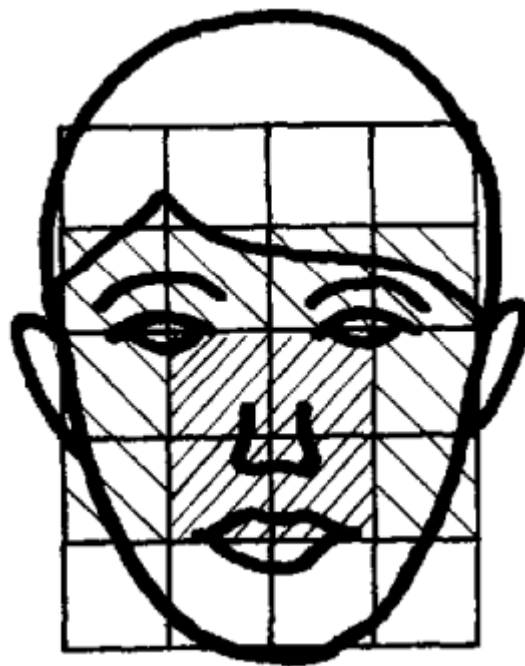


Figure 2.6: An illustration of the rules applied in the first step. The center part of the face (darkly shaded) has four cells with a uniform grey level; the upper round part of a face (lightly shaded) has a uniform grey level etc.

2.4.2 Feature-invariant methods

Feature invariant methods focus on the structure of invariant to pose, viewpoint or lighting intensity facial features. In this approach, basic physical characteristics of the face are used as indicators in various identifiers, for the appropriate choice of data patterns, which subsequently are extracted from the image. Most of the times this approach is performed to detect and extract discriminative features. Then statistical models are used to define the connection between the extracted data and afterward the presence of a face in the input image. Usually, color, edge or texture detection techniques performed on the image and then face and facial features are located.

In his research, Sirohey [78] used the information of the edge map of the image and preprocessing techniques, defined the contour of the face and separated from the background clutter. Han et al. [79] proposed a morphology-based detection model which considers the eyes as the most salient feature. They essentially performed morphological operations to identify eye-analogue pixels and segment the eyes (see Fig. 2.7). Based on color techniques, Huang et al. [80] developed a skin tone filter to identify the facial area while Zhang et al. [81] in their research used centroids of the image in RGB color space to identify one or more faces. Texture-based methods are focusing on the discriminative elements of the facial texture by separating it from the rest of the image. Dai and Nakano [82] proposed a texture-based model which also consists of color techniques to perform face detection from complex backgrounds. They transformed the image from RGB to YIQ color space, and they observed that the orange-like areas of the image (face included there) were enhanced only by using the “I” component.

Even if they are simple, feature-based approaches have poor performance when it comes to confronting problems such as illumination, noise, and occlusion. For example, a shady face in an image could create a significant number of edges, which subsequently leads to a false detection.

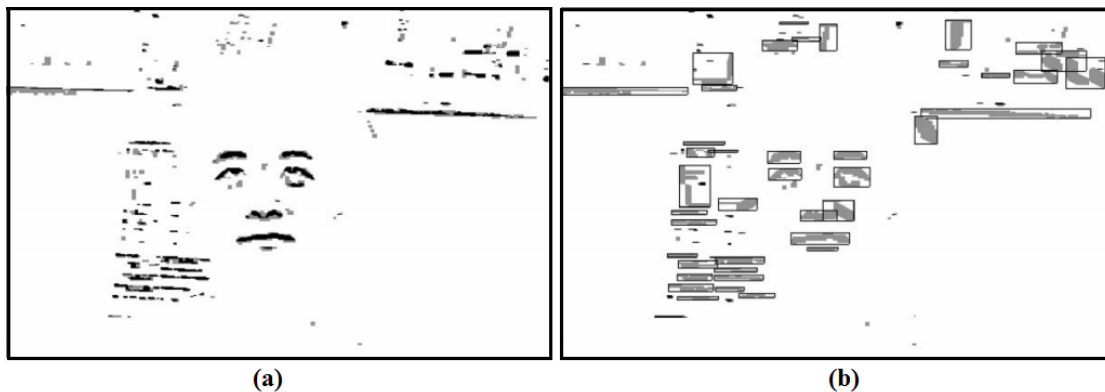


Figure 2.7: The eye-analogue segmentation [79]. (a) The eye-analogue pixels and segments, (b) The eye-analogue segments at the labeling stage.

2.4.3 Template matching methods

Template matching methods differ from other approaches due to the training session that precedes the face and facial features detection. In the training stage, face templates are created by landmark points, curves or volumes, which represent the facial features. Facial components, as well as the face contour labels, are defined manually. The judgment for the existence of a face in the image is achieved based on the correlation values. Template-based methods can be separated into two types, predefined and deformable templates.

In predefined templates method, usually, portions of the face are stored and then correlated with the testing image, to determine the probability of a face existence and location. One of the first experiments with predefined templates, introduced by Sakai et al. [83]. In their work, they used line extraction and pattern detection techniques on eight-level gray digitized pictures, for eyes, mouth and face contour detection. Sinha [84] propose a remarkable model, in which 3-D objects were recognized based on ratio templates. Influenced by Sinha's research, Scassellati [85] developed an algorithm that detects the eyes in a real-time environment. Later on, K Anderson and McOwan [86] based on the ratio template algorithm, proposed a real-time system which performs single face tracking under complex conditions (i.e. crowd places and other inclusions). The modified 'golden ratio' as they named their system, had four stages: search for possible faces using the ratio template algorithm; employ supplementary operators to define the probability of a true face location; probabilities recalculation according to the image motion and matching history and lastly to select the most face-like location.

On the other hand, the deformable templates are aggregate since they can change size and orientation to fit in the image and detect a face and its features. Well-known approaches are the Active Shape Model [87] (ASM) and the Active Appearance Model (AAM) [88]. Cootes et al. [89], used the ASM to detect and locate a face within images. In particular, they employed statistical feature detectors to identify possible features and subsequently based on reshape and alignment methods to define if they are true features. Initially, they created a training set of labeled images, in which facial features are circumscribed by landmark points (see Fig. 2.8). Then principal component analysis (PCA) is performed on the landmark vectors to result in a mean shape model. Once applied to a new image, the model adjusts its local components, gradually fit on the facial features, and face contour. The expansion of previous technique, AAM, proposed by Cootes et al. [88] due to the disadvantage of the first (i.e. evaluate only information about the shape), they introduced a new model AAM which also considers the texture information.

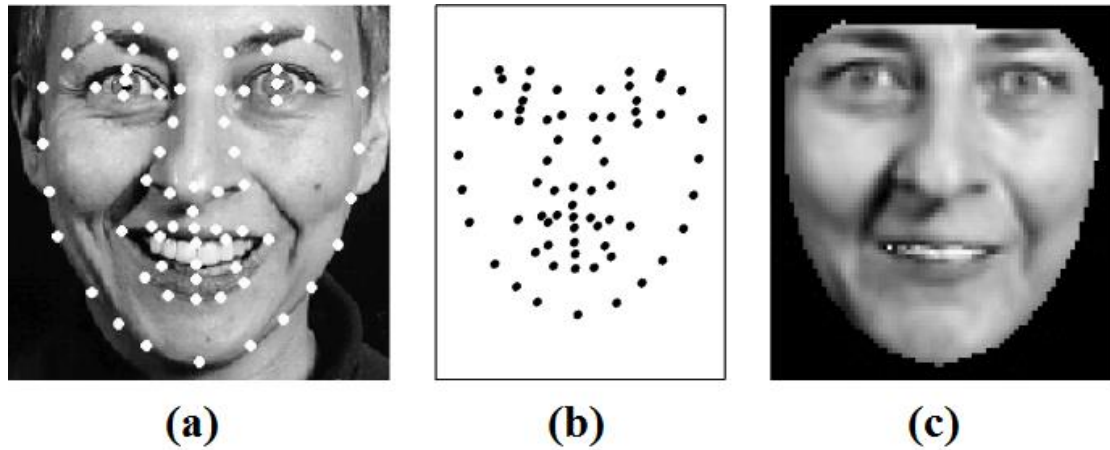


Figure 2.8: Features' marking process. (a) Labeled training image; (b) The landmark points and (c) Shape-free patch of the face [88].

Template matching methods offer more flexibility and accuracy than the previous two techniques that were analyzed. However, they are extremely demanding since the user has to label every image in the training set before he performs it. Last and not least such methods had great results when they were applied on faces that are included in the learned subspace [90], but inaccurate when they were applied to an original face. For example, if a model was trained with sad faces, it will fail to fit in all the characteristics of a happy face.

2.4.4 Appearance-based methods

Appearance-based techniques store various information, such as the intensity level of the pixels in the face image. Then, they compared or fed into classifiers to detect a new face or facial characteristic [91]. In contrast to the other approaches, appearance-based methods, perform statistical analysis, usually combined with machine learning to conclude if an object (i.e. face) exists. Most of the existed models initially perform normalization of the image (i.e. resize, alignment or cropping), features extraction (i.e. texture descriptor in a pixel-based manner) and data learning or classification.

Some papers [92, 93], used PCA on their training set (see Fig. 2.9) to obtain the eigenfaces [94] which capture variations of several facial characteristics such as the hair, the eyes, and the face position. The new face could be decomposed into the eigenfaces as a linear combination of principal components.

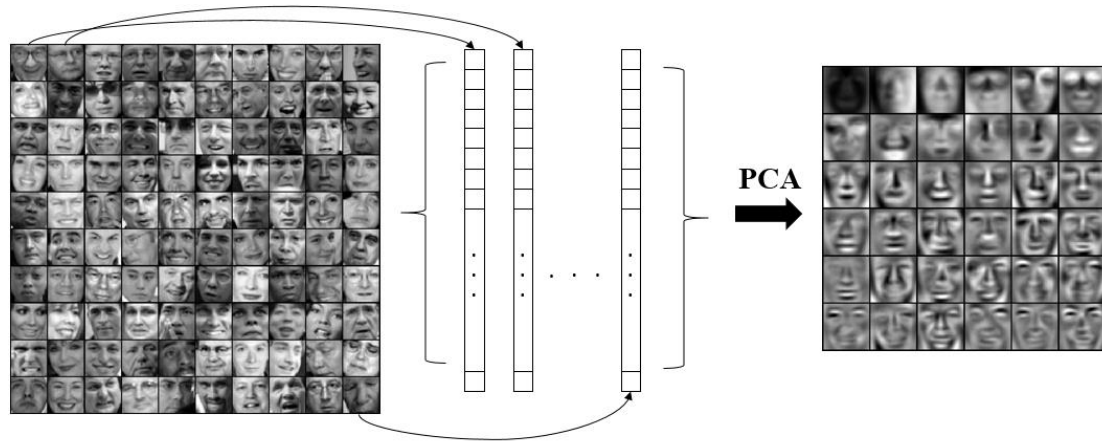


Figure 2.9: Eigenfaces production process [95, 96]. The original 100 faces are “unfolded” into each of the vectors and then PCA applied to the whole array to produce the 36 eigenvectors.

Rowley et al. [97], proposed a frontal face detection model based on neural networks, which could handle the problem of orientation in a cluttered background. First, they classified each possible face region based on its pose and then performed image plane normalizations to align them. Subsequently, the regions fed into several neural networks to evaluate them as a face or non-face. Other similar approaches used support vector machines [98] and hidden Markov models (HMM) [99].

In 1995, Freund [100], introduced the concept of Adaptive Boosting (Adaboost), a machine learning met-algorithm which combines other learning algorithms to create one boosted classifier. This scheme is a conjunction of ‘weak learners’, as they named it, have been used in many object detection projects. In their work Papageorgiou et al. [101] instead of the usual image intensities features, they adapted the Haar wavelets [102], to build a pedestrians detection algorithm. Influenced by the latter, Viola and Jones [103] developed their object detection algorithm which has adopted in various projects for face detection, such as [7, 104-106].

In this research Viola-Jones’ algorithm was employed for face and facial characteristics detection because of its processing time (its speed enables for real-time face detection applications) and robustness in feature selection and its scale and location invariability.

Viola and Jones algorithm

Viola and Jones developed a system for object detection with main characteristics the fast and accurate evaluation of the image. In their model proposed three main concepts: The integral image, the efficient use of the Adaboost classifier and connection of complex classifiers to form one cascade scheme.

As mentioned earlier in this chapter, Haar wavelets that used in Papageorgiou et al. [101] model, inspired Viola and Jones to introduce the so-called Haar-like features

in their detection algorithm. In their simplest form, Haar-like features are two adjoining rectangles (black and white color) which applied on the image and calculate the difference of the sum of the pixels' intensities within the rectangle. In particular, in facial characteristics recognition, the eyes region usually is darker than the forehead region. Selected Haar-Like patterns (see Fig. 2.10) of different size applied on these facial areas and features are extracted.

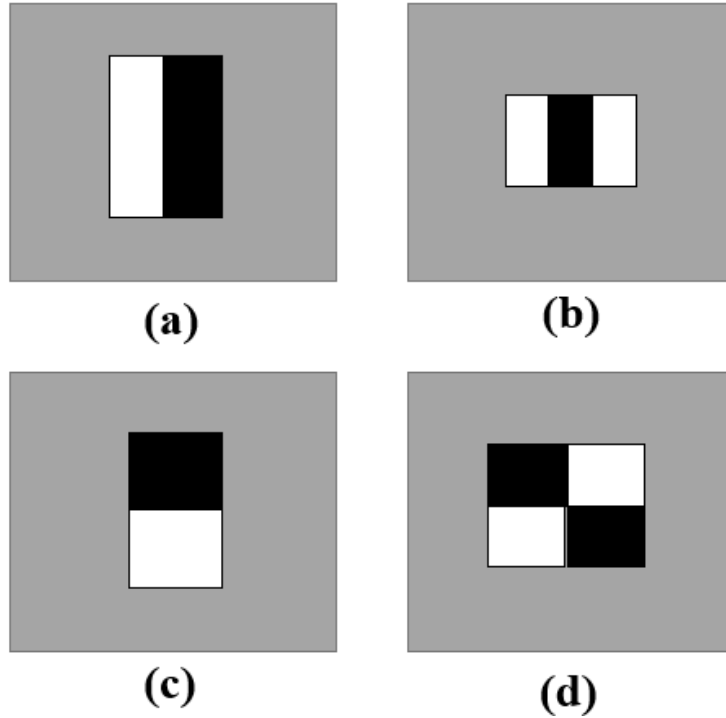


Figure 2.10: Example of four Haar-like patterns used in Viola-Jones algorithm. The size and position of a pattern can change while its black and white rectangles keep their ratio the same. In (a) and (c) two-rectangle patterns are presented, in (b) and (c) three-rectangle are presented and (e) a four-rectangle.

In order to reduce the computational time, Viola and Jones introduced a new concept, that is, the integral image. Essentially, the integral image represents the sum of the pixels above and to the left of a point x, y and is calculated as:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (2.1)$$

where $ii(x, y)$ is the integral image and $i(x', y')$ is the original image. Instead of summing up all pixels within a rectangle, the next pair of recurrences is calculated:

$$s(x, y) = s(x, y-1) + i(x, y) \quad (2.2)$$

$$ii(x, y) = ii(x-1, y) + s(x, y) \quad (2.3)$$

where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$ and $ii(-1, y) = 0$. Hence, it is possible to find the sum of every rectangle in a four array references (see Fig. 2.11). In the case of two-adjoin and four-adjoin rectangle there will be six and nine array references respectively.

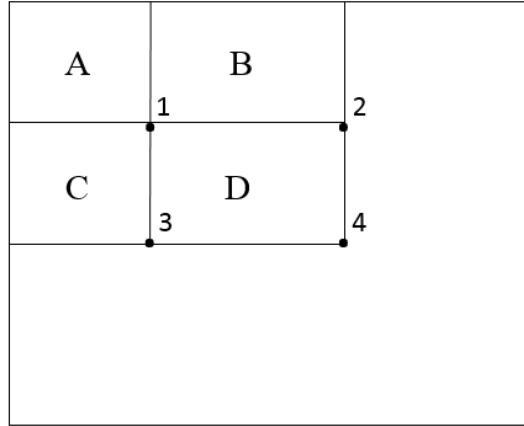


Figure 2.11: The four array references [103]. The rectangle A's sum of pixels represented by the integral image at the point 1. At the point 2 the value is A+B, at the point 3 is A+C and at the point 4 is A+B+C+D. Thus, the sum of the pixels in D is $4 + 1 - (2 + 3)$.

During the features extraction process, more than 160,000 features were produced. Hence, Viola and Jones used a variant of Adaboost learning algorithm for features selection and training. As mentioned earlier in this chapter, Adaboost combine 'weak classifiers'. The idea behind this concept is that the first classifier operates on the training data while is not expected to gain an exceptional score. Then, the examples are re-weighted to focus on the misclassified data. The final 'strong' classifier is composed of some weighted 'weak' classifiers followed by a threshold. In this case, the threshold is used to keep or discard Haar-like features, depending on their values.

In a cascade scheme of classifiers (see Fig. 2.12), initially, Adaboost was chosen to reduce the majority of sub-windows before complex classifiers were used. Hence, the computational times, as well as the number of false positive results, were greatly decreased.

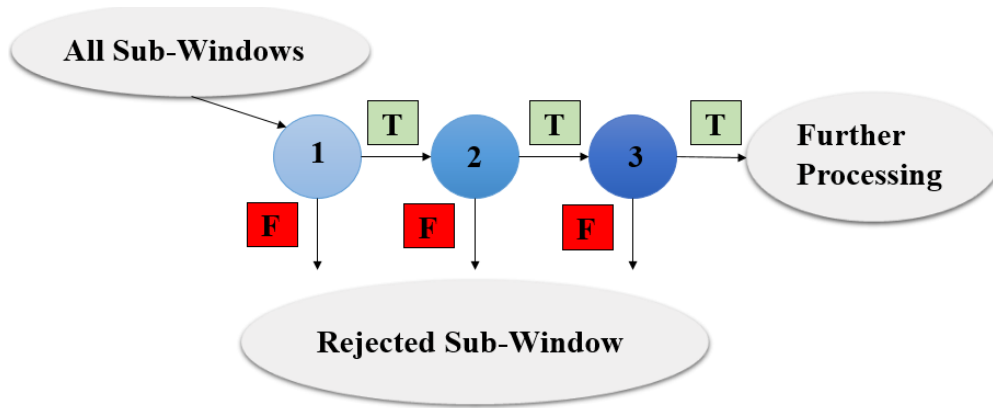


Figure 2.12: Cascade Classifier Scheme [103]. Every sub-window passes through a series of classifiers. The first classifier rejects a large number of examples without limited processing. The next tiers discard more examples until the number of sub-windows sufficiently reduced. Further processing can be further stages of the cascade or a different detection system.

2.5 Facial feature extraction

After the detection and localization within the image, the face should be analyzed in order to derive determinant information of the facial expression and then to proceed to emotion classification and recognition. Hence, the next task is the feature extraction process.

Methods for feature extraction can be separated into three categories [107]: global- or component based; static- or sequence-based and 2-D or 3-D face representation. In literature, researchers usually combined elements of these methods to develop their feature extraction algorithm.

In global approaches features spanning the whole face area while in component approach usually patches of the face are examined. In sequence approach, usually the face is recorded over time, and emotions are classified depending on the facial expression intensity (i.e. onset-peak-offset), with the neutral emotion being the start point. Although this technique provides useful information about the object, it has a considerable disadvantage. The fact that it relies on the neutral emotion as a reference point cannot be used in a realistic situation in which humans' facial expressions are unpredictable. On the other hand, in static approaches, the face is examined regarding the instantaneous facial expression. Thus, the final decision depends on the amount of information the system has about the human face. The argument regarding 2-D (i.e. view-based) and 3-D (i.e. volume-based) representation is directly connected with the system's processing time. In other words, in 3-D representation, the human face might overcome difficulties such as the faces pose, but it is computationally expensive in both face recognition and feature extraction stage.

Features can be categorized as geometric- or appearance based [107] and there is a clear distinction between the methods which are employed in each category in order

to extract the features. In particular, geometric features are usually based on the shape of facial components (i.e. eyes, mouth, nose, etc.). On the other hand, appearance features are based on the appearance (i.e. skin texture) of the face. Table 2.3 illustrates an overview of significant methods that have been employed by the computer vision community for facial expression feature extraction.

Methods	Category		Advantages	Disadvantages
	G	A		
Gabor features		✓	Robust to illumination.	High dimensionality. Alignment is required.
Haar-like features		✓	Computational efficiency.	High dimensionality. Restricted rotation up to 45°.
Local Binary Pattern (LBP) features		✓	Sensitive for subtle facial expression.	Very high dimensionality. Alignment is required.
Histograms of Oriented Gradients features		✓	Sensitive for subtle facial expression. Contrast normalization.	High dimensionality. Alignment is required.
Active Shape Models	✓		No face alignment required.	Training set required. Incapable of capturing subtle facial motion using low-resolution model.
Kalman filters and particle filters	✓		No face alignment required.	Incapable of capturing subtle facial motion using low resolution model. Computationally expensive.
Active Appearance Model	✓	✓	No face alignment required. Combines appearance and geometric features.	Training set required. Computational complex.

Table 2.3: Overview of significant methods that have been used in features extraction process. (Note: In the category column, G: Geometry, A: Appearance. The discrimination depends on which of these categories these methods were mostly used).

2.5.1 Geometric feature-based approaches

Geometric feature-based approaches focus on the location of facial features (eyes, mouth, nose, etc.) and then information about the shape or discriminative relations between them is examined. Usually, these methods are connected with frame-based FER systems and they can categorize into model-based and model-free.

In model-free methods, feature points are set on high contrast areas (see Fig. 2.13) to elucidate the deformation of them. For example, Zhang et al. [108] used Kalman filters to track the pupils of the eyes and then using anthropometric statistics they determine the rest of facial characteristics' location. By measuring the deformation of these regions during the time, they were able to recognize the human emotion through a training-testing process. Related work proposed by Valstar [109] et al. who used Gabor-feature based boosted classifiers to detect the facial characteristics in the first frame, and particle filtering to track the position of them in the following frames.

Model-free methods offer a low-dimensional representation of the feature, which in turn boosts the classification's processing time. Although, they lack in the expression analysis task since they cannot detect subtle alterations of the texture.

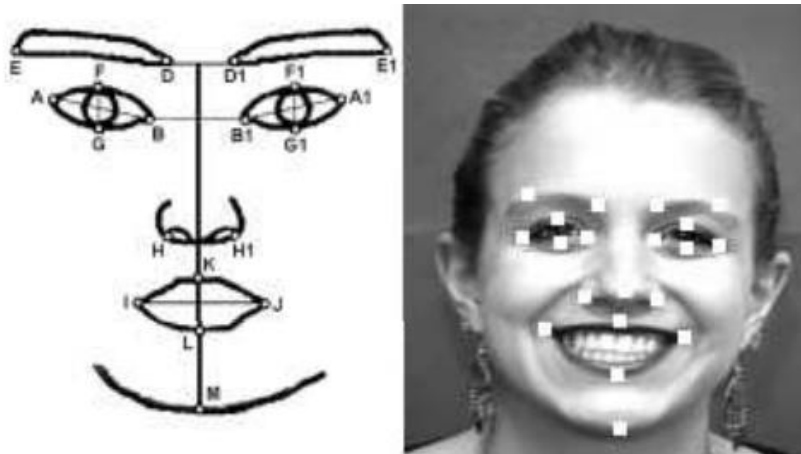


Figure 2.13: Fiducial facial points [109]. The letters A-M represent the points on the edges of the facial characteristics.

In model-based geometric methods, instead of setting or detecting the feature points, usually, statistical models are employed to fit the facial components. Their advantage over the model-free geometric models is the accurate tracking of the fiducial points and that was the reason they preferred more in expression analysis.

Initially, different ASMs which can be used for 2-D or 3-D image analysis, proposed in various FER schemes. ASMs or “Smart Snakes” can take the shape of any object (i.e. a human face). In a new image are repeatedly attempt with deformations to fit on the object of interest and therefore represent it with several vertices. Recently, Shbib et al. [110] employed ASMs to create a model of the face based on 68 feature points. Then they performed face segmentation of faces of CK+ database and extracted

the fiducial points by fitting their model on the faces. The system measured the dislocation of these points and using SVMs estimated the facial expression. Related work proposed by Suk et al. [111] who utilized ASMs and SVMs to develop a mobile application for real-time FER. In general, ASMs can be applied successfully in high-resolution images. However, in low-resolution images, they cannot capture subtle motion. Another disadvantage of ASMs is that they focus only on shape constraints.

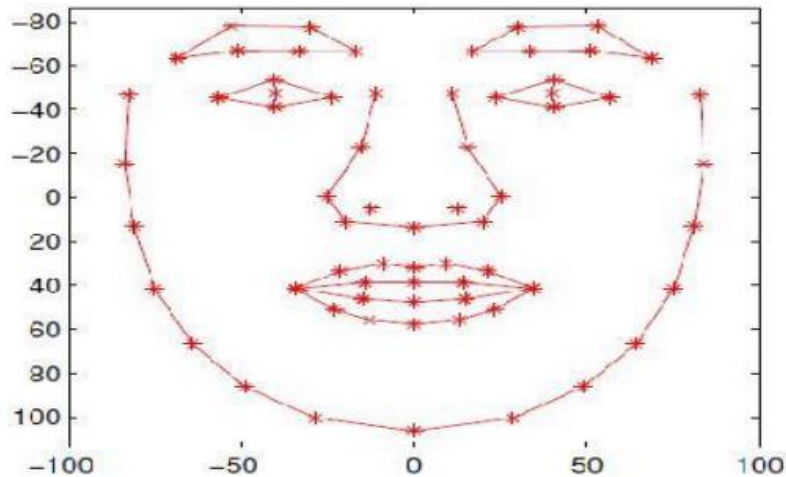


Figure 2.14: ASM Facial Features from Shbib et al. [110].

On the other hand, Active Appearance Models utilize both geometric and appearance features. AAMs have a training stage in which landmark points are set on the model. The difference with ASMs is that allow linear appearance variation. This essentially means that the appearance can be expressed as a base appearance (i.e. the initial image) plus a linear combination of m appearance images (see Fig. 2.15).

Datcu et al. [112] proposed an FER system based on AAMs. They initially detected the face using VJ algorithm and then they applied AAMs to the face area in order to locate the facial characteristics effectively. Asthana et al. evaluated AAM method for facial expression recognition. In their research, they used variants of AAM algorithms in the same scheme (i.e. face recognition – same dataset and classification method). They concluded that Iterative-Discriminative (ID) approach boost the fitting performance. The benefit of such methods is that there is no need for face alignment while the main drawback is that they required a setup stage at the first frame and they are computationally expensive.

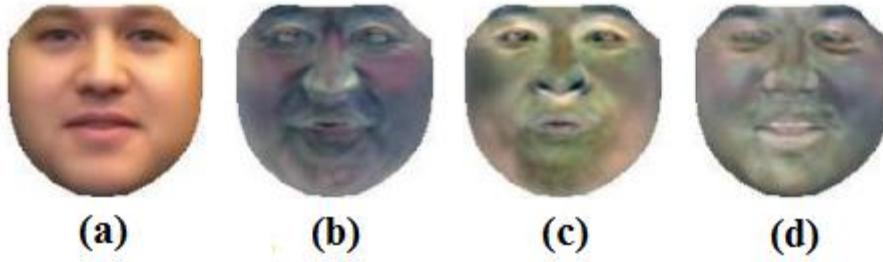


Figure 2.15: The linear appearance variation of an independent AAM [113]. The face in (a) represents the base (i.e. initial face) and (b-d) are represent a linear combination of m appearance images on the same pixel area.

2.5.2 Appearance feature-based approaches

Appearance feature-based approaches focus on the appearance (i.e. skin texture) of the whole face or local regions in order to extract information. Unlike geometric methods, the appearance methods do not extract the facial motion from the dislocation of landmark points on the face. They determine the deformation of the texture if they are sequenced-based or they observe the pixels-relation in still images. Therefore, they are numerous techniques they can perform such analysis. Some of the most significant approaches included: Gabor Filters, Local Binary Patterns, and Histograms of Oriented Gradients.

Gabor filters

The Gabor filters [114] were widely used by the computer vision community for feature extraction, texture analysis [115] and stereo disparity estimation [116] (i.e. binocular depth estimation). Daugman [117] proved in his research that the family of “2-D Gabor filters” can accurately model the function of the two-dimensional receptive field profiles were found experimentally in cortical simple cells. The impulse response of these filters is given by multiplying a Gaussian envelope function with a complex oscillation.

Hence, the 2-D Gabor filter’s general impulse response in spatial domain and frequency domain is given by the following functions [118],

$$h(x, y; f, \theta) = \frac{1}{\sqrt{\pi\sigma_1\sigma_2}} \cdot e^{-\frac{1}{2}\left(\frac{R_1^2}{\sigma_1^2} + \frac{R_2^2}{\sigma_2^2}\right)} \cdot e^{i(f_x x + f_y y)} \quad (2.4)$$

where (x, y) is the image location, σ_1, σ_2 are the standard deviation of the round Gaussian function on the x- and y-axes, and f, θ is the radial center frequency and the

orientation of the Gabor filter respectively. Moreover $R_1 = x \cos \theta + y \sin \theta$, $R_2 = x \sin \theta + y \cos \theta$, $\sigma_1 = \frac{c_1}{f}$, $\sigma_2 = \frac{c_2}{f}$, $f_x = f \cos \theta$, $f_y = f \sin \theta$, c_1 and c_2 are constant numbers. The coefficient $\sqrt{\pi\sigma_1\sigma_2}$ maintains the equivalency of the energies of different Gabor filters to 1.

$$H(u, v; f, \theta) = 2\sqrt{\pi\sigma_1\sigma_2} \cdot e^{\left(-\frac{1}{2}(\sigma_1^2(S_1-f)^2 + \sigma_2^2 S_2^2)\right)} \quad (2.5)$$

where (u, v) are the modulation parameters, $S_1 = u \cos \theta + v \sin \theta$, and $S_2 = -u \sin \theta + v \cos \theta$. Two parts a real and an imaginary constitute the complex Gabor filter (see Fig 2.16) and they define the symmetry as even and odd respectively.

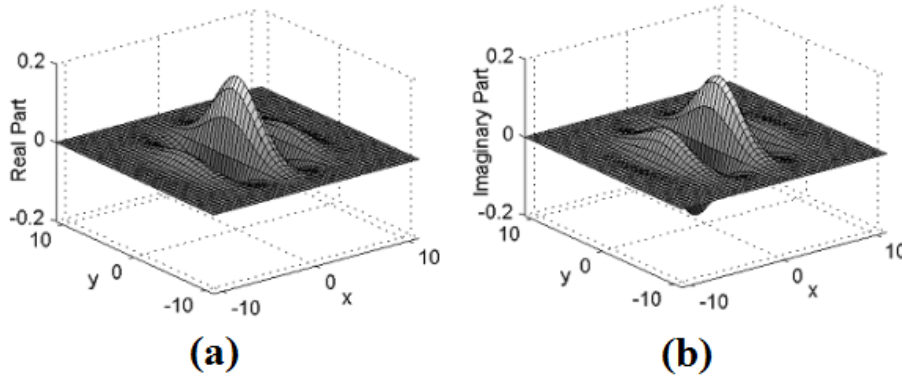


Figure 2.16: Complex Gabor filter ($f = \frac{\pi}{3}$, $\theta = 0$ and $c_1 = c_2 = \pi$) [118]. (a) The real part's graphic representation and (b) the imaginary part's graphic representation.

In facial feature extraction process Gabor wavelets mostly preferred. Gabor wavelets are associated with a Gabor filter bank, (i.e. usually predefined filters) of pre-selected frequencies and orientations in order to avoid a time-consuming process which would include all the 40 possible filters [119]. As mentioned earlier in this chapter Gabor filter has a real and imaginary part. Thus, in order to classify the extracted features, the magnitude of the complex numbers is calculated. Deng et al. proposed Gabor filters for the facial features extraction. In their work, they used five frequencies and eight orientations to demonstrate that the Gabor filters could represent local features very well. Recently, Ou et al. [120] proposed a system which selects 28 facial points and Gabor wavelet filter with five frequencies and eight orientations to extract features in low-quality images is applied. Littlewort et al.[121] had also used Gabor filters for the feature extraction from the CK+ dataset images. Although their video-based FER system had remarkable accuracy, it was computational complex.

Local binary patterns

Local Binary Patterns belongs to the broad category of visual descriptors. The data were extracted using this technique, usually, are fed to a classifier for training and testing schemes. In 1990, He et al. [122] proposed the Texture Spectrum which includes a variety of measures for texture discrimination. Ojala et al. [123], based on He's work, introduced LBP, a powerful method for local texture analysis. In their model reduced the original 6561 texture units to only 256. The process (see Fig. 2.17) in the simple case of a 3x3 neighborhood the pixels are thresholded based on the center pixel value. Then the thresholded image is multiplied by predefined weights of each pixel. The sum of these pixels (i.e. 169) is the final value assign to this texture unit. In particular, the LBP descriptor can be described by the functions (2.6-2.8) [124].

The neighborhood's pixel value differences are calculated as,

$$s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{P-1} - g_c) \quad (2.6)$$

where c is the center pixel value, P : is the number of the sampling points in an evenly spaced circular neighborhood and $s(z)$ is the thresholding function given by,

$$s(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases} \quad (2.7)$$

Therefore, the general function of the LBP operator is,

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (2.8)$$

where the differences of the neighborhood are represented by a P -bit binary number (i.e. for a 3x3 neighborhood is $P=8$) and 2^p different values for the LBP code.

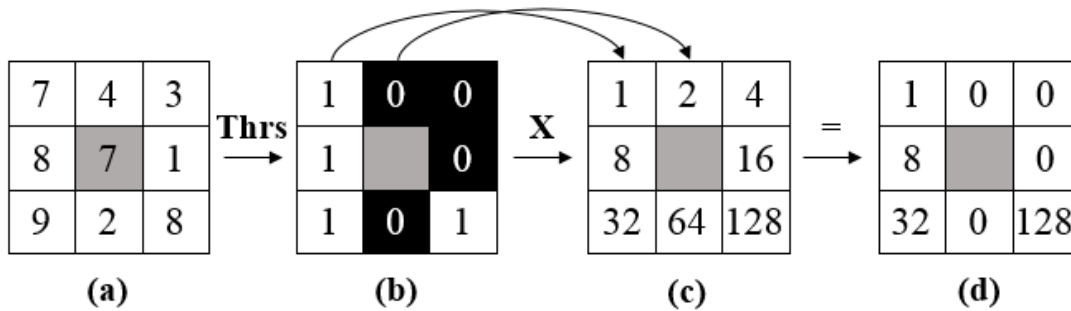


Figure 2.17: The simplest case of LBP in 3x3 pixels neighborhood [123]. (a) The initial image, the center pixel's value used as a threshold, (b) The thresholding image, (c) Weights of each pixel, (d) The resulted image after the LBP implementation.

In literature, various projects used the LBP descriptor for facial expression recognition. Feng et al. [125] proposed a system which utilizes LBP descriptor to recognize facial expressions from static images. In particular, they subdivided the face into 10×8 non-overlapping blocks and they extract the LBP features from each block. Then, they concatenate the extracted histograms into one feature vector to proceed to the classification of the data. Later on, Liao et al. [126] in their FER algorithm combined three different techniques, LBP; Tsallis entropy and the null-space based LDA to extract texture features from the human face. They tested their algorithm using different combinations of these three techniques in in four different image resolution having remarkably high rates. In the same manner, Zhao et al. [127] used an extension of LBP called volume local binary patterns (VLBP), to combine motion and appearance features. Essentially the difference between VLBP and simple LBP is that the analysis is extended to the spatiotemporal domain using dynamic textures. Recently Shan et al. [128] have used simple LBP and boosted-LBP features in their FER system. They also examined the LBP behavior in low-resolution images, and their results were promising. Recently Happy et al. [10] introduced an exceptional emotion recognition algorithm that utilized the LBP histogram features of the salient patches in the face. In order to be precise they used various bin-widths of LBP histograms.

In this research, due to its processing time and promising performance in texture analysis, LBP descriptor is one of the feature extraction methods is used.

Histograms of oriented gradients

In 2005, Navneet Dalal and Bill Triggs [129], presented their novel work at the Conference on Computer Vision and Pattern Recognition (CVPR). They developed a feature descriptor, called the Histograms of Oriented Gradients. Initially, the intention for this descriptor was to detect humans in an ambiguous scene, under complex conditions. Later on, this approach adopted by the research community for various object detection purposes. HOG descriptor operates in grayscale images (intensity: 0-255) and uses a sliding detection window which is moved around the image. At each position of the detection window, a HOG descriptor is computed and then is fed to an SVM classifier, which classifies it as either a “person” or “not a person”. To compute the HOG features the detection window is separated into a fixed number of cells. Within every cell, the gradient vector at each pixel is computed and put into an n-bin histogram. Then, block normalization follows in order to make the results invariant to illumination variation. HOG descriptor can be described by three sub-processes: gradient vectors computation, histograms of cells extraction and block normalization.

❖ Gradient Vectors

A gradient vector [130] is essentially given by the computation of the derivative in the x-direction and the y-direction respectively and is expressed as,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f(x, y)}{\partial x} \\ \frac{\partial f(x, y)}{\partial y} \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \end{bmatrix} \quad (2.9)$$

where, $f(x, y)$ is the image function and $\begin{bmatrix} f_x \\ f_y \end{bmatrix}$ is the gradient vector. Also, the gradient vector has magnitude and direction which are given by,

$$|\nabla f(x, y)| = \sqrt{f_x^2 + f_y^2} \quad (2.10)$$

$$\theta = \tan^{-1} \frac{f_x}{f_y} \quad (2.11)$$

When the HOG descriptor is applied on the image, the gradient vectors of the exterior pixels cannot be computed since some of the neighboring pixels do not exist. Thus, an effective solution is to pad the contour of the image with zero-valued pixels. An example of the gradients computation is given in Fig. 2.18. A small cell of the image is isolated and magnified so that the pixels can be distinguished by the human eye. In this example, a random pixel is chosen. To compute the gradient vector, the 4-connected neighbors¹ of the pixel of interest are figured and subtracted from each other in x- and y-coordinate respectively. Then based on the equations (2.10) and (2.11) above, the magnitude and the angle of the gradient vector are computed.

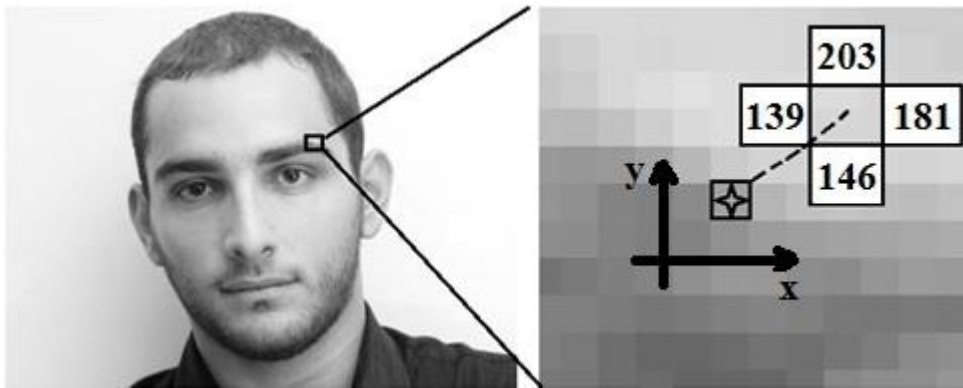


Figure 2.18: Gradient Vector, Magnitude, and Angle.

¹ These pixels are connected horizontally and vertically ($x \pm 1, y$) or ($x, y \pm 1$) is connected to the pixel at (x, y).

In particular, for this example, which represents part of the experimental process of this research, the computation was performed as,

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} 181 & - & 139 \\ 203 & - & 146 \end{bmatrix} = \begin{bmatrix} 42 \\ 57 \end{bmatrix} \quad (2.12)$$

$$M = \sqrt{42^2 + 57^2} = 71 \quad (2.13)$$

$$\theta = \tan^{-1} \frac{42}{57} = 0.6350 \text{ or } 36.5^\circ \quad (2.14)$$

where m and θ are the magnitude and the angle of the vector respectively.

❖ Histogram of cells

As mentioned in the introduction of this section, the gradients are computed within every cell of the image and then are put into an n-bin histogram. In the original paper, Dalal and Triggs used a 9-bin histogram. A similar process followed to compute the HOG features in the example in Fig. 2.19. The histogram ranges from 0 to 180 degrees, so there are 20 degrees per bin. For each gradient vector, its contribution to the histogram is given by the magnitude of the vector (so stronger gradients have a higher impact on the histogram). The benefit of this technique is the detailed pixel-based representation of the image in a data-efficient manner. Even if it is computational-costly to process the image in a pixel-basis, same time there is a reduction from 64 components to only 9-bin histogram values. A common confusion in HOG descriptor [131] is the realization of what information is given by the histogram. The histogram does not encode where each gradient is within the cell; it only encodes the “distribution” of gradients within the cell.

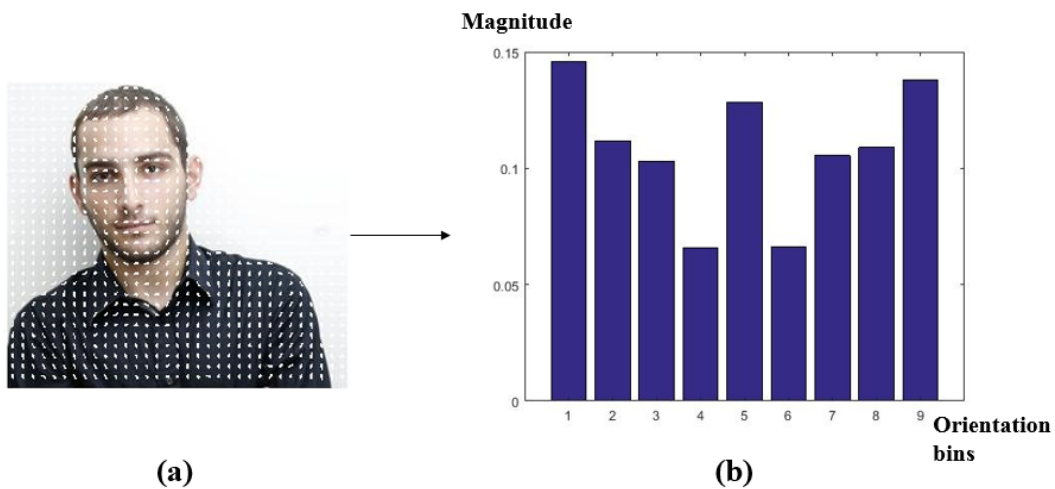


Figure 2.19: Example of the HOG descriptor of 8×8 cell size and 9 orientation bins operation on a 237×264 image. (a) HOG visualization, (b) Histogram of gradients.

❖ Block normalization

In order to make the intensity of the image invariant under different illumination, block normalization is performed. For example, an image 3x3 given by,

$$I(x, y) = \begin{bmatrix} I_1 & I_2 & I_3 \\ I_4 & I_5 & I_6 \\ I_7 & I_8 & I_9 \end{bmatrix} \Rightarrow \begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} I_2 - I_8 \\ I_4 - I_6 \end{bmatrix} \quad (2.15)$$

where I_5 is the point of interest and $\begin{bmatrix} I_x \\ I_y \end{bmatrix}$ is the gradient vector. If an amount would be added or subtracted to every pixel of the image that would not affect the gradients computation, (see equation 2.16). Although, in real conditions, in which all the pixel values would be multiplied by an arbitrary amount (see equation 2.17), the new gradients would be different.

$$I(x, y) = \begin{bmatrix} I_1 + 1 & I_2 + 1 & I_3 + 1 \\ I_4 + 1 & I_5 + 1 & I_6 + 1 \\ I_7 + 1 & I_8 + 1 & I_9 + 1 \end{bmatrix} \Rightarrow \begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} I_2 - I_8 \\ I_4 - I_6 \end{bmatrix} \quad (2.16)$$

$$I(x, y) = \frac{1}{2} \times \begin{bmatrix} I_1 & I_2 & I_3 \\ I_4 & I_5 & I_6 \\ I_7 & I_8 & I_9 \end{bmatrix} \Rightarrow \begin{bmatrix} I_x \\ I_y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \times I_2 - \frac{1}{2} \times I_8 \\ \frac{1}{2} \times I_4 - \frac{1}{2} \times I_6 \end{bmatrix} \quad (2.17)$$

Instead of normalizing each gradient vector or each histogram of the cell, it is computationally efficient to perform block normalization. The blocks used by Dalal and Triggs consisted of 2 x 2 cells. The blocks had “50% overlap” and the process is performed by concatenating the histograms of the four cells within the block into a vector of 36 components (4 histograms x 9 bins per histogram). Then, the vector is divided by its magnitude in order to normalize its values. The effect of the block overlap is that each cell will appear multiple times in the final descriptor, but normalized by a different set of neighboring cells. (Specifically, the corner cells appear once, the other edge cells appear twice each, and the interior cells appear four times each).

Orrite et al. [132] proposed an emotion recognition model based on HOG feature analysis. Initially, they utilized a Bayesian formulation to compute the edge distribution within the dataset. Hence, they were able to define the most discriminate locations for each of the emotion. Then, they used a decision tree to cluster and merge the feature classes. HOG features were calculated using the log-likelihood maps in every branch

of the tree and finally, the data were fed to an SVM classifier. In three of the six emotions, the performance of their algorithm was impressive. In their emotion recognition algorithm, Dahmane et al. [133] used HOG method to extract the texture features. They accumulate the gradient magnitudes for a set of orientations of 1-dimensional histograms over a size-adaptive dense grid. A novel work presented recently by Carcagnì et al. [15], who used the HOG descriptor on the whole face region to classify seven emotions. Their system outperformed previous FER attempts and their results were very promising.

Inspired by Carcagnì's work, for the feature extraction part of the algorithm proposed in Section 3, HOG descriptor is used. However, instead of use a holistic method, in this paper salient patches are detected and analyzed.

2.6 Facial expression classification and recognition

The utilization of machine learning algorithms constitutes the last step in FER systems. Machine learning studies how to learn automatically to make accurate predictions based on past observations. There are two types of learning, supervised and unsupervised. In supervised learning, the categories of each observation are pre-defined and given (i.e. labeled training data) to the classifier. On the other hand, in unsupervised learning training data are not labeled and categories must be identified by the classifier.

Machine learning can be further categorized according to the resulting output. In some cases is required to perform regression, when the outputs are continued and not discrete. In other cases, the data are divided into groups depending on their characteristics and that is what is called clustering. In FER systems, usually, classification is used. Classification is the process that the input data are divided into two or more classes, and the learning algorithm has to create a model that assigns the new data to one or more of these classes (see Fig. 2.20). Based on the mechanism each classification method using, three sub-categories can be formed: Template-based methods, neural-network based methods and statistical classification methods. Some of the most known classification algorithms, which employed in FER research, are presented in table 2.4.

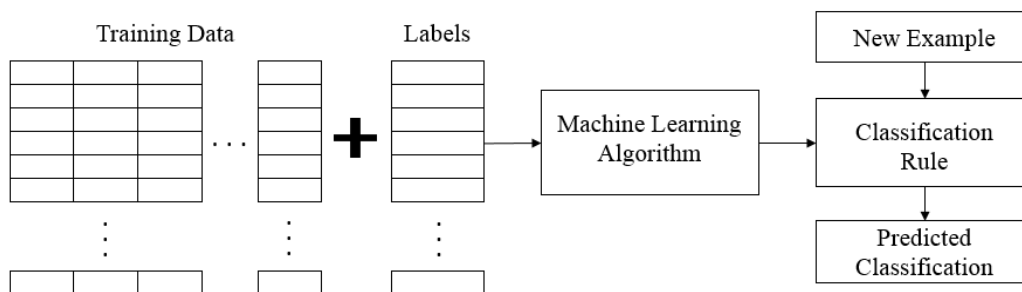


Figure 2.20: Classification Algorithm

Method	Comments
Template matching	Assigns samples to the most similar template. Normalization is required.
Neural Network	Composed of 'neurons' or nodes - that mimic the properties of biological neurons. Weighted connections between the neurons create a classification model.
Naïve Bayes	Assign pattern to the class with the highest estimated posterior probability.
Classification Trees	Features represented by nodes.
Support Vector Machine	Expand the margin between two classes (i.e. two-class SVM).
Linear Discriminant Analysis	Model the differences (i.e. differenced projection vectors) between classes.

Table 2.4: Overview of significant classification methods were used in FER systems.

2.6.1 Template matching

Template matching techniques are utilized to find areas of an image that match to a template. There are two main components: the source and the template image. The source image is the object of interest, which is tested in order to find similarities to the template. The learned templates in their simple form are the average of the training images in each class (i.e. a template for joy emotion, another for anger emotion etc.). As it mentioned earlier in Section 2.4.3, the template matching techniques, are demanding and in most cases, manually labeling is needed. In general object detection-recognition problems (i.e. face detection), the results are reasonable. Although, when the information is based on subtle changes (i.e. facial features formation), template methods are incapable of offering good accuracy due to the fact that the averaging process leads to a smoothing of significant facial details [134]. Previous attempts of expression recognition based on template matching methods are [93, 135-137].

2.6.2 Artificial Neural Networks

Artificial Neural Network (ANN) or usually in literature Neural Network (NN) is a very common technique for facial expression classification. The NN method has its roots in human biology since it imitates the way that the human neurons are processing the information and solve problems. In machine learning NN, refer to the interconnections between neurons in a specific number of hidden layers of each system. The synapses in each layer constitute the parameters of the system for the manipulation of the data and usually, they called weights. Numerous types-approaches of NNs such as Feedforward neural network (FNN), Recurrent neural network (RNN), Static neural network employed in the past in FER systems. NNs proved a promising method to map facial

expressions based on image features. However the decision of the appropriate network size was always complicated [138] and sometimes led to poor results. An example of how the neural network classifies the six basic emotions with a full-face approach and feature-based approach is presented in Fig. 2.21.

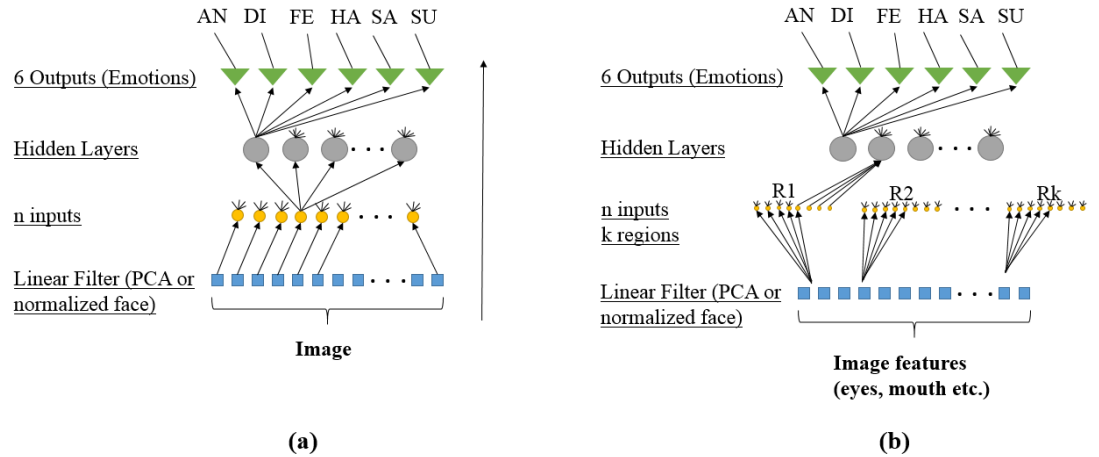


Figure 2.21: Neural Networks classification for six basic emotions (AN: Anger, DI: Disgust, FE: Fear, HA: Happy, SA: Sad and SU: Surprise). (a) A full-face approach, (b) A feature-based approach.

Gargesha et al. [139] based on static approaches, Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) networks to classify the six basic emotions and the neutral state. Their data were extracted based on geometric feature approach. In particular, the Euclidean distances for the contour points of the face along with differences in inter-feature distances constituted the features were fed into the neural network. Ma et al. [138] utilized feedforward NNs approach to classifying five emotions and the neutral state. They particularly used the two-dimensional discrete cosine transform (DCT) to detect facial features and subsequently they fed the NN classifier using only one hidden layer. A recent approach for emotion analysis proposed by Katratwar et al. [140], who used Gabor filtering for feature extraction. Then they train their system using MLP network along with backpropagation of error. Other related approaches based on NN classification methods are [141-143].

2.6.3 Naïve Bayes

Naive Bayes (NB) classifiers belong to the broad category of probabilistic classifiers, and they are based on Bayes theorem. NB classifiers are operating on the assumption that the value of a specific feature is independent of the value of any other feature, given

the class variable. Essentially, NB is a conditional probability model which takes a given feature vector and assign to this instance probabilities. For example given a vector,

$$x = (x_1, x_2, \dots, x_n) \quad (2.18)$$

where x has n features (independent to each other), then the assigned probabilities are,

$$p(C_k | x_1, x_2, \dots, x_n) \quad (2.19)$$

where there are K possible outcomes or classes C_k . However, the above formulation cannot deal with a large number of features. Therefore, the above formula needed to become more flexible. According to the Bayes theorem, the conditional probability can be reconstructed as,

$$p(C_k | x) = \frac{p(C_k)p(x | C_k)}{p(x)} \quad (2.20)$$

where $p(C_k | x)$ represents the posterior, $p(C_k)$ represents the prior and $p(x | C_k)$ represents the likelihood probabilities. The NB classifier is an integration of the probability model and a decision rule. Usually, follow the hypothesis of the most probable result, that is called MAP (maximum a posteriori), the decision rule is used. A Bayes classifier predicts the class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \arg \max p(C_k) \prod_{i=1}^n p(x_i | C_k) \quad (2.21)$$

There are many FER systems using NB classifier or similar methods, which are based on Bayes theorem. Although NBs were criticized [144] for their poor performance, recent work [145, 146] shows that they are still preferred in research. Also, exceptional systems using Bayesian or alternatives of them classifiers proposed by, [147, 148].

2.6.4 Classification trees

A classification tree is a general category in machine learning which includes several classification algorithms: Bagging decision trees, Random Forest, Boosted Trees and Rotation Forest. A simple tree structure constitutes branches and leaves. Branches represent a combination of features that end to the leaves, which in turn represent the class labels. Nodes from the first (i.e. root of the tree) to the last (i.e. leaves) play the most important role. Essentially, in each node, a single rule is used to divide the data into disjoint subsets. A simple example of classification tree in an FER system with

three emotions and the neutral state is presented in Fig. 2.22. Initially, all the data are fed to the first decision node. Anger and happy emotions are separated from Sad and then they compare with each other. In the end, every emotion is compared with the neutral state.

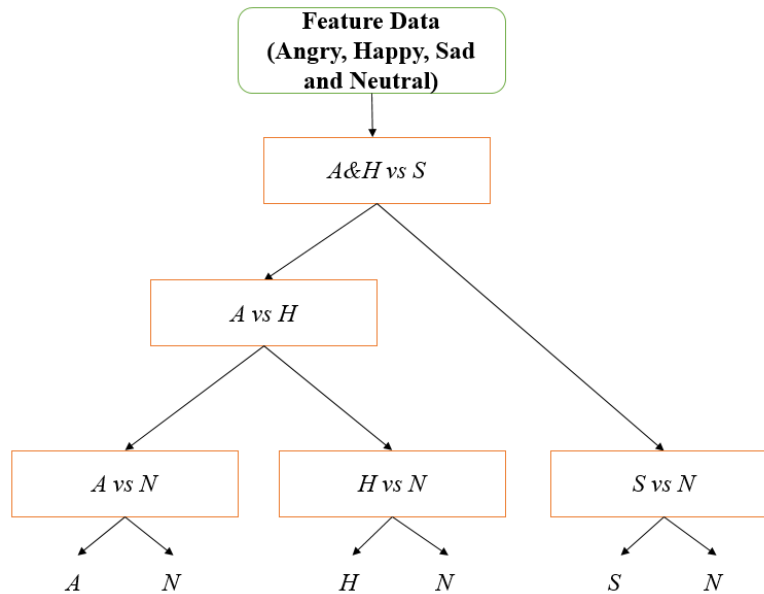


Figure 2.22: Classification Tree example. Three emotions and the neutral state are classified using a decision tree classifier with five nodes and 11 branches.

Since classification trees can be employed through several approaches, there are numerous FER systems based on them. In literature, various types such as decision trees [149, 150], boosted and bagging trees [151], and random forests [152], models have been utilized.

2.6.5 Support vector machine

In 1963, Vapnik and Chervonenkis [153] invented Support Vector Machine (SVM) which is one of the most popular classifiers in visual pattern recognition. SVMs are mainly used in supervised learning. Initially were performing linear classification and were operating with two-class problems. In 1992, Boser et al. [154] introduced the “kernel trick”, which enable SVMs to map their inputs into high-dimensional feature spaces and essentially perform a non-linear classification.

The basic idea behind the SVMs operation is that classification for a two-class problem is achieved by maximizing the width of the empty area (margin) between the two classes [139]. The margin specifies the distance between the discrimination hypersurface in n-dimensional feature space and the closest training patterns, which are called support vectors. Thus, for a two-class linearly separable, with n-dimensional

normalized feature vectors x , where $x \in [0,1]$ and related classes ω , where $\omega \in \{-1,1\}$ a separating hyperplane (see Fig. 2.23) can be defined as,

$$w \cdot x + b = 0 \quad (2.22)$$

where w is the normal vector and $\frac{b}{\|w\|}$ is the perpendicular distance from the hyperplane to the origin. In order to maximize the margin, two parallel hyperplanes are defined as,

$$w \cdot x + b = 1 \quad w \cdot x + b = -1 \quad (2.23)$$

covering the support vectors and no training data between them. To ensure that no training patterns exist between the two hyperplanes, the following inequality is used for all the training data x_i ,

$$\omega_i(w \cdot x_i + b) \geq 1 \quad (2.24)$$

From the above formula is clear that maximization of the margin requires minimization of $\|w\|$.

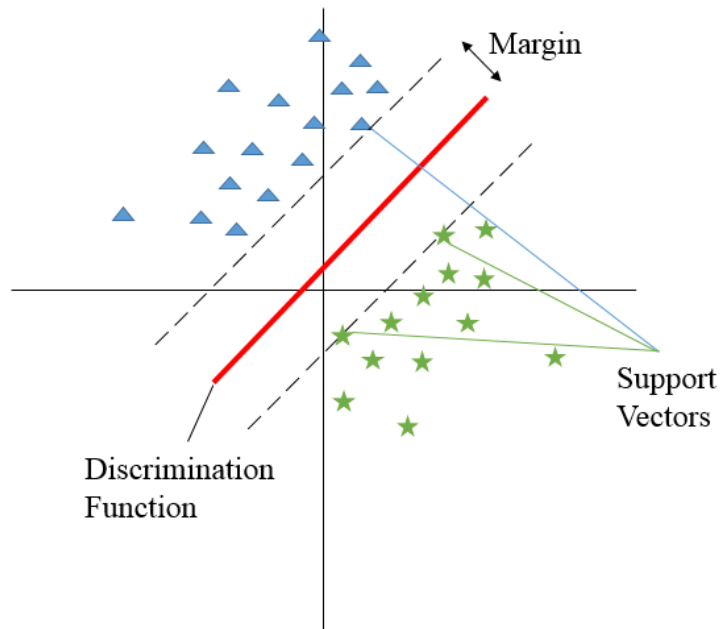


Figure 2.23: Simple two-class classification example of Support Vector Machine. In this example, an optimal linear discrimination function maximizes the margin between the data of the two separated classes. The optimal hyperplane explained by the function of the support vectors.

SVMs are commonly used in FER research since they offer impressive results. They can also be beneficial when many separating hypersurfaces exist, as it always finds the optimal. In other words when the feature vector is large SVM have proved very useful method. Moreover, unlike other classifiers, SVMs constitute a good approach to avoid the problem of overfitting. Some of the most recent research in FER based on SVM classification introduced by [155-157] while in the past impressive work proposed by [158-161]. In this research, because of their performance especially with high-dimensional feature vectors, SVM classifier is one of the classification methods is used.

2.6.6 Linear discriminant analysis

Linear Discriminant Analysis is a classification method, which assumes that different classes produce data based on different Gaussian distributions. LDA is also known as Fisher discriminant analysis, named after its developer R. A. Fisher [162]. LDA is based on the concept of searching for a linear combination of variables (i.e. predictors) that best separates two classes (i.e. targets) [163]. In order to capture the idea of separation of the classes, Fisher developed the following score function (2.27),

$$Z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_d x_d \quad (2.25)$$

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad (2.26)$$

$$S(\beta) = \frac{\overline{Z_1} - \overline{Z_2}}{\text{Variance of } Z \text{ within groups}} \quad (2.27)$$

where β are the linear model coefficients, C is the pooled covariance matrix and μ_1, μ_2 are the mean vectors. The problem of maximization of score can be solved by calculating the linear model coefficients and pooled covariance matrix from the following equations,

$$\beta = C^{-1}(\mu_1 - \mu_2) \quad (2.28)$$

$$C = \frac{1}{n_1 + n_2} (n_1 C_1 + n_2 C_2) \quad (2.29)$$

where n_1, n_2 are the number of elements in group respectively. Finally, new data point classification is achieved by projecting it to the maximum separating direction. Then if it conforms to the following equation is classified as C_1 , otherwise as C_2 ,

$$\beta^T \left[x - \left(\frac{\mu_1 + \mu_2}{2} \right) \right] > \log \frac{p(c_1)}{p(c_2)} \quad (2.30)$$

where $p(c_1), p(c_2)$ are the class prior probabilities. An example of a two-class LDA classification is presented in Fig. 2.24.

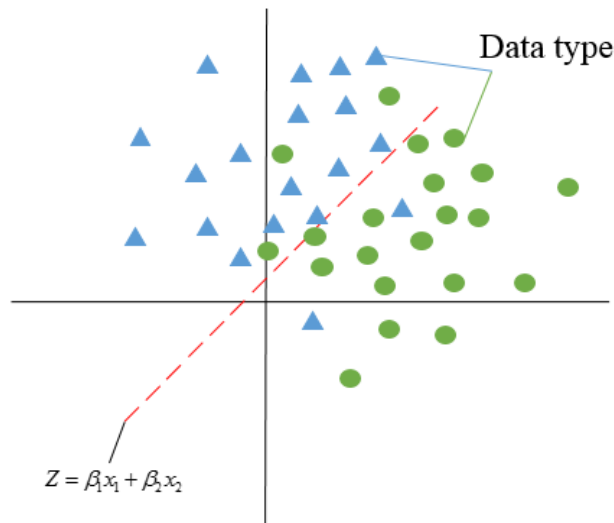


Figure 2.24: Simple two-class classification example of Linear Discriminant Analysis. In this example, LDA is used to find an optimal linear model that best separates two classes.

In literature, LDA classification approach has been widely used in face detection as well as in FER research and it is considered as a fast and accurate method. Some of the most recent FER systems, which employed LDA classification, are [126, 128, 164-166]. In this research, LDA classification preferred for its performance on high-dimensional data vectors.

2.7 Summary

In this chapter, we surveyed the most important methods used in face detection, feature extraction and classification schemes. The combination of these methods got great success on the collected databases, however, for a practical automatic expression system, the existing approaches still could not confront the problem in the real data such as face misalignment and undetected facial characteristic. In this thesis, we proposed the VJ algorithm for face detection, a geometric solution for face alignment, HOG and LBP for feature extraction, and SVM and LDA classifiers to handle these problems.

Chapter 3

Emotion Recognition System (ERS)

3.1 Introduction

This Chapter includes the description of the steps followed in order to create the Emotion Recognition System. The ERS analyzes the facial expression of frontal face images, with the implementation of Matlab 2016b. Facial expression analysis, in generic images, requires the design of an algorithm, which combines four basic functions: face detection, salient regions detection, feature extraction and data classification.

The methods of the face and facial feature detection were selected based on the robustness and latest reports in literature (see Section 2.4.4). Furthermore, the process of feature extraction, which is the most important step in automatic facial expression-emotion recognition systems, was examined using two different techniques and a variety of parameter combinations. As discussed in Section 2.5, in the past many feature extraction approaches have been employed. However, as specified in Section 1.2, this research was based on the social science findings and particularly Ekman's [52, 54] proposed model of human emotions. Ekman proved that during an emotion stimulation, different combinations of muscles are triggered in order to form a facial expression (see FACS in Section 2.2.4). Therefore, this research implemented two appearance-based methods, HOG and LBP, based on their exceptional performance in texture analysis. These powerful descriptors were applied into three salient regions of the face (mouth, eyes and glabella) in order to capture subtle information of the skin formation for each of the six basic emotions: anger, disgust, fear, joy, sadness and surprise. Then, the feature data were fed into two different classification algorithms, LDA and SVM resulting in the four most efficient ERS models: three based on each facial characteristic separately and one based on all three facial characteristics. A simple schematic overview of the proposed framework is presented in Fig. 3.1.

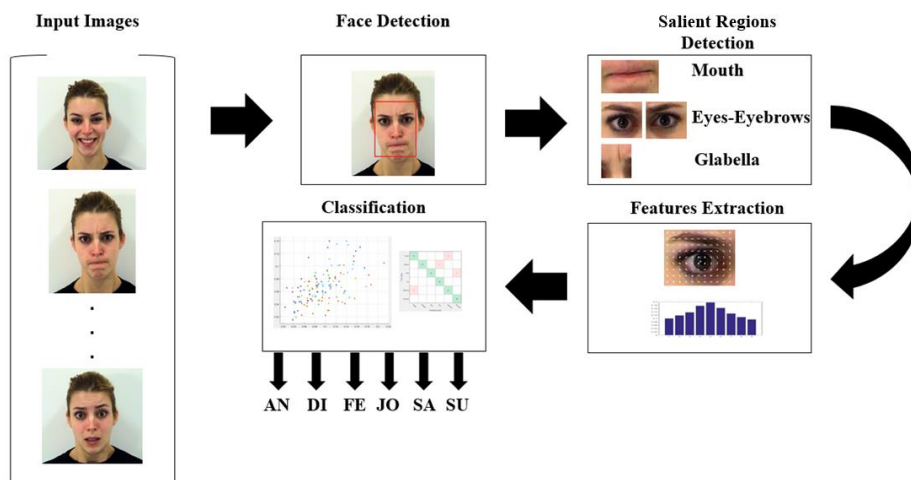


Figure 3.1: Schematic overview of the proposed framework.

The following sections present the procedure of the construction of the ERS algorithm. Section 3.2 presents the proposed algorithm and briefly analyses the principal components. In Section 3.3, face and salient regions detector algorithms are discussed. Moreover, Sections 3.4 and 3.5 include the analysis of the feature extraction algorithms, HOG and LBP respectively and present the experiments on two different classification methods. Based on the classification performances, the most accurate and efficient models are compared with state-of-art methods.

3.2 Algorithm overview

A block diagram of the proposed algorithm is shown in Fig. 3.2. Initially, on the setup process, the user defines the feature extraction method. Then, all the available images in the checking folder are saved in a cell array. For each of the input images, possible faces $f_i \in i = 1, \dots, n$ are detected. Based on the position, $[(x_i, x_j), (y_i, y_j)]$ each face is cropped and analyzed independently. The cropped face is passed to the alignment checking sub-function, in which left and right eyes detection is performed on the upper half of the image. Depending on the position of the eyes' centers, the face is aligned if required.

After the successful face detection, the algorithm continues with the detection of three salient regions (mouth, eyes and glabella). The human facial characteristics are usually uncovered. However, it is possible that occlusions can lead to false results. Therefore, this algorithm is designed to take into account, the maximum available regions by the following process:

- 1) The face image is passed through a homomorphic filter to correct the non-uniform illumination (see Section 3.3.3).
- 2) Considering only the lower half of the face image, the mouth location is detected and cropped. If the mouth cannot be detected, an identification flag for the mouth is defined as false.
- 3) Focusing on the upper half of the face image and knowing the locations of left and right eyes, which were detected earlier in the alignment process, the eyes' centers are defined. Having the position of the eyes in the image, eyebrows and glabella locations, are calculated geometrically. Eye and eyebrow are cropped together in one image while glabella is cropped independently. If the eyes cannot be found, then the corresponding flag for the eyes is defined as false.
- 4) If both the above flags are false, the user is informed that the detection was unsuccessful. Otherwise, the feature extraction process begins with all the available regions.
- 5) For each of the available facial characteristic, HOG or LBP feature are extracted. Finally, a feature vector is formed by concatenating all the calculated data.

The concatenated feature vector is fed into the classifier. Depending on the available facial regions, one of the following schemes (MEG-classifier, M-classifier or

EG-classifier) is enabled and the facial expression is classified as anger, discussed, fear, joy, sad or surprise emotion.

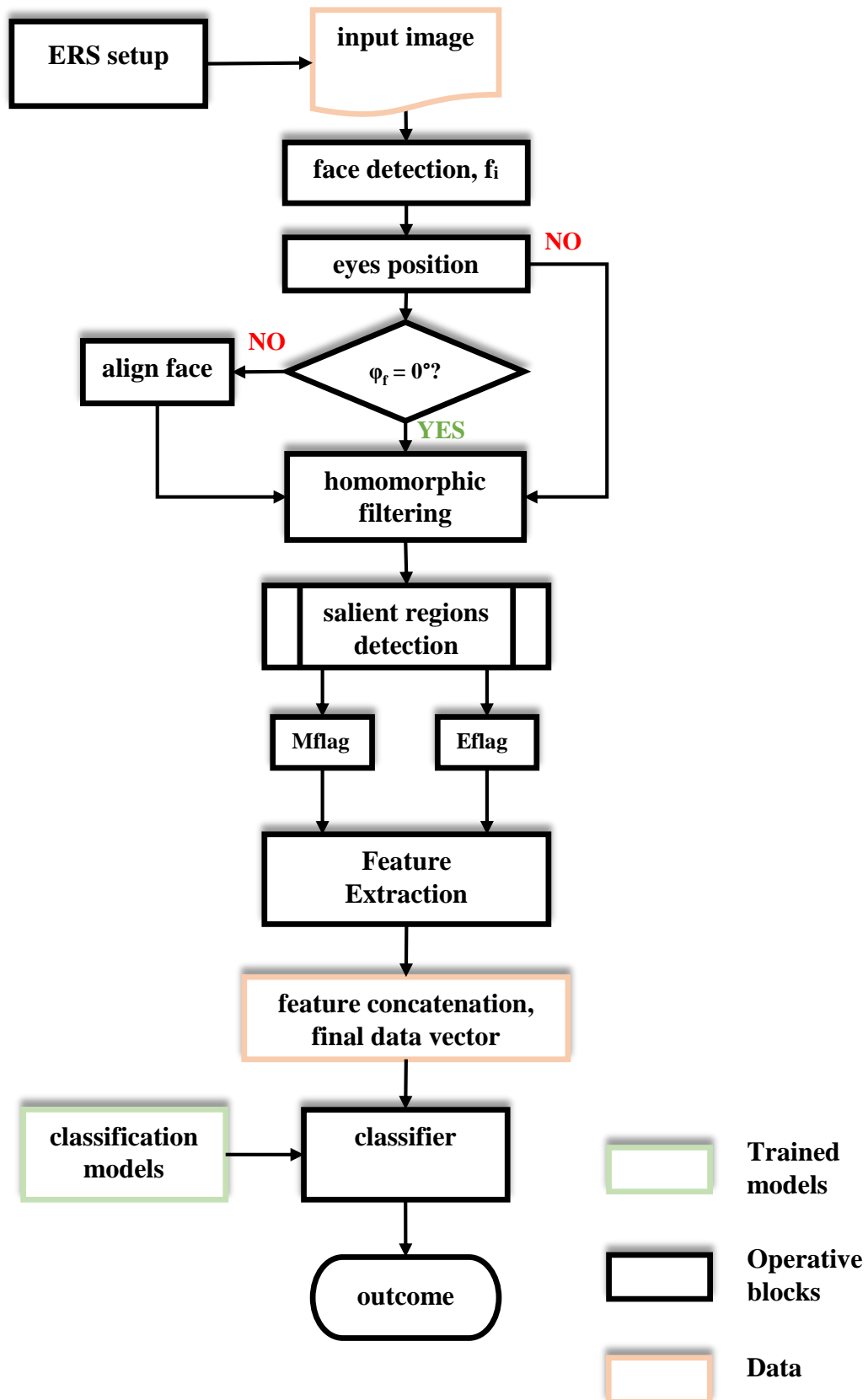


Figure 3.2: Block diagram of the ERS algorithm.

3.3 Face and salient regions detection

The decision of a suitable face detection method was the first challenge of this research. Several techniques were applied on ADFES [63] and CK+ [66] databases, in order to determine which of these is the most robust. Initially, skin color techniques were used but due to their low performance, they were constituted as a secondary method in the proposed system. On the other side, face detector model based on VJ algorithm [103] achieved accurate results. Therefore, VJ algorithm considered as the most appropriate method for this research. In order to be more flexible, the algorithm was designed to detect multiple faces in the input image and align the faces through a geometrical process using the eyes' centers.

In the next step, the salient facial regions are detected. In Khan's [167] psycho-visual experiment, gaze facial targets of human observers who made judgements about emotions were analyzed. Mouth and eyes were considered as the most informative facial regions for emotion recognition. In this research, aiming to mimic the human judgement techniques, three salient regions (mouth, eyes – eyebrows, and nose) were analyzed and their detection performed using the updated versions of the VJ algorithm [168, 169].

In the next subsections, the face and salient regions detection techniques used in the ERS algorithm are described.

3.3.1 Face detection based on skin color

In order to separate the face region, skin color segmentation was required. Therefore, the region of interest segmented in RGB color space, based on the rules stated by Kovac et al. [170]. In particular, the set of RGB values (R, G, B) were classified as skin region if the pixel values satisfy the following inequality,

$$R > 40 \ \& \ G > 40 \ \& \ B > 20 \ \& \ \max\{R, G, B\} - \min\{R, G, B\} > 15 \quad (3.31)$$

$$|R - G| > 15 \ \& \ R > G \ \& \ R > B$$

Then, by thresholding in color space, skin regions were determined. The possible skin regions were highlighted with rectangles and a filter was created to discard the non-face regions. Essentially, the simplest method to build such a filter is to examine the shape of the region. Thus, a specific range of width to height ratio of the face rectangle region is defined by Ibrahim [8] as follows:

$$0.83 \leq \frac{\text{width}}{\text{height}} \leq 1.27 \quad (3.32)$$

The above method is considered as time efficient; however, the detection accuracy is low. This is because objectives with pixel values similar to skin color can lead to false detections. In addition, its application to color images only, is narrowing the database selection process to those databases including exclusively color images. Therefore, the above technique was used as a secondary technique, providing an alternative solution in case the primary face detection approach, based on the VJ algorithm, fails to find a face in the input image. An example using skin color technique to detect a human face is presented in Fig. 3.3. By applying the aforementioned RGB set of rules on the given image a new binary image occurs, representing the skin region with white color and the non-skin region with black color. Then small objects are eliminated with morphological opening. The morphological open operation is an erosion followed by a dilation, using the same structuring element for both operations. By using the Equation 3.32, the region of the neck is removed and the final face is defined using a bounding box.

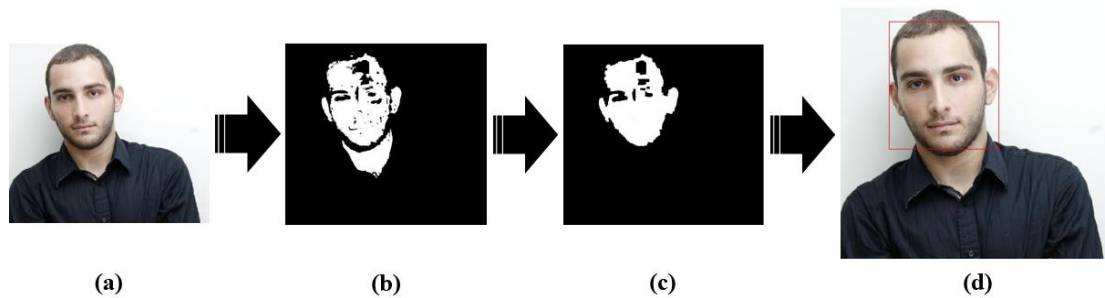


Figure 3.3: Skin-color-based method applied successfully on a face image. (a) The original image, (b) Image after the skin filtering, (c) Face region, (d) Original image with a bounding box around the face region.

3.3.2 Face detection based on VJ algorithm

The primary face detection method used in this project is based on the object detection algorithm developed by Viola and Jones [103]. The VJ algorithm was characterized as robust and fast in previous face detection systems [7, 104-106]. As mentioned in Section 2.4.4, Haar-like feature are extracted and then the integral image technique is implemented in order to reduce the computational time. Then, AdaBoost algorithm, using a weighted linear combination of weak classifiers, performs the feature selection. This approach reduces the amount of feature data significantly. Finally, to extract the possible faces, a cascade scheme of classifiers which is becoming increasingly complex (more feature are used in each step) is employed.

The “Computer Vision Toolbox” of Matlab 2016b includes a cascade object detector (`vision.CascadeObjectDetector`) [171], able to perform object detection using the VJ algorithm. This algorithm is capable of detecting faces (i.e. default parameter), eyes, mouths, nose or the upper part of the human body. Human faces approximately rotated by ± 15 degrees are possible to be captured. Depending on the object parameter, pre-trained classification models are selected. The option to create a new classification model is provided to the user. However, building a new face detector model requires a

considerable amount of face images from various databases to be trained and tested. The latter is not related to the purpose of this research. For that reason, the default model was used.

The procedure of face detection using the VJ algorithm is shown on Table 3.1. The detector is initialized, and a system object **faceDetector** based on face detection is created. In the given image, **Img** the VJ algorithm returns the location of possible faces **numFaces** by using the step function. A bounding box **bbox**, $M \times b_y - 4$ matrix, which M defines the number of the possible faces, containing the detected objects location. Each row of the **bbox** contain the exact location of each face as it follows,

$$BBOX = [x, y, width, height] \quad (3.33)$$

The possible faces are then cropped as f , and face alignment (see the next subsection) is performed if required. The aligned face image **Af** is stored for the next step which is the salient regions detection. Face detector was 100% accurate in ADFES database while in CK+ database only 11 out of 389 faces were not detected. Due to processing time reasons, the cropped face images were resized to 256×256 , in order to keep most of the information about the facial characteristics texture.

Algorithm 1: Face Detection based on VJ algorithm

input : single or multiple faces image, *Img*

output : face images, *Af*

BODY:

1. **faceDetector** = vision.CascadeObjectDetector
 2. **bbox** = step (faceDetector, Img(i))
 3. **numFaces** = size(bbox,1)
 4. **for** $i \leftarrow 1$ to *numFaces* **do**
 5. $f = \text{crop}(\text{Img}, \text{bbox})$
 6. **find** the eyes location using the ‘LeftEye’ and ‘RightEye’, Classification models of CascadeObjectDetector algorithm
 7. **align** face, f using the location of the eyes center as in Eq. 3.4 and **update** the aligned face, $Af(j)$
 8. **end**
-

Table 3.1: The Face Detection algorithm of ERS

An example of the face detection process for 22 faces of the ADFES database is presented in Fig. 3.4. The faces from the anger folder of the ADFES database were successfully detected and cropped based on the abovementioned process.

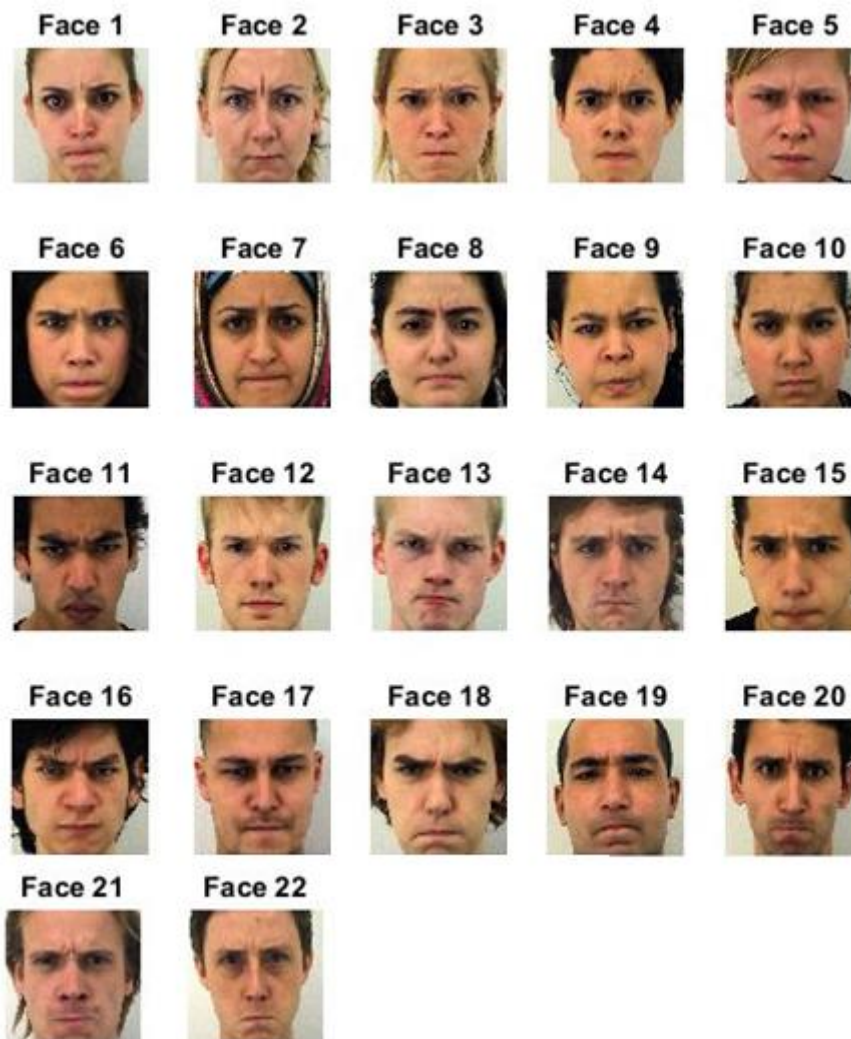


Figure 3.4: Example of face detection process using VJ algorithm.

❖ Image alignment

Recalling the aforementioned restriction of this algorithm that the face detection can be achieved with at most ± 15 degrees rotated faces, another issue occurs. Most of the databases include aligned and frontal posed human faces. However, in non-experimental cases, pictures may not be aligned and hence faces might be rotated. In this research, the classification model was specifically built based on database images. Thus, face alignment techniques had to be employed, in order to avoid a misclassification issue. Initially, color techniques have been used to segment the skin from the image background, so that the head position and orientation could be defined, although, the results were not encouraging. Thus, face alignment based on the eyes position was selected.

For this reason, the cascade object detector was employed to detect the left and right eye locations, within the upper right and upper left regions of the image respectively. Then the centers of the eyes are used as reference points for the image alignment. This process was very fast because the regions' size was limited. The geometrical equation to compute the angle between the eyes' centers is given by,

$$angle = \tan\left(\frac{\Delta y}{\Delta x}\right) * \left(\frac{180}{\pi}\right) \quad (3.34)$$

where Δy and Δx indicate the difference between the eyes in y and x axis respectively. After the detection, the coordinates of the eyes are stored for the salient regions detection. The following Fig. 3.5 presents an example of face alignment. The initial face image rotated by 10 degrees. By applying the proposed technique, the face successfully aligned based on the eyes' location.



Figure 3.5: Face alignment example.

3.3.3 Salient regions detection

The detection of salient regions is performed using the same object detector as for the face localization. However, because the eyes coordinates were determined in a previous stage of this algorithm, only the mouth detection is required. The reasons why salient regions were not detected in the first place, without performing a face detection are:

- 1) Computational time is reduced. In particular, by knowing the face location, the salient region detection is limited to only a specific region of the image.
- 2) The VJ algorithm uses the Haar-like feature to describe faces and facial characteristics. Face detection is a “coarse” scan in a large image. However, for facial characteristics such as the eyes, the detection is more detailed. This can lead to misclassifications. In addition, the possibility more than one faces existing in a generic image make this process more complex.
- 3) An algorithm's malfunction from a possible occlusion can be prevented. Sometimes due to occlusions (i.e. sunglasses or covered mouth), the information of the face regions can be limited. Thus, the algorithm is designed

to overcome this problem by examining only the detected regions. If a region fails to be detected the classification model is changed. The only case the algorithm cannot make judgements is when mouth, eyes and glabella regions cannot be detected.

Therefore, for accuracy and computational time issues, the face is detected independently.

❖ Homomorphic filtering

The salient regions detected in this research are mouth, eyes (including the eyebrows) and glabella. The first two regions are detected based on the VJ algorithm and the third using a geometrical approach. Before the salient region detection, homomorphic filtering is performed on the face image in order to remove the non-uniform illumination existing in images. The idea of using such a filter especially in the detection of salient regions occurred after repeated misclassifications. Delac et al. [172] proposed a sub-image homomorphic filtering technique for improving facial identification. Through their method, they increased significantly the recognition rate on the grayscale FERET database [69]. Similar projects [173-175] used techniques of homomorphic filtering for image enhancement with great success.

The illumination-reflectance model of image formation describes the intensity of each pixel (i.e. the amount of light reflected by a point on the object), as the product of the illumination of the scene and the reflectance of the object,

$$I(x, y) = L(x, y)R(x, y) \quad (3.35)$$

where I is the image, L is the illumination of the scene, and R is the reflectance of the scene. In order to counterbalance the non-uniform illumination, the illumination L has to be removed and the reflectance R has to be retained. Usually, illumination varies gradually across an image. On the other hand, the reflectance can coarsely change especially on the edges of the objects. Hence, considering the illumination as a noise signal, the difference between the two components L and R is the key element examined in order to separate them from each other. First, the multiplicative components are transformed to additive components in log domain using the following equations,

$$\ln(I(x, y)) = \ln(L(x, y)R(x, y)) \quad (3.36)$$

$$\ln(I(x, y)) = \ln(L(x, y)) + \ln(R(x, y)) \quad (3.37)$$

Then, a high-pass filter (i.e. Gaussian, Butterworth, and Chebyshev) is used in the log domain to remove the low-frequency illumination component and at the same time to retain the high-frequency reflectance component. An illustration of a homomorphic filter is depicted in Fig. 3.6 and an example is illustrated in Fig. 3.7. As seen in Fig. 3.7 if the input image is true-color (RGB), is converted to a grayscale one. The gradual change in illumination of the left image has been corrected considerably in the right

image. In particular, the facial characteristics are clear (i.e. the right ear is visible, while the left cheek's non-uniform illumination is compensated).

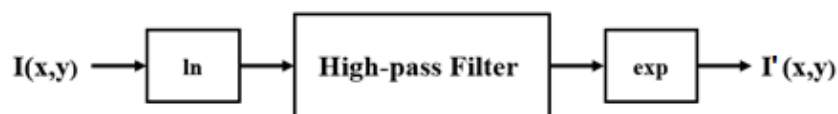


Figure 3.6: Homomorphic filtering

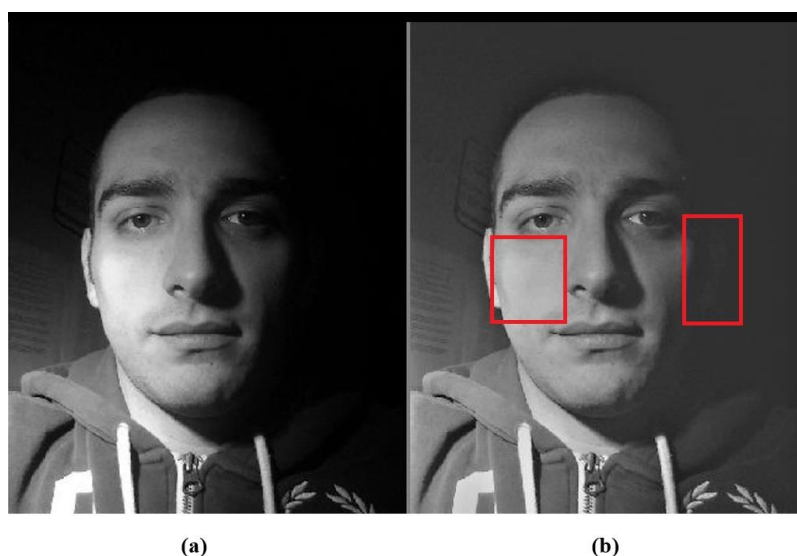


Figure 3.7: Example of Homomorphic filtering. (a) The original face image. (b) The face image after the homomorphic filtering application.

❖ Salient regions

After reducing the effect of illumination on the face's image, the salient regions detection is performed. Due to the fact that the salient regions' sizes vary from face to face, while deforming differently depending on the type of expression, a fixed area around their centers has to be captured. Moreover, in order to be precise with the evaluation of the facial expression, the texture-based techniques used in this algorithm, must be applied on equal size regions. For those reasons, a common size, which covers all the possible cases for the mouth, eyes and glabella regions without a critical loss of information, was decided. The maximum height and width for a face image of size 256×256 pixel, were determined empirically, by taking into account the dimensions of the detected region from both ADFES and CK+ databases. According to the face dimensions, in Table 3.2, the size of the regions along with the emotions caused the maximum deformation, are presented. The deformation calculated based on both width and height. Mouth region is presented its maximum deformation (64x96) during the joy and surprise emotions, eye region (64x64) during the surprise emotion while glabella has maximum deformation (56x40) during disgust and anger emotions.

Region	Size	Emotion
Mouth	64x96	max width: Joy max height: Surprise
Eyes	64x64	max width: Surprise max height: Surprise
Glabella	56x40	max deformation: Disgust and Anger

Table 3.2: Regions size – Emotion caused the maximum deformation.

Firstly, for the mouth detection, VJ algorithm is applied to the lower half of the face. The mouth region is cropped according to the fixed size dimensions, considering its center as the reference point. Then, for the eyes region detection, VJ algorithm is applied to the upper half of the face and the eyes' regions are cropped according to the fixed size dimensions, considering their center as the reference point. In order to avoid a second detection, the eyes coordinates are provided from the alignment process. Finally, glabella's region center is calculated from the mean of the straight line, which lies at the eye center. The glabella region is cropped to the fixed size dimensions, considering its center as the reference point. An example of the salient regions (confined in yellow rectangles), is presented in Fig. 3.8. In both images, the faces are joyful and the mouth area is deformed at its maximum on 'x' axis. On the other hand, in the eyes' region is clear that the upper limit offers more space for a possible eyebrows deformation.

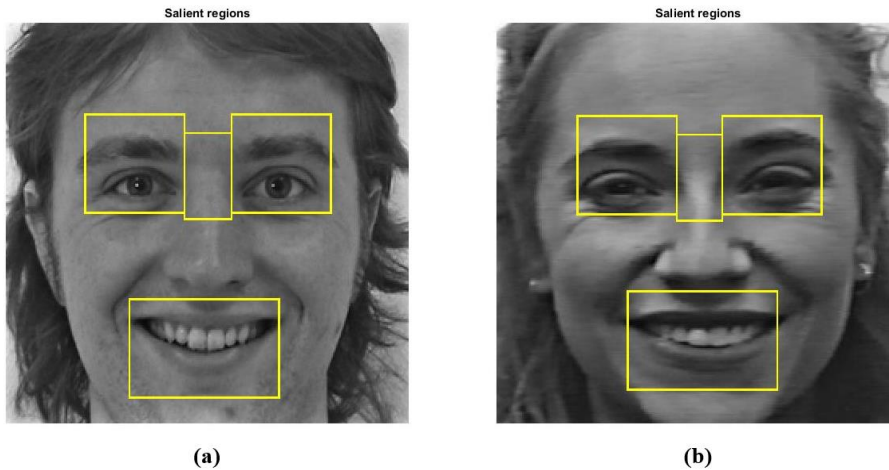


Figure: 3.8: Salient regions in bounding boxes. (a) Face from the ADFES database, (b) Face from the CK+ database.

In Figure 3.9, the detection of the salient region for three different emotions (joy, anger and sadness) is presented. As it noticed the mouth's appearance is unique for every emotion. The eyes' regions have a similarity in joy and anger emotions while the nose is identical for joy and sadness emotions. Later, in the experimental process the impact of each region during the emotion recognition will be tested and discussed in detail.

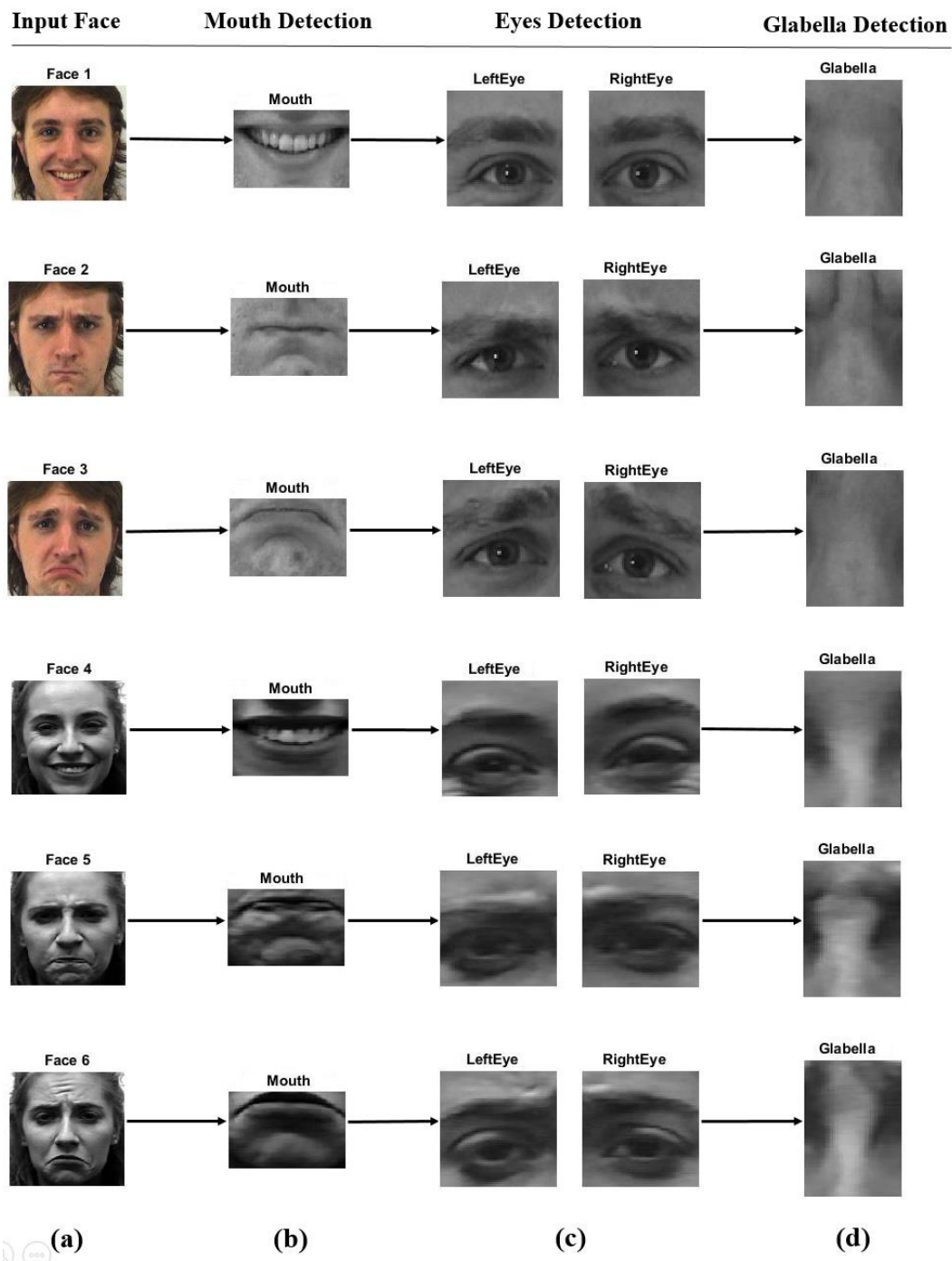


Figure 3.9: Salient regions detection with VJ algorithm. (a) Two individuals from CK+ and ADFES databases, (b) Mouth region detection, (c) Left and right eyes (including the eyebrow) region detection and (d) Glabella region detection.

3.4 ERS using HOG descriptor

As mentioned in Section 2.2.4, Ekman and Friesen [54] defined that emotion produced from the stimulation of different facial muscles. Usually, the deformation on several

facial characteristics is visible. Mouth, eyes and glabella regions detected in the previous step of this algorithm, constitute the most informative sources during a facial expression. A powerful feature extraction technique, which examines the shape characteristics of the skin's formation, is histograms of oriented gradients. HOG descriptor is based on the accumulation of gradients' directions over the pixel of a predefined area called the cell. The accumulated values are subsequently registered in a 1-D histogram. By concatenating all the histograms values, a feature vector is constructed. In this research, shape information was extracted from the available regions using HOG descriptor.

3.4.1 HOG Descriptor's parameters assessment

HOG descriptors were used in the past for human detection in generic images. A significant advantage of HOG over the other techniques is its exceptional performance without image pre-processing. Thus, it was chosen as one of the approaches, applied on ERS. Steps for the HOG feature extraction are as follows:

- 1) Gradient computation is performed by applying a 1-D central difference filter $[-1, 0, 1]$ and $[-1, 0, 1]^T$. Their directions can be between -180 and 180 degrees. At the image borders, gradients are computed using forward difference, which essentially replicates the pixel values.
- 2) Cell histograms are created. Each pixel within the cell is assigned to one of the available orientation bins based on its value. Therefore, cell size and orientation bins are the parameters, which directly influenced the performance of the HOG descriptor. Various combinations of parameters have been tested for each of the salient regions.
- 3) Cell grouping into spatially connected blocks. In this way, the gradients magnitude is locally normalized and the effects of illumination and contrast are confronted. The block size is set to 2×2 cells and the block overlap is set to the half of the block size. Finally, HOG descriptor vector is built by the concatenation of the components of the normalized cell histograms from all the blocks.

In order to find the optimal values for the HOG parameters, i.e. the best configuration to derive the most discriminative feature for each region, 25 combinations of them were tested. In particular, the accuracy of each component using two classification methods was measured in order to evaluate their performance. Table 3.3-3.5 report the number of feature from where each combination was produced. By increasing the cell size, the algorithm captures large-scale spatial information and the processing time is decreased. The significance of this parameter was emphasized by Deniz et al. where HOG descriptors were extracted from a regular grid, and a fusion of them at different scales demonstrated the best model for face recognition. On the other side, increasing the number of the orientation bins enhances the capability to encode finer orientation details. However, the size of the feature vector is also increased, which requires more time to process. For these reasons, the parameters' selection has to perform rigorously, with respect to the system's processing time.

Cell Size	Orientation bins				
	5	9	15	20	55
3x3	12,400	22,320	37,200	49,600	136,400
5x5	3,960	7,128	11,880	15,840	43,560
8x8	1,540	2,772	4,620	6,160	16,940
12x12	560	1,008	1,680	2,240	6,160
15x15	300	540	900	1,200	3,300

Table 3.3: Number of feature for the mouth area based on cell size and orientation bins combinations.

Cell Size	Orientation bins				
	5	9	15	20	55
3x3	16,000	28,800	48,000	64,000	176,000
5x5	4,840	8,712	14,520	19,360	53,240
8x8	1,960	3,528	5,880	7,840	21,560
12x12	640	1,152	1,920	2,560	7,040
15x15	360	648	1,080	1,440	3,960

Table 3.4: Number of feature for the eyes area (including both eyes) based on cell size and orientation bins combinations.

Cell Size	Orientation bins				
	5	9	15	20	55
3x3	4,080	7,344	12,240	16,320	44,880
5x5	1,400	2,520	4,200	5,600	15,400
8x8	480	864	1,440	1,920	5,280
12x12	120	216	360	480	1,320
15x15	40	72	120	160	440

Table 3.5: Number of feature for the glabella area based on cell size and orientation bins combinations.

An example of how the HOG feature extraction process was performed and evaluated for cell size equal to eight and nine orientation bins is illustrated in Fig. 3.10. As indicated in Table 3.2, the maximum deformations occurred during: a joyful emotion as in Fig. 3.10(a) for the mouth region, a surprise emotion as in Fig. 3.10(b) for the eyes region and a disgust emotion as in Fig. 3.10(c) for the glabella region. Based on this explanation, in the second row of Fig. 3.10 is demonstrated the ability of the HOG descriptor to derive information about the shape of the texture. In particular,

HOG feature are visualized using a grid of uniformly spaced angle histogram plots [176]. The cell size and the size of the image determine the grid dimensions. Each rose plot shows the distribution of gradient orientations within a HOG cell. The length of each petal of the rose plot is scaled to indicate the contribution each orientation makes within the cell histogram. The plot displays the edge directions, which are normal to the gradient directions. Viewing the plot with the edge directions allows understanding the shape and contours encoded by HOG. To obtain the final HOG descriptor (see the third row of Fig.3.10), histograms of gradients (HOG) within each block are concatenated with half-block overlap.

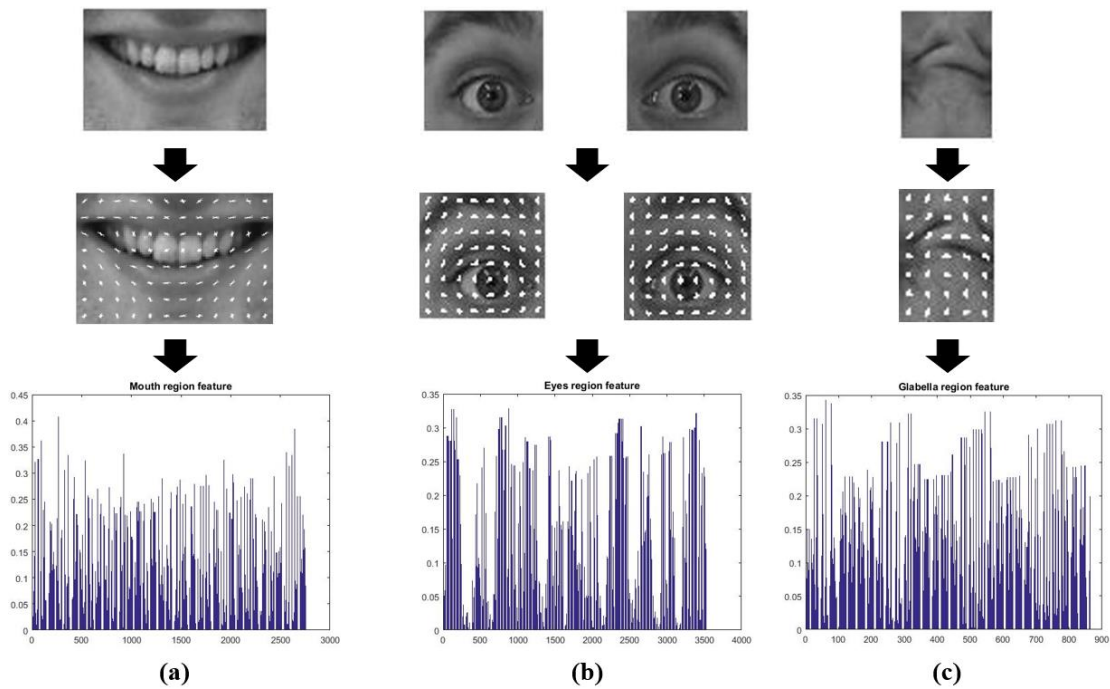


Figure 3.10: Feature extraction using HOG descriptor (cell size: 8, orientation bins: 9) in three steps: Input image, HOG visualization and concatenated histogram of cells. (a) Joy emotion, (b) Surprise emotion and (c) Disgust emotion.

Fig. 3.11 presents a schematic example of feature extraction for the left eye region. There are 64 cells, which cover the image and every block includes four cells, which overlap by half. Therefore, the final descriptor is constructed by 49 blocks and each block has 36 values (4×9 histograms), resulting in 1,764 feature. In the next subsection, the classification parameters for the experimental process are discussed.

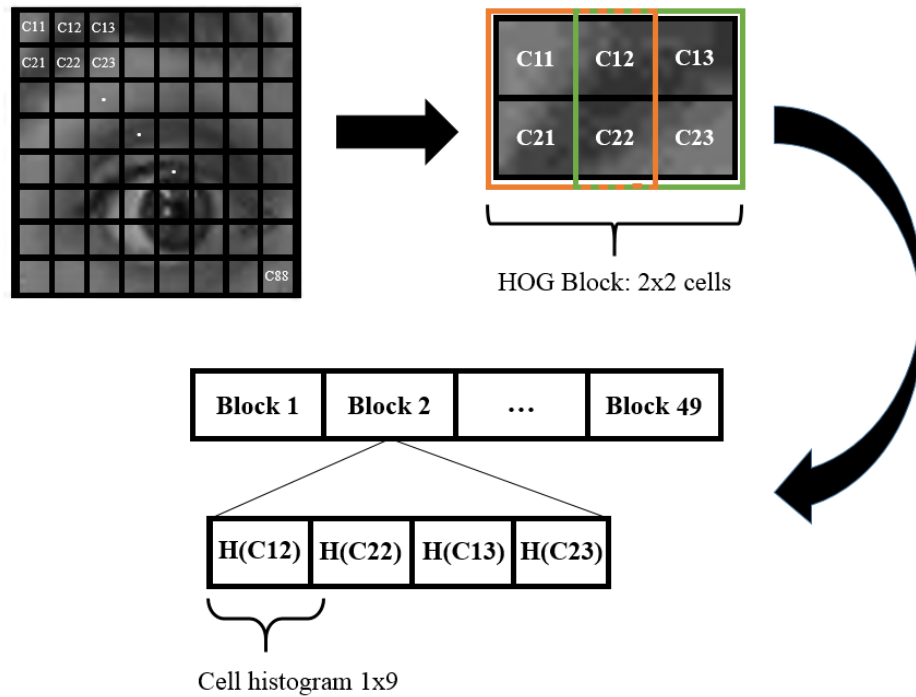


Figure 3.11: A Schematic example of HOG extraction in the right eye.

3.4.2 Classification parameters assessment

The regions of interest were acquired automatically using the VJ algorithm and processed to extract the HOG feature vector. ERS system was designed to employ the classification model, depending on the available facial regions. The MEG-classifier was trained with all the available regions (mouth, eyes and glabella) while the M-classifier and EG -classifier were trained considering only the mouth and eyes-glabella regions respectively. The classification techniques utilized for training and testing purposes were the LDA and SVM. Both approaches are fast and memory-friendly. The idea of SVM is to create a hyperplane in between the data sets in order to indicate in which class it belongs to (see Section 2.6.5). On the other hand, LDA assumes that different classes generate data based on different Gaussian distributions (see Section 2.6.6). Another reason why these two techniques were selected is their ability to handle a significant amount of data. The training process performed using LDA with diagonal covariance and linear SVM with the one-vs-one multiclass method.

For the testing process, the average recognition rate was computed using the 10-fold cross-validation technique. Essentially, a feature vector was divided randomly into 10 equal subsets. Nine subsets were used for training while the remaining subset was used for testing. The process was repeated ten times until all the subsets used exactly once for testing. The accuracy was calculated from the average recognition rate of all ten folds. However, in order to be more precise, the whole process was performed three times with different folds format. The final recognition rate determined from their average.

In most cases, the feature vector had notable large amount of data. Thus, principal component analysis has been applied for dimensionality reduction. PCA's explained variance of the data was calculated by decreasing it gradually from 95% to 35%. The contribution of PCA on the classification process was critical concerning the performance of the system. Essentially, the original space reduced (i.e. retaining the most important variance), to the space spanned by a few eigenvectors. Data redundancy and time complexity issues are confronted. In order to determine the effectiveness of this technique, computational time of the LDA classifier in training and testing process with and without PCA is presented in Table 3.6. It is noticed that the use of PCA reduces significantly the computational time in both training and testing processes. Computational time is always affected from the equipment's condition and age. Therefore, it is worth mentioned that for this research a Samsung laptop with Intel Core i3 and memory 4GB was used.

Cell	Bins	Feature Number	Train PCA-on	Train PCA-off	Test PCA-on	Test PCA-off
8	9	864	6.4	3.8	0.1	0.1
5	9	2,520	10.9	11.0	0.4	0.5
3	20	16,320	69.6	258.1	4.4	13.2

Table 3.6: Three models from glabella region are evaluated. Training was performed for 330 images of CK+ database. Testing was performed on one image of CK+ database. Train and test columns contain computational time in seconds.

The following subsections, present the experiments performed by utilizing CK+ and ADFES databases. LDA and SVM classifiers were applied to all the possible combinations of the HOG parameters in order to result in the most efficient model.

3.4.3 HOG training and testing using the CK+ database

For the experimental process, two databases were trained and examined. However, considering the popularity in literature, training operation is describing using CK+ database[66]. For 378 images, the detection process has performed successfully. However, it was important to balance the dataset before executed the classification, in order to prevent from a misleading. Therefore, a study carried out using 330 faces. CK+ database includes both posed and non-posed (spontaneous) sequences of expressions (see Section 2.3). The experiments performed in this research utilized only the last images of each sequence (i.e. apex of the expression). Moreover, in this database, validated emotion labels are added to the metadata. For these reasons, an algorithm that stores the last picture of each sequence and identifies the associated emotion had to be developed. Table 3.7, presents the number of images was used, for each of the six basic emotions. Furthermore, in the following sections confusion matrices were used to for the better understanding of the system's performance. The color intensity defines the

size of the percentage. Thus, the deeper the color, the largest percentage is presented. For example, a light green color describes a range of 0% - 35%, a green color a range of 36% - 65% while a deep green color a range of 66% - 100%.

Emotion	No of images
Anger	55
Disgust	56
Fear	53
Joy	59
Sadness	48
Surprise	59

Table 3.7: Number of images per emotion.

In order to determine what combination of parameters provides the most informative model, each region was trained and tested individually. Table 3.8 demonstrates that the best recognition rates were achieved, using SVM and LDA. LDA offered a slightly better accuracy than SVM. Therefore, for comparison and reporting results, LDA is preferred. Fig. 3.12-3.17 present the overall behavior of the regions studied, along with the confusion matrixes of their best model. These figures in conjunction with Table 3.8 lead to the following observations:

- 1) Mouth region constitutes the major source of information. It was expected that a mouth model would dominate expression recognition since its shape is almost unique for all of the six basic emotions. As it can be seen in Table 3.8, LDA classification model has an exceptional accuracy, 84.7%. In Fig. 3.12, the combination of parameters for five PCA variation levels demonstrates that the best results occur when the cell size is equal to 12x12, the orientation bins are 20 and the PCA variation is at 95%. Although similar accuracy is noticed in other combinations, the chosen one is produced by the smallest feature vector. In Fig. 3.13, the confusion matrix for the mouth region illustrated. From the results, it is clear that some emotions share common characteristics. For example, for anger and sadness the mouth is closed, and likewise, for fear and joy the mouth is open and deformed horizontally. Thus, the few misclassifications appeared between them in confusion matrix, are reasonable.
- 2) Eyes' model using LDA classifier has moderate performance at 63.9%. Fig. 3.14 shows that the performance rate for the most combinations of parameters lies between 61% and 64%. Therefore, the selection of them has to be based on the number of feature extracted. The model that associates a reasonable number of feature and competitive accuracy has cell size is equal to 15x15, number of orientation bins equal to 20 and its PCA variation is at 75%. The connection between the performance and the emotions can be determined by observing the Fig. 3.15. Eye and eyebrow as one region is informative for emotions such as surprise and joy. However, their shape for emotions like fear and sadness is very similar, and lead to misclassifications (i.e. $\pm 30\%$ false

negatives between them). The same exactly stands for anger and disgust emotions. Therefore, eyes contribution in facial expression analysis is low.

- 3) Glabella model has significantly low performance at 54.8%. Fig. 3.16 presents the overall performance of all the possible combination of parameters for the Glabella region. The lower levels of variation (i.e. 75% and 55%) show higher accuracy. This is justified by the fact that most of the data have similar values since glabella region deforms notably only in two of the six emotions. Therefore, in order to discover the changes in this region more data are required. The model that describes reasonably the Glabella area has cell size equal to 8x8 (i.e. more information), 15 orientation bins and PCA variance at 55%. However, the average accuracy of this model does not reflect the information provided from its confusion matrix in Fig. 3.17. In particular, joy and disgust emotions have exceptional prediction rates while sadness and fear have a terribly low prediction and their false negative rates is almost equally divided.

	SVM (%)	C	B	LDA (%)	C	B
Mouth	82.8	12	15	84.7	12	20
Eyes	61.5	12	55	63.9	15	20
Glabella	53.2	8	15	54.8	8	15

Table 3.8: Best accuracy of each region for LDA and SVM. C and B represent the cell size and orientation bins respectively.

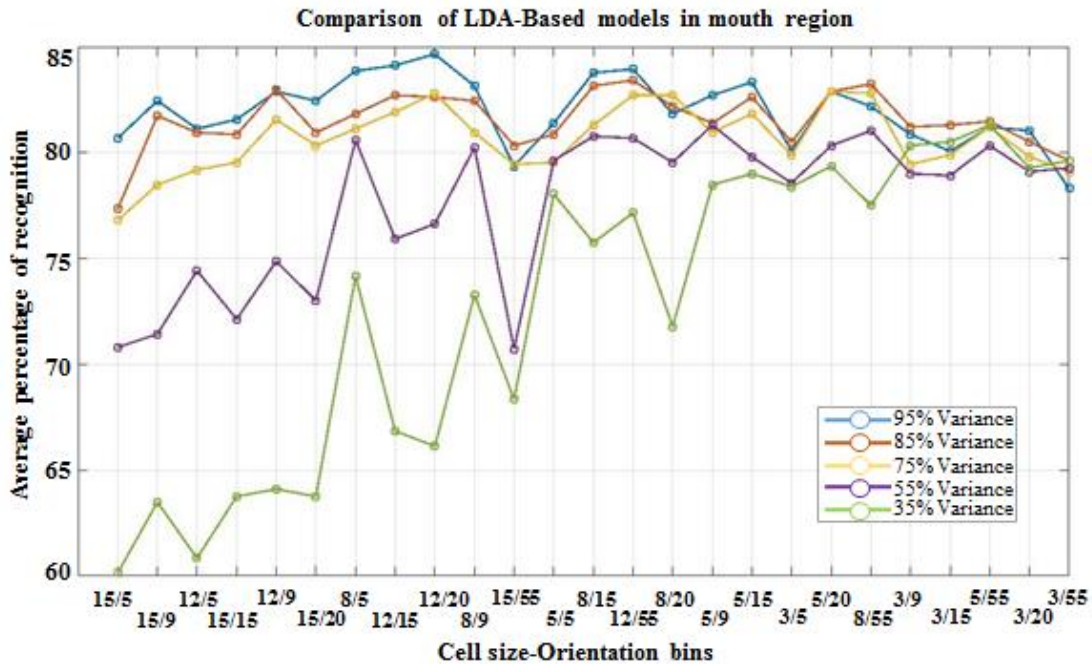


Figure 3.12: Mouth region. Average recognition rates evaluation for all the possible combinations of HOG descriptor's parameters with five PCA variance levels. Parameters are placed based on feature vector's size (min – max).

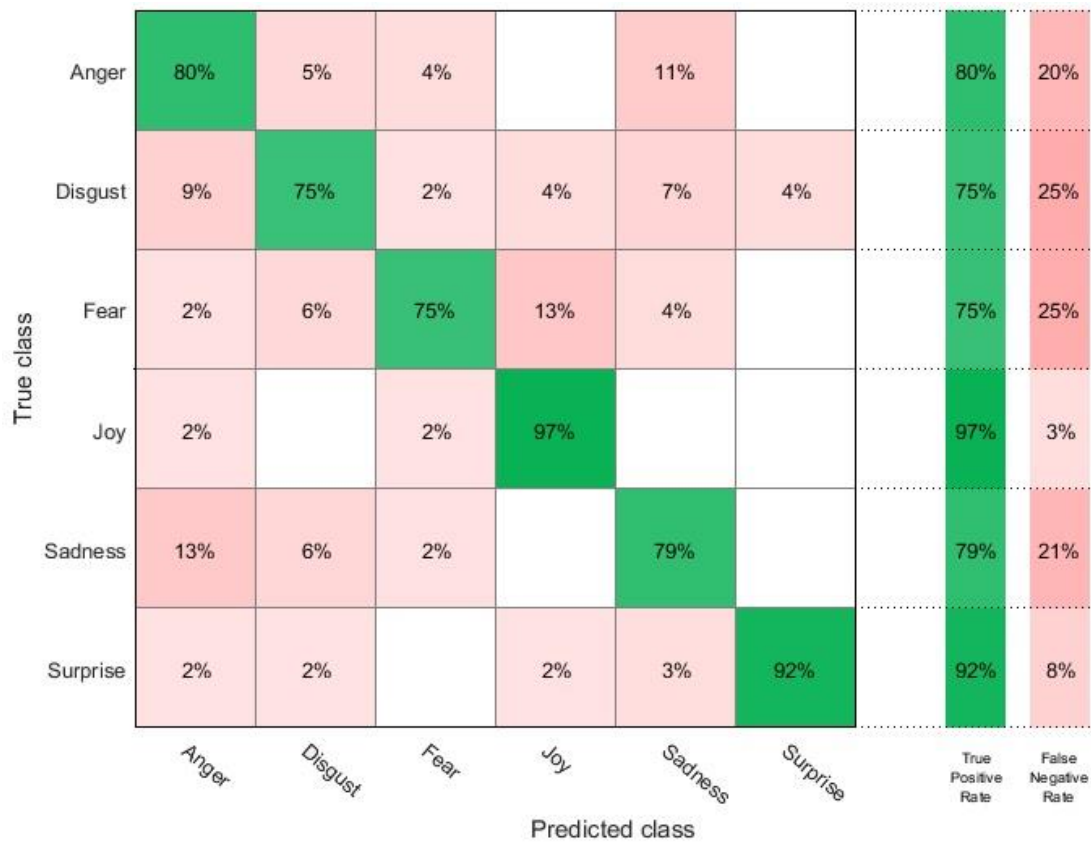


Figure 3.13: Mouth region’s confusion matrix.

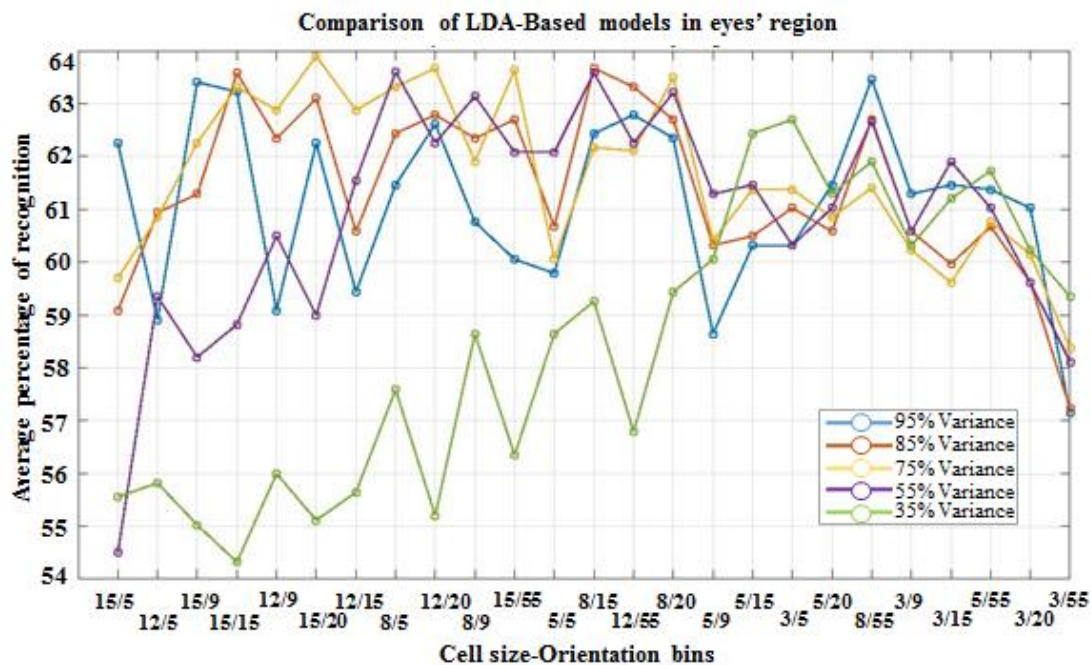


Figure 3.14: Eyes’ regions. Average recognition rates evaluation for all the possible combinations of HOG descriptor’s parameters with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max).

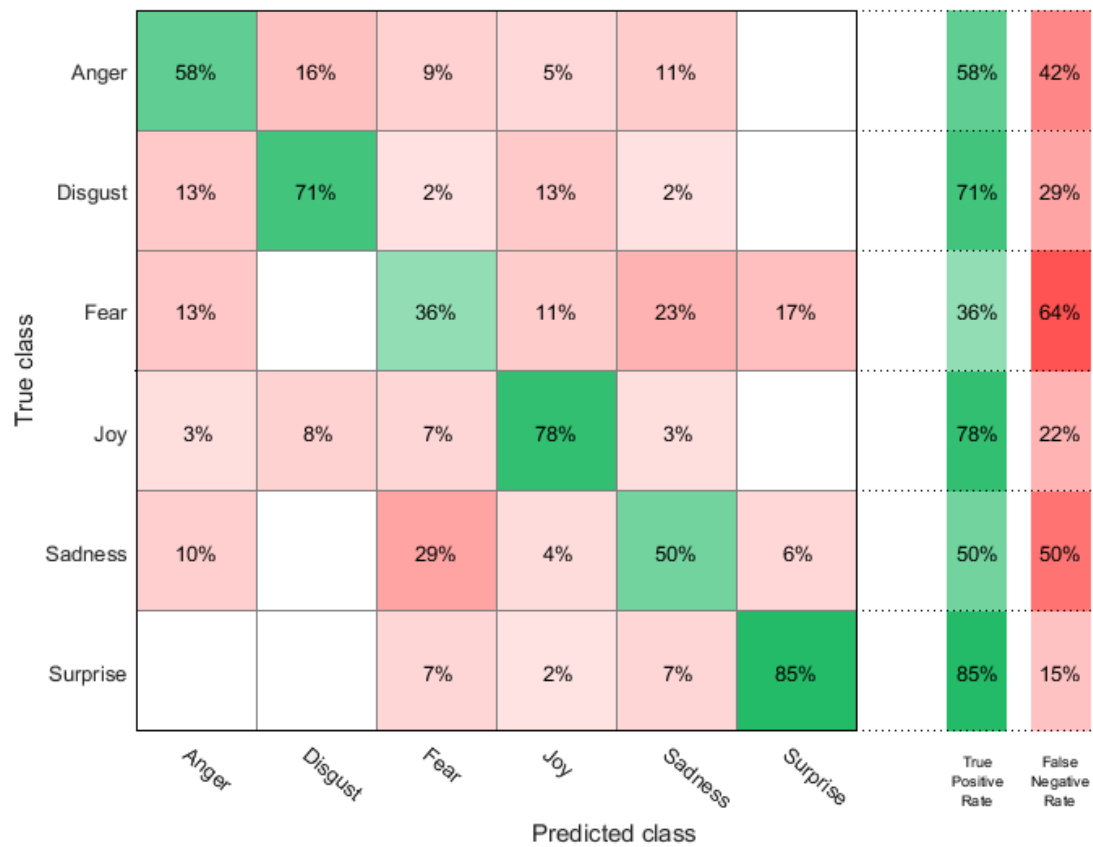


Figure 3.15: Eyes regions’ confusion matrix.

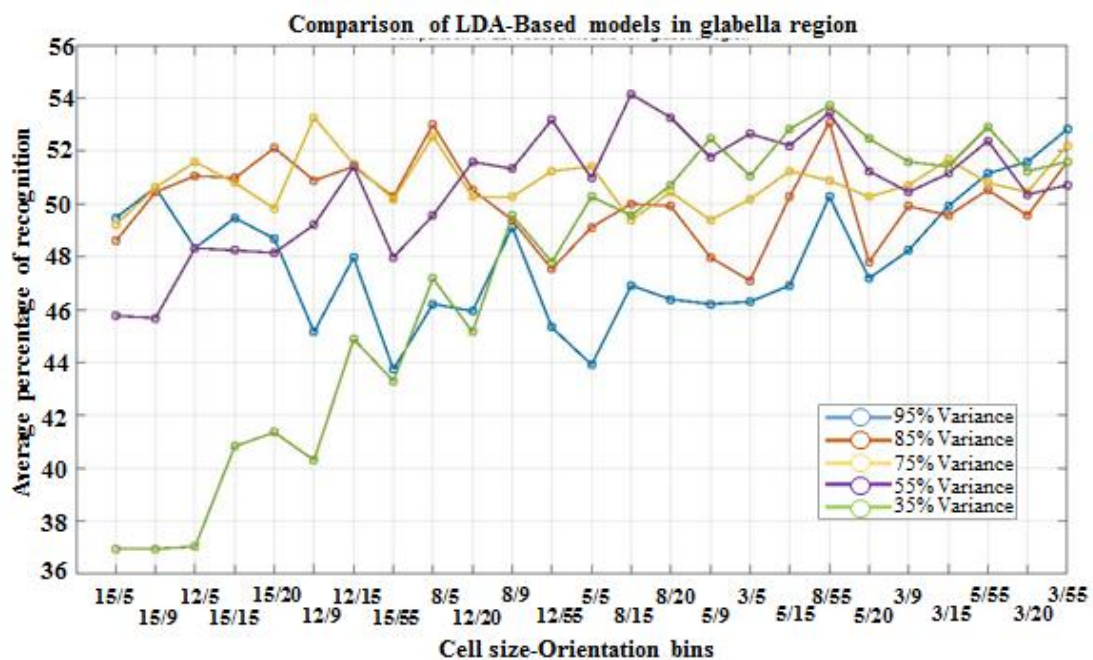


Figure 3.16: Glabella region. Average recognition rates evaluation for all the possible combinations of HOG descriptor’s parameters with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max).

True class	Anger	58%	7%	9%	16%	9%		58%	42%
	Disgust	9%	80%	11%				80%	20%
	Fear	11%	2%	32%	19%	19%	17%	32%	68%
	Joy	5%	2%	8%	78%	3%	3%	78%	22%
	Sadness	27%		19%	23%	13%	19%	13%	88%
	Surprise	2%	2%	15%	10%	12%	59%	59%	41%
		Anger	Disgust	Fear	Joy	Sadness	Surprise	True Positive Rate	False Negative Rate
		Predicted class							

Figure 3.17: Glabella region's confusion matrix.

Therefore, based on the above results and the intention to create a system that operates under partial occlusion, three classification models are proposed. Depending on the availability in detection stage, if both mouth and eyes flags are positive the MEG-classifier is activated. On the other side, if one of these is negative then M-classifier or EG-classifier is activated.

The MEG-classifier was trained based on the models which provided the best recognition rate. The M-classifier remained the same as presented in the previous paragraph. However, due to their moderate performance, glabella and eyes' most informative feature were concatenated to form a new model the EG-classifier. Table 3.9 shows the results of the ERS system using 10-fold cross validation, for the three possible cases. In the first case, MEG-classifier achieved accuracy of 90.3% and although the feature vector dimensionality was at first large with 5120 components, after PCA employed, is reduced significantly to 211 components (PCA variance = 92%). In the second case that only mouth region is considered, the accuracy remains high at 84.7%. Initially the feature vector formed by 2240 components but PCA decrease them to 235 (PCA variance = 97%). Finally, in the third case that the eyes' and glabella regions are analyzed the accuracy dropped to 67.0%. By using PCA the original 2280 components are reduced to 37 (PCA variance = 64%). Furthermore, Fig. 3.18-3.19 present the confusion matrixes for the EG-classifier and MEG-classifier while M-classifier confusion matrix is the same as in Fig. 3.13.

As discussed earlier, the proposed framework demonstrated an exceptional average recognition rate of 90.3% in MEG-classifier model. A comparison of the confusion matrixes results (Fig. 3.13, Fig. 3.18-3.19), shows that the MEG-classifier model enhanced system's accuracy in every emotion. Particularly interesting, was the disgust, joy, sadness and surprise emotions with 91%, 97%, 92% and 92% recognition rates respectively. On the other hand, M-classifier model with a remarkable average recognition rate of 84.7% was proved very robust in the case that eyes fail to be detected. Finally, EG-classifier model with 67.0% average recognition rate constitutes a slightly better option than their original models (i.e. eyes' and glabella on their own) in the case that the mouth area is not detectable.

By following the same training and evaluation processes, ADFES has demonstrated similar results with CK+ database. In particular, MEG-classifier achieved 90.2% recognition rate, while M-classifier and EG-classifier the accuracy was 87.50% and 69.70% respectively. In order to examine the capability of emotion recognition in ERS, the databases will be tested on each other in the next experiment.

Model	SVM (%)	LDA (%)
MEG-classifier	87.6	90.3
M-classifier	82.7	84.7
EG-classifier	61.5	67.0

Table 3.9: Best accuracy of each model using LDA and SVM.

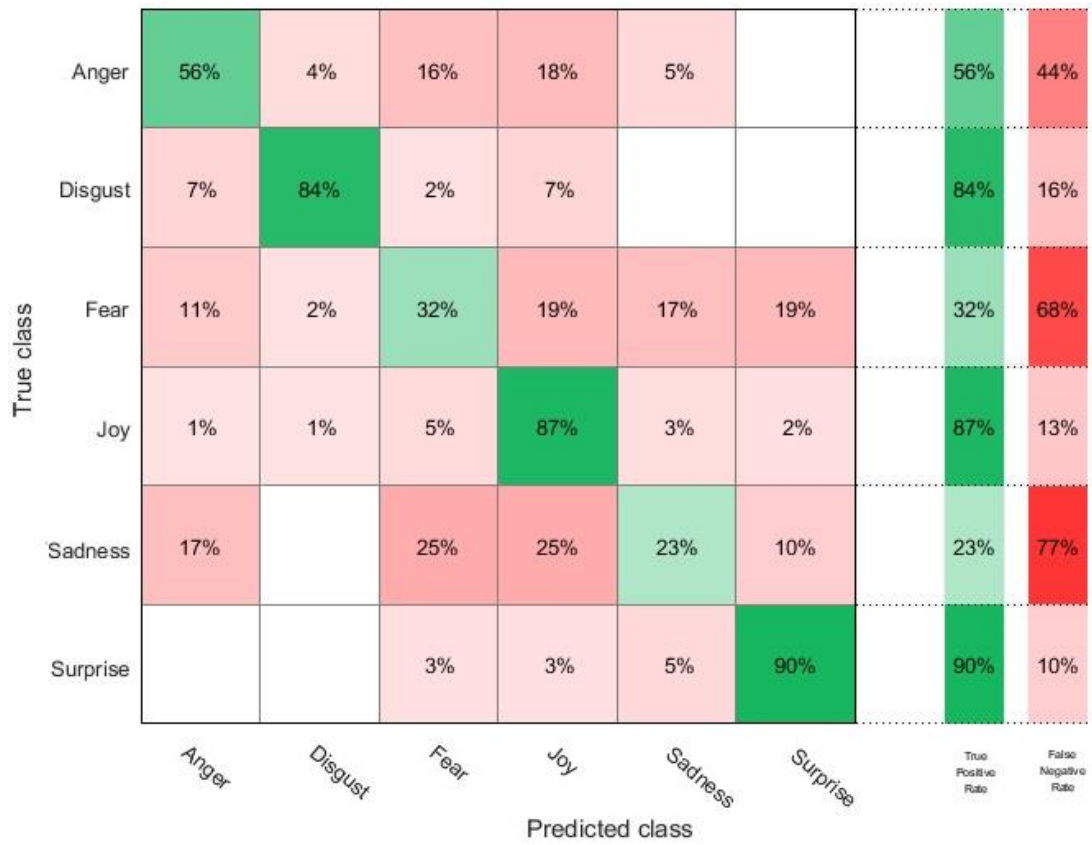


Figure 3.18: EG-classifier confusion matrix.

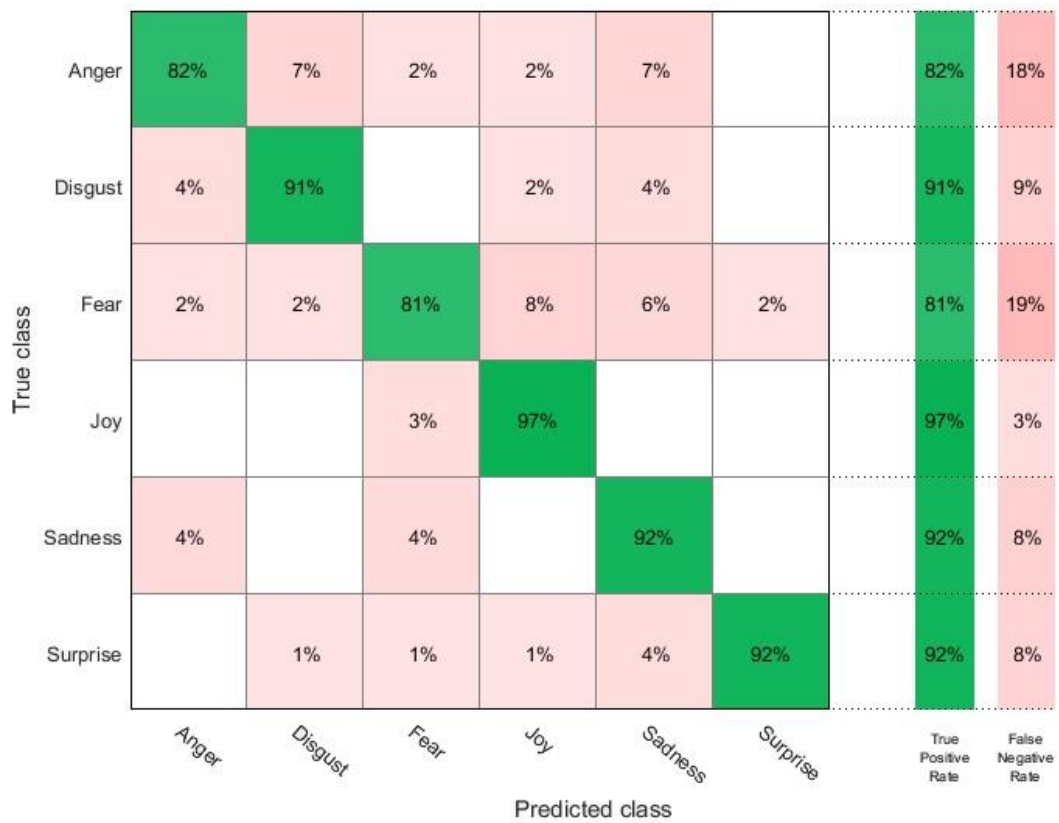


Figure 3.19: MEG-classifier confusion matrix.

3.4.4 Generalization of the algorithm

The second experiment's aim was to study the performance of the proposed emotion recognition system, by testing a new database on the previously trained models. Therefore, ADFES database is tested on the CK+ database and vice versa. This process has helped to evaluate the dynamic of ERS framework if it would be used in real life images.

Using the LDA classification method and based on the most efficient parameters for each database, two different schemes are introduced in this experiment:

- 1) In the first scheme, images from the CK+ database were utilized for the training and images from the ADFES database were used for the testing. Table 3.10 shows the results obtained from this process. The results obtained using MEG-classifier model. The second column shows the accuracy achieved by training CK+ database and testing it with its own images, while the third column represents the accuracy obtained by testing ADFES database.
- 2) In the second scheme, images from ADFES database were used for the training and images from CK+ database were utilized for testing. In Table 3.11, the results of this process are presented. The results obtained using MEG-classifier model. The second column shows the accuracy achieved by training ADFES database and testing it with its own images, while the third column represents the accuracy obtained by testing CK+ database.

Emotions	Intra-training (%)	Train: CK+ Test: ADFES (%)
<i>Anger</i>	82.7	63.7
<i>Disgust</i>	91.5	68.2
<i>Fear</i>	81.8	50.0
<i>Joy</i>	97.5	95.5
<i>Sadness</i>	93.8	86.4
<i>Surprise</i>	94.2	95.5
Average	90.3	76.6

Table 3.10: Average recognition accuracy for each emotion training CK+ and testing ADFES.

Emotions	Intra-training (%)	Train: ADFES Test: CK+ (%)
<i>Anger</i>	91.4	47.3
<i>Disgust</i>	91.7	96.4
<i>Fear</i>	82.5	90.6
<i>Joy</i>	95.0	100
<i>Sadness</i>	95.6	75.0
<i>Surprise</i>	86.4	92.2
Average	90.4	83.6

Table 3.11: Average recognition accuracy for each emotion training ADFES and testing CK+.

The above results show that the system is capable of performing emotion recognition not only on an experimental basis. However, they also demonstrate the requirement for more data. For example, in the first scenario where the CK+ database utilized for the training process, three emotions (i.e. joy, sadness and surprise) achieved great accuracy, while two emotions' (i.e. anger and disgust) recognition rate was moderate and one emotion's accuracy (i.e. fear) was poor. On the other side, when ADFES database used for training purpose, poor result appeared only in one emotion (i.e. anger). Four emotions (i.e. disgust, fear, joy and surprise) had exceptional accuracy while one emotion's (i.e. sadness) accuracy was acceptable.

As mentioned in Section 2.3, for an effective facial expression recognition algorithm a database of sufficient size is required. This experiment has helped to realize the importance of the size, quality and universality of the database, in emotion recognition analysis. ADFES database had slightly better recognition rate due to the quality of the images.

3.4.5 SelfieDat: A new database

In order to determine the performance of the ERS system under real life's circumstances, a new dataset based on selfie-images was developed. The people participated in this experiment were friends and colleagues who were interesting in this research and asked to perform the six basic emotions under arbitrary light conditions and pose, using their personal cell phone's camera. The dataset named SelfieDat includes 12 individuals. Essentially, SelfieDat images were tested using the pre-trained classifiers. Two parameters of ERS are evaluated: the adequacy and robustness of the trained database (i.e. CK+) and the capability of the system to perform emotion recognition with various resolutions. Some of the characteristics that distinguish this database from other are:

- 1) The hardware means are ordinary and arbitrary. Therefore, some of the images have high resolution and other low resolution.
- 2) The candidates were not posed in a predefined stance. Some of them were slightly turned their faces to various angles.
- 3) The light conditions and the distance of the face from the camera were arbitrary. Some images were taken with natural light and other with camera's flash and faces were close and other far from the camera's lens.
- 4) Data were collected from people of various ages. In particular, the younger candidate was 18 years old while the oldest was 64 years old.

In Fig. 3.20, samples for each basic emotion of SelfieDat are illustrated. The complete database can be found in Appendix A.

The experimental process was divided into three parts. As mentioned earlier in Section 3.4.3, depending on the detected regions three classification models can be activated: MEG-classifier, M-classifier and EG-classifier. In each part, one of these classifiers was utilized. The images were tested using the models produced from CK+ database. Table 3.12 presents the results obtained from MEG-classifier, for each of the basic emotions and the final recognition rate. The results obtained using MEG-classifier model. The second column shows the accuracy achieved by training CK+ database and testing it with its own images, while the third column represents the accuracy obtained by testing SelfieDat database.

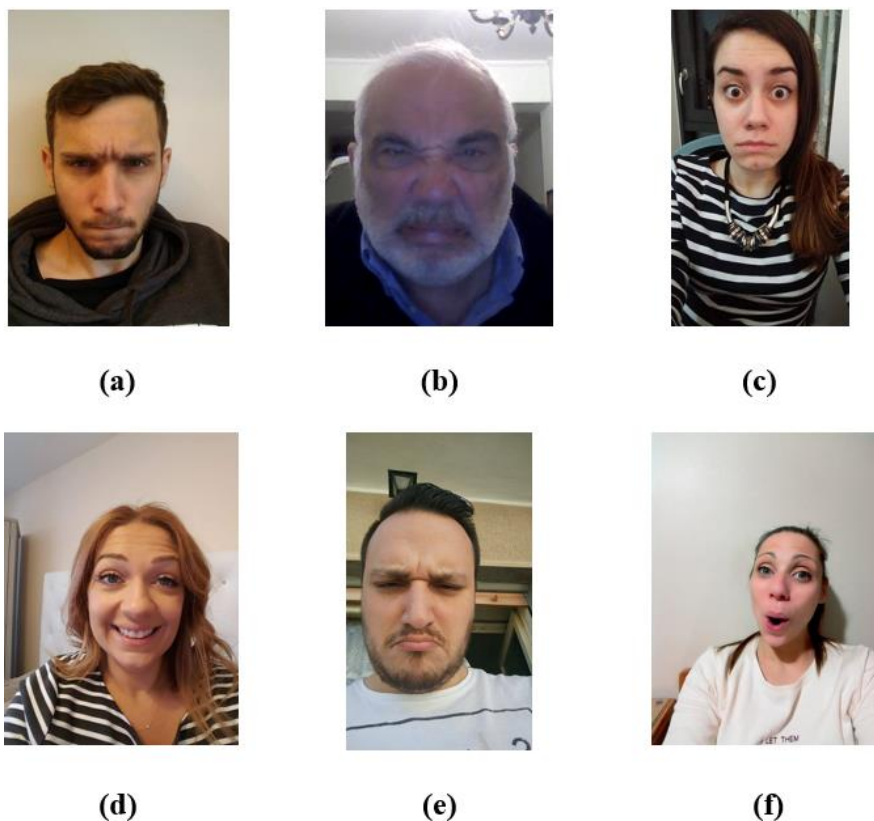


Figure 3.20: Sample of the participants in selfie-experiment. (a) Anger, (b) Disgust, (c) Joy, (d) Sadness and (e) Surprise.

Emotions	Intra-training (%)	Train: CK+ Test: SelfieDat (%)
<i>Anger</i>	82.7	100
<i>Disgust</i>	91.5	66.7
<i>Fear</i>	81.8	75
<i>Joy</i>	97.5	91.6
<i>Sadness</i>	93.8	75
<i>Surprise</i>	94.2	100
Average	90.3	84.7

Table 3.12: Average recognition accuracy for each emotion training CK+ and testing SelfieDat.

The above results demonstrate the performance of ERS, in a challenging real life scenario. For three of the emotions (anger, joy and surprise) the proposed framework had a particularly high performance. This can be explained by the fact that the expressions formed during these emotions are identical in every human. On the other side, for the rest of the emotions (disgust, fear and sadness), the expressions vary. For example, some people show their disgust by opening their mouth and exposing their tongue while other closing their mouth and frowning (see Fig. 3.21). Most of the individual in CK+ database they closed their mouth during the disgust emotion. Therefore, the information collecting from this area can lead to a misclassification such as the surprise emotion. A good solution to this issue, which would increase the accuracy of the system, is a massive and universal database.



Figure 3.21: The disgust emotion (samples from SelfieDat). In the left image, the expression is more intense (i.e. mouth is open and frowning) while in the right image the mouth is closed.

M-classifier and EG-classifier models' accuracy was 58.3% and 51.4% respectively. In particular, the mouth region's classifier had high accuracy rate (83.3%) in surprise and joy emotions while the fear emotion was extremely low (16.7%). The eyes and glabella regions' classifier had moderate performance quite similar in all of the six emotion. These results show the interdependence between these two models. In

Table 3.13, the results of M-classifier and EG-classifier are presented. The second column shows the accuracy for M-classifier, while the third column presents the accuracy for EG-classifier.

Emotions	M-classifier (%)	EG-classifier (%)
<i>Anger</i>	75	66.7
<i>Disgust</i>	50	33.3
<i>Fear</i>	16.7	66.7
<i>Joy</i>	83.3	33.3
<i>Sadness</i>	41.7	33.3
<i>Surprise</i>	83.3	75
Average	58.3	51.4

Table 3.13: Average recognition accuracy for each emotion for M and EG classifiers.

3.4.6 A comparison with state-of-art

In this section, the results achieved using the proposed method, are compared with the most recent State-of-the-Art FER algorithms [9, 10, 12, 13, 177], using the same database (i.e. Cohn-Kanade database). In emotion analysis field, several databases have been used in the past for training and testing purposes. In order to evaluate the ERS performance, CK+ database was preferred in this research. CK+ has three significant elements: reasonable size, sequences of images begin with a neutral expression and proceeds to a peak expression and is referred in the majority of similar works in literature. However, the selection of the subsets (i.e. images per emotion) and image position varied. For example, Carcagni et al. [9] preferred the fourth image of each sequence while other approaches the last. To be as fair as possible, in this project the database was separated into approximately equal subsets and only the last image of each sequence was used. For comparison purposes, the results obtained from the MEG-classifier model (LDA classification) were used.

Table 3.14 reports a comparison of other related works with the best results achieved using the ERS. It is important to mention that the overall recognition rate of ERS was competitive and particularly high compared with other significant recent research, such as [9, 10]. In particular, ERS's performance for emotions such as disgust, joy, sadness and surprise was similar or better than its competitors. However, according to Table 3.14, the proposed system slightly suffers to recognize anger and fear. That occurred because during these emotions the facial characteristics' deformation is similar to each other. However, except for this limitation, the proposed framework is comparable with other state-of-the-art methods in terms of expression recognition accuracy.

Emotion	Carcagni et al. [9] (%)	Happy et al. [10] (%)	Zhong et al. [177] (%)	Poursaberi et al. [12] (%)	Yang et al. [13] (%)	ERS (proposed) (%)
<i>Anger</i>	88.6	87.8	71.4	87.1	81.18	82.7
<i>Disgust</i>	89.0	93.3	95.3	91.6	55.29	91.5
<i>Fear</i>	100	94.3	81.1	91.0	81.39	81.8
<i>Joy</i>	100	94.2	95.4	96.9	94.90	97.5
<i>Sadness</i>	100	96.4	88.0	84.6	74.36	93.8
<i>Surprise</i>	97.4	98.5	98.3	91.2	91.08	94.2
<i>Average</i>	95.8	94.1	88.3	90.4	80.00	90.3

Table 3.14: Comparison with the state-of-art methods for six basic emotions. Bold indicates the best results.

Recent years have witnessed significant progress in FER algorithms. In [9] which presents the best overall accuracy, the HOG descriptor was used for feature extraction. In this work instead of facial patches, the entire face was utilized. Data were extracted from areas that remain static during the emotion's activation, similarly with ERS's operation. However, a significant drawback in their algorithm is the amount of redundant information that extracted which leads to a large feature vector. In section 3.4.3, is noticed that glabella region's performance was much lower than eyes and mouth. Therefore, it is assumed that glabella area as well as other areas in the human face could be removed from the final feature vector in order to improve the system's overall performance. That would be beneficial in terms of accuracy and computational time. Moreover, impressive results were presented in [10]. In their system, the LBP descriptor was applied in feature selection process. Data extracted from 19 facial patches around the eyes, nose and mouth regions. Whereas the performance of their work was slightly better than in ERS, their algorithm involves four facial region detections and data extraction from 19 different facial patches. On the other hand, ERS requires only two facial characteristics' detections and feature extraction from four facial patches.

Based on the above observations and the novel framework of ERS that allows recognizing an emotion with less than two facial regions the proposed method constitutes one of the most robust solutions in FER analysis.

3.5 ERS using LBP descriptor

In previous Section 3.4, information about the shape of the facial patches was extracted by using HOG descriptor and the results were promising. In this section, in order to broad the scope of this research, another appearance-based method called Local Binary Patterns is utilized. By following the same process as in Section 3.4, LBP descriptor's capability to acquire information of the skin formation during the emotional arousal was studied. Each facial region was examined separately. Based on the results, the best

combination of the parameters was used to design the final system. The latter achieved, by concatenating the most accurate feature vectors.

3.5.1 LBP Descriptor's parameters assessment

LBP is a visual descriptor used in various works in the past for classification in computer vision field. It was first introduced by Ojala et al. [123] in 1994 and since then became one of the most powerful tools in texture analysis. Thus, it was chosen one of the approaches, applied on ERS. Steps for the LBP feature extraction are as follows:

- 1) The image is first divided into a specific number of cells.
- 2) Within each cell, each pixel has to be compared with a specific number of neighbors. When the center pixel's value is greater than its neighbors, it becomes '0'. Otherwise, it becomes '1'. Therefore, in the case there are eight neighbors, an 8-digit binary number occurs.
- 3) The histogram of each cell is computed by taking into account the number of occurrences for each value. In its simplest structure, a histogram contains 256 values. The histograms are normalized and the concatenation of them provides the feature vector of the entire image.
- 4) In this research, the extension of the original operator, called uniform pattern [178] has been utilized. In order to reduce the size of the feature vector, a simple rotation-invariant descriptor is used. This method was occurred by the fact that several binary patterns appear more frequently in texture analysis than others. A local binary pattern is uniform if it contains at most two bitwise transitions from '0' to '1' or from '1' to '0' during a circular cross. For example, 00001000 is a uniform pattern because it has only two transitions, while 01001000 is not since it contains four transitions. The LBP histogram bins contain one position for every uniform pattern, and the non-uniform patterns are assigned to a single bin. Using this technique is achieved a size reduction from 256 to only 59 bins.

As in HOG descriptor's analysis, the purpose of this experiment is to find the best LBP parameters' combination for each of the salient regions. LBP descriptor has three important parameters: the cell size, the number of neighbors and the radius of circular pattern. In practice, the first two parameters had significant influence in systems performance while the last did not alternate the results. Therefore, for this experiment, the accuracy from different combinations of cell size and number of neighbors are presented. The numbers of feature produced from each combination, for each of the examined facial characteristics, are shown in Table 3.15-3.17. As it is expected, the number of features increases as the cell size decreases and the number of neighbors increases. For example, for the mouth region (see Table 3.15) that has a size of 64x96 LBP descriptor produces only 354 feature if cell size is equal to 32 and number of neighbors equal to 8. On the other hand, by decreasing the cell size to half (i.e. 16x16),

four times more features will be extracted. The latter can reduce the system's processing time. Therefore, a reasonable amount of data in conjunction with an acceptable performance has to be determined.

	Number of Neighbors	
Cell Size	8	16
4x4	22,656	93,312
8x8	5,664	23,328
16x16	1,416	5,832
32x32	354	1,458
64x96	59	243

Table 3.15: Number of feature for the mouth area based on cell size and number of neighbors combinations.

	Number of Neighbors	
Cell Size	8	16
4x4	30,208	124,416
8x8	7,552	31,104
16x16	1,888	7,776
32x32	472	1,944
64x64	118	486

Table 3.16: Number of feature for the eyes area based on cell size and number of neighbors combinations.

	Number of Neighbors	
Cell Size	8	16
4x4	8,260	34,020
8x8	2,065	8,505
14x10	944	3,888
28x20	236	972
56x40	59	243

Table 3.17: Number of feature for the glabella area based on cell size and number of neighbors combinations.

The feature extraction process using the LBP descriptor is described by Fig. 3.22. In this illustration, the cell size and number of neighbors were 16x16 and 8 respectively.

Depending on the maximum deformation caused during an expression performance, the second row of Fig. 3.22 presents the histograms occurred from the concatenation of the feature from the cells. Unlike HOG descriptor, LBP's function partitions the input image into non-overlapping cells. Therefore, the number of cells and neighbors are the only factors involved in descriptor's capability to extract information. By increasing the cells size, local detail is lost, while by increasing the number of neighbors more detail is captured.

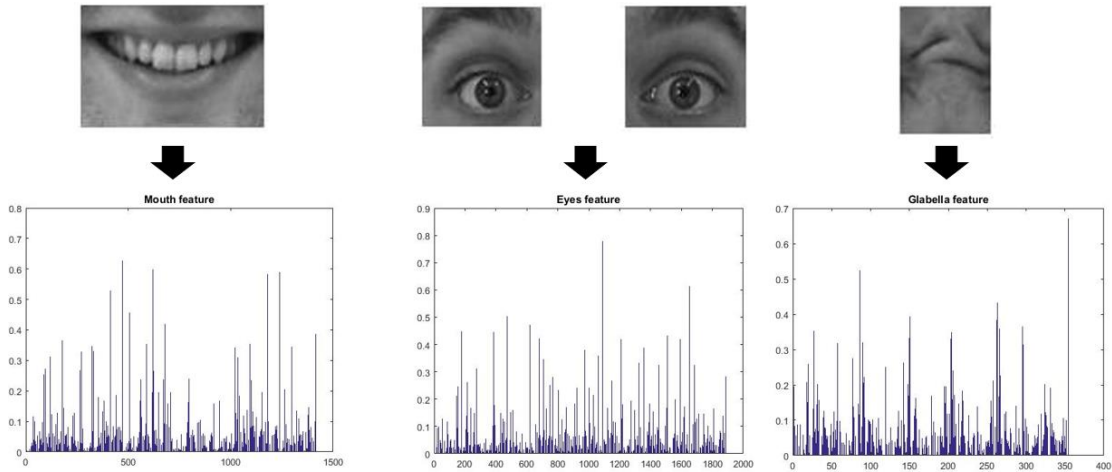


Figure 3.22: Feature extraction using LBP descriptor (cell size: 16, number of neighbors: 8) in two steps: Input image and concatenated histogram of cells. (a) Joy emotion, (b) Surprise emotion and (c) Disgust emotion.

Fig. 3.23 presents a schematic example of the feature extraction process using LBPs, for the left eye region. The image is broke down into 16 cells and each cell is analyzed independently. For each pixel in every cell, eight neighbors were thresholded. As can be seen in the figure each cell produces a separate histogram of 59 components. The number of features is calculated according to the following formula,

$$NumFeature = c \times b \quad (3.38)$$

where c is the number of cells and b is the number of histogram's bins. The latter is calculated by,

$$b = (P \times P - 1) + 3 \quad (3.39)$$

where P is the number of neighbors. This formula is applied only to the extended version of LBP where uniform pattern property is used. According to the above formula for a 64×64 window of the eye's region, there are 944 features.

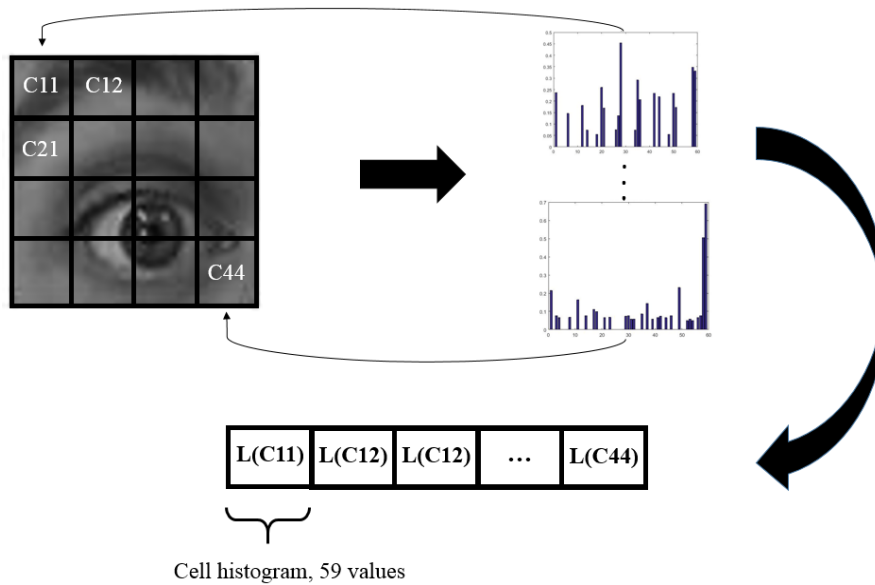


Figure 3.23: A schematic example of LBP extraction in the right eye.

In order to evaluate the system's performance experiments using CK+ and ADFES databases were performed. In the following subsections, the results obtained from LDA and SVM classifiers using all the possible combinations of the LBP parameters are evaluated, in order to determine the most efficient model. The classification parameters are identical with the experiments performed for HOG descriptor (see Section 3.4.2).

3.5.2 LBP training and testing using the CK+ database

By following the same process as in Section 3.4.3, CK+ database was separated into approximately equal size subsets for each basic emotions. Therefore, 330 images were utilized for training purposes (see Table 3.7).

In order to result in the most efficient parameters' combination for this system, experiments were performed for mouth, eyes and glabella regions' separately. In Table 3.18, the best accuracy for each of the above regions using LDA and SVM classification are presented. The results of both classification methods were similar. However, the combinations were demonstrated the highest accuracy in SVM concentrate a large amount of data. Therefore, in order to reduce the processing time, LDA classification is considered as the most suitable. Algorithm's performance is affected by two major factors: the combination of the feature parameters and the PCA's explained variance. Fig. 3.24-3.29 present the overall behavior and the best model's confusion matrix for

each facial region. From the results illustrated in these figures and Table 3.18, the following observations can be stated:

- 1) Mouth region has a great impact on system's performance. Due to the numerous muscles and its unique contribution to every expression, mouth constitutes the main source of information. In Table 3.18, mouth model achieved by itself an accuracy of 83.4%. Moreover, in Fig. 3.24 where the overall accuracy is presented is clear that the PCA variance levels lines follow a common course on the "x-axis" which describes the combinations of the cell size and the number of neighbors. The most accurate and size acceptable model occurred when the PCA variance is 85%, the cell size is 16x16 and the neighbors are 8. In Fig. 3.25, the mouth model's confusion matrix shows that joy, surprise, fear and disgust accuracies are particularly high over 80% while sadness and disgust are close to 75%. It was expected that the most misclassifications occurred between anger and sadness as well as between fear and joy. The reason for this issue is that the above pairs of emotions share common characteristics in the mouth region.
- 2) Eyes' model has an acceptable performance rate of 68.3%. Moreover, it is noticed that is increased in comparison to the analogous model in Section 3.4.3. Fig. 3.26 shows that the PCA variance level of 85% dominates over the other and the most efficient combination of cell size and neighbors are 16 and 8 respectively. As it can be seen in Fig. 3.27, this model presents a particularly high accuracy especially in joy, surprise and disgust emotions while fear and sadness emotions are quite low. The most misclassifications occurred between fear and sadness, fear and surprise and sadness and anger emotions. Unlike with HOG, LBP descriptor was demonstrated better accuracy in this particular region.
- 3) Glabella region had extremely low contribution accuracy of 56.3%. In Fig. 3.28, most of PCA variation levels follow the same performance in every combination of the LBP parameters. However, the model that describes reasonably the Glabella area has cell size equal to 14x10, 16 neighbors and PCA variance at 55%. According to the Fig. 3.2, in most of the emotions with only exception the emotion of disgust the accuracy is low. Essentially, the frowning expression of an individual during feelings such as disgust and anger provides useful information. However, this is not applicable in the rest of the emotions, for which this region does not present any wrinkles.

	SVM (%)	C	N	LDA (%)	C	N
Mouth	84.1	4	8	83.4	16	8
Eyes	65.2	8	8	68.3	16	8
Glabella	56.0	4	16	56.3	14x10	16

Table 3.18: Best accuracy of each region for LDA and SVM. C and N represent the cell size and number of neighbors respectively.

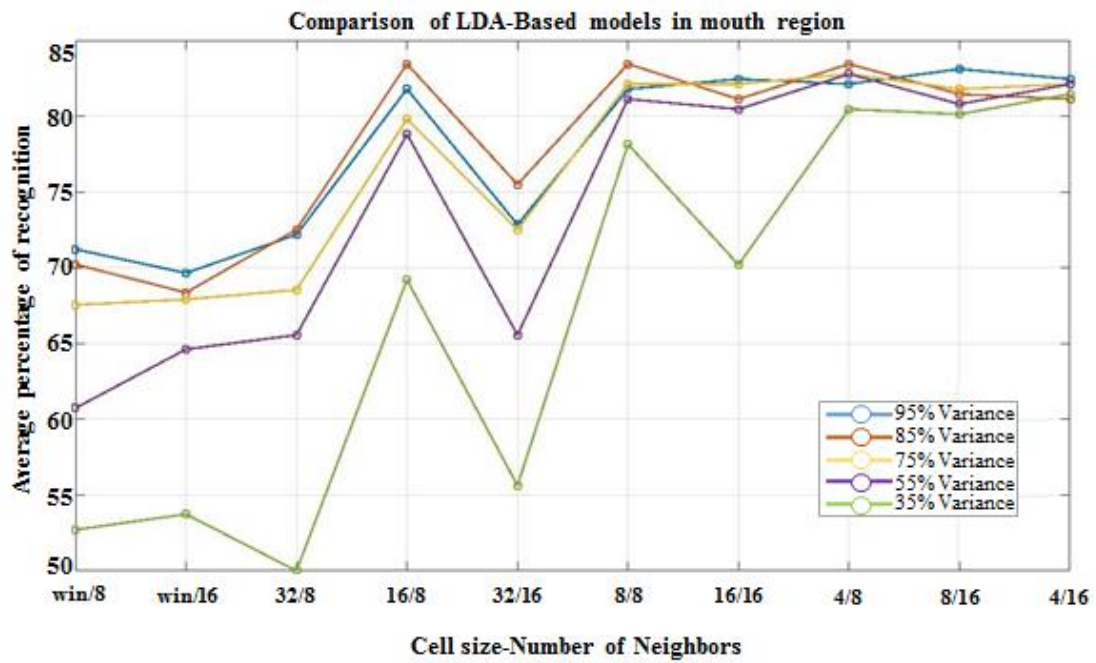


Figure 3.24: Mouth region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max).

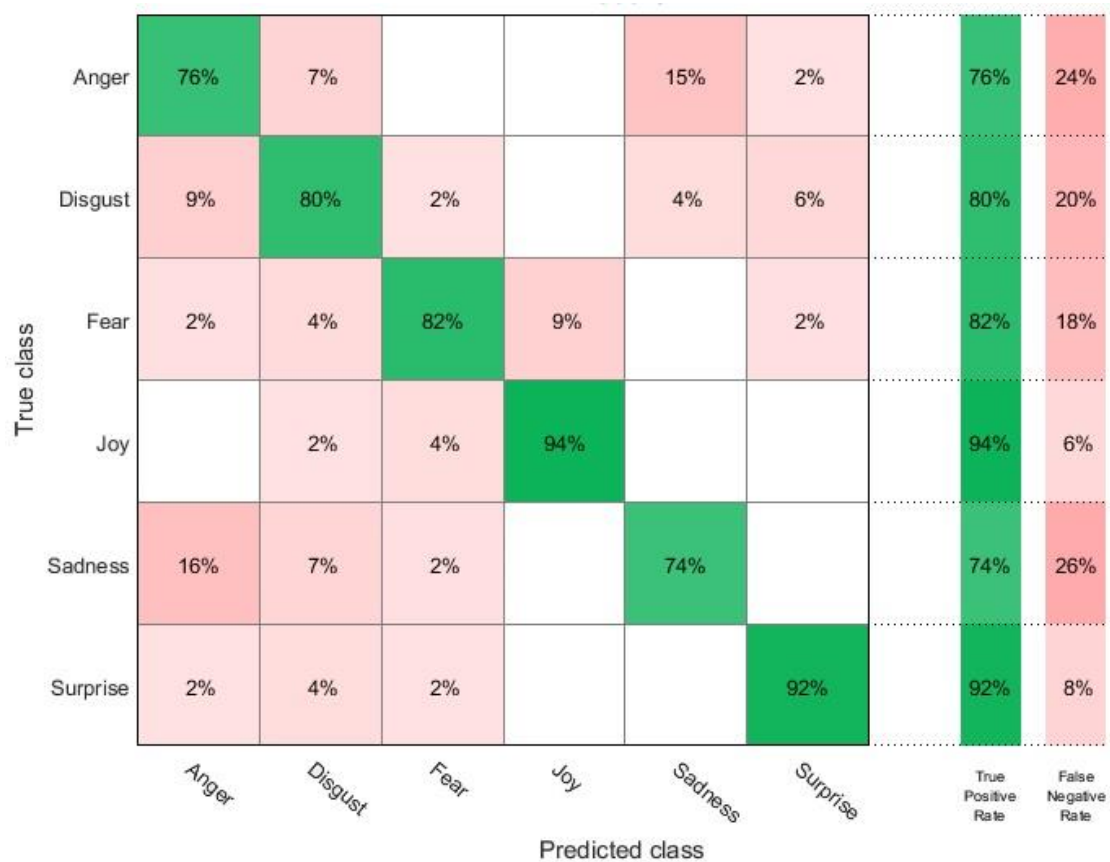


Figure 3.25: Mouth region’s confusion matrix.

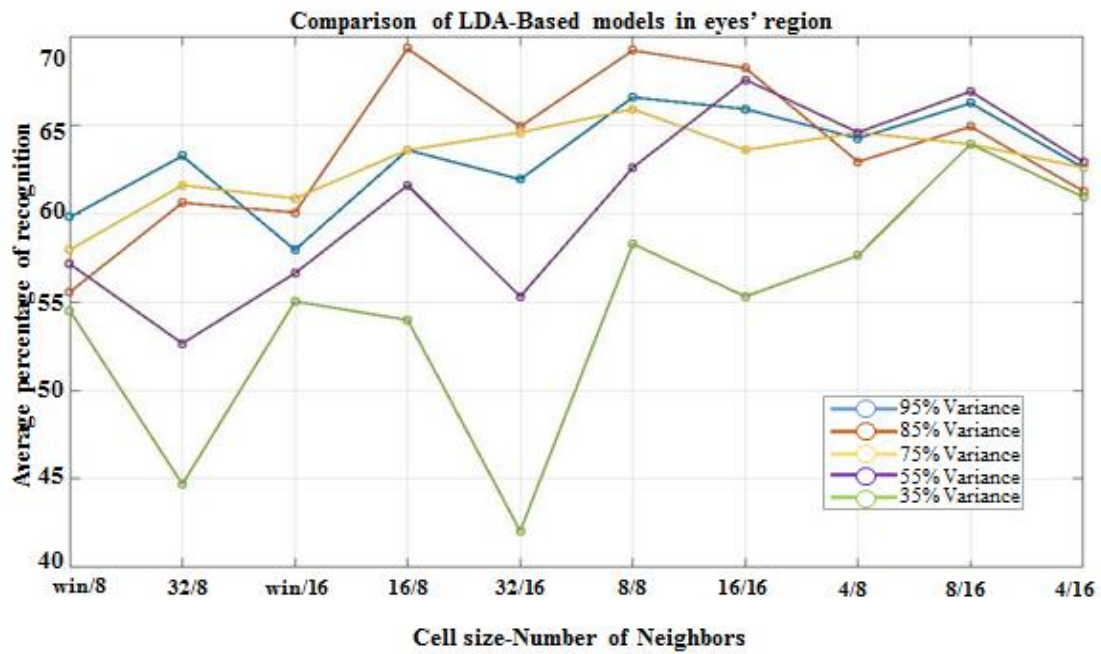


Figure 3.26: Eyes region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max).

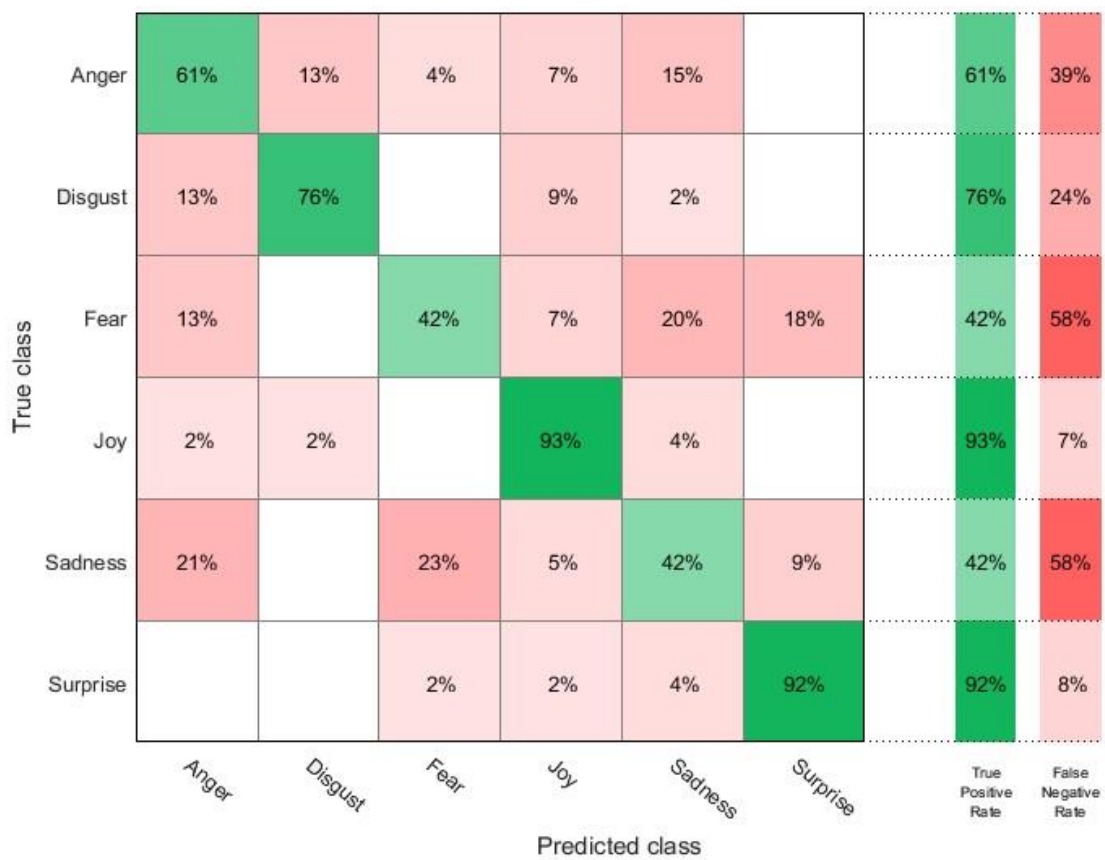


Figure 3.27: Eyes regions’ confusion matrix.

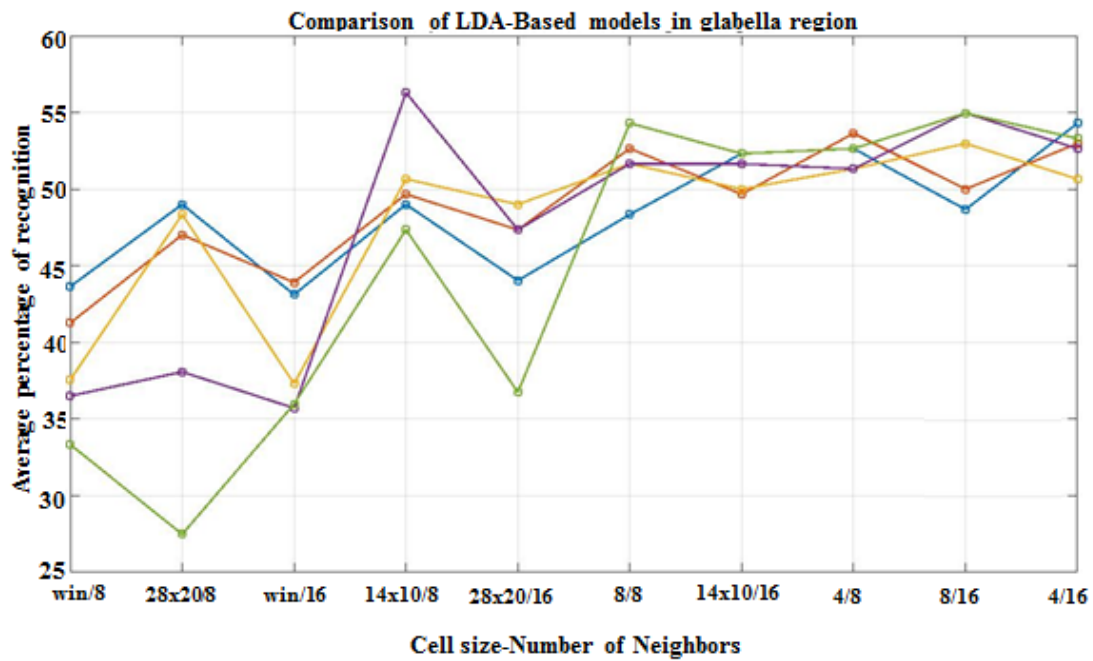


Figure 3.28: Eyes region. Average recognition rates evaluation for all the possible combinations of LBP descriptor’s parameters (win is referred to the entire image) with five PCA variance levels. Parameters are placed based on feature vector’s size (min – max).

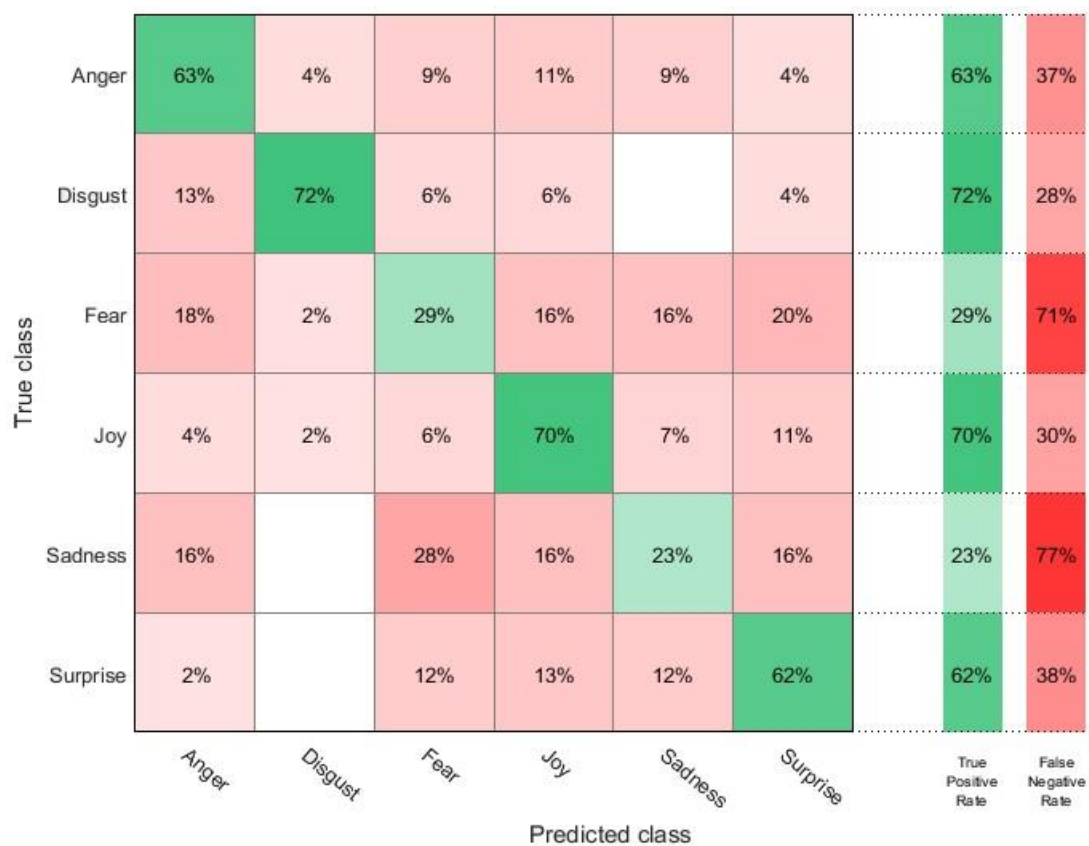


Figure 3.29: Glabella’s region’s confusion matrix.

In order to design an effective and robust algorithm, which confronts the partial occlusion problem, three classification models are proposed. In the same manner as in Section 3.4.3, depending on the availability of the detected regions, ERS activates one of the following models: MEG-classifier (all regions are detected), M-classifier (only the mouth region is detected) and EG-classifier (only eyes and glabella regions are detected). Based on the above results, the parameters provided the best accuracy were selected to extract the most informative feature from each region. The data were concatenated together, to form a new feature vector. The MEG-classifier model determined after several tests were performed on the concatenated feature vector with SVM and LDA classification methods. The same process was followed with eyes and glabella regions to determine the EG-classifier. In Table 3.19, the most effective models are presented. In the first line, the LDA classifier of MEG demonstrates the superiority of this model over the M or EG. The accuracy is slightly increased in comparison to the corresponding model in HOG case. M-classifier was described in the previous section. However, by observing the overall results it is clear that mouth offer most of the information regarding the emotion. Due to the low performance in glabella the region, EG-classifier does not increase the recognition rate.

Model	SVM (%)	LDA (%)
MEG-classifier	89.7	90.6
M-classifier	84.1	83.4
EG-classifier	67.2	68.5

Table 3.19: Best accuracy of each model for LDA and SVM.

Fig. 3.30-3.31 present the confusion matrices generated from the EG-classifier and MEG-classifier models. The latter has certainly increased the recognition rate of the algorithm. In particular, emotions such as joy, surprise and disgust had remarkable accuracy, 100%, 96% and 94% respectively. However, the concatenation of the feature from eyes and glabella regions did not sufficiently improve the algorithm's recognition rate. Especially for the fear and sadness emotions, the accuracy was quite low 40% for each of them. Many misclassifications occurred between these two emotions since the shape of the eyes is similar.

Therefore, the MEG-classifier using LBP descriptor constitutes a robust and efficient solution in emotion analysis. On the other side, the EG-classifier had moderate performance. A comparison with the HOG descriptor's results (see Section 3.4.3) demonstrates that the recognition rate is slightly better. Both techniques had exceptional outcomes, but in general, LBP was superior to HOG in all available models.

The same process was adopted using the ADFES database. The results were related to the CK+ with 93.7%, 88.9% and 72.7% recognition rates in MEG-classifier, M-classifier and EG-classifier respectively. In the next section, experiments will be performed on ADFES and SelfieDat databases in order to determine if the proposed algorithm can achieve high rates in real life's scenario.

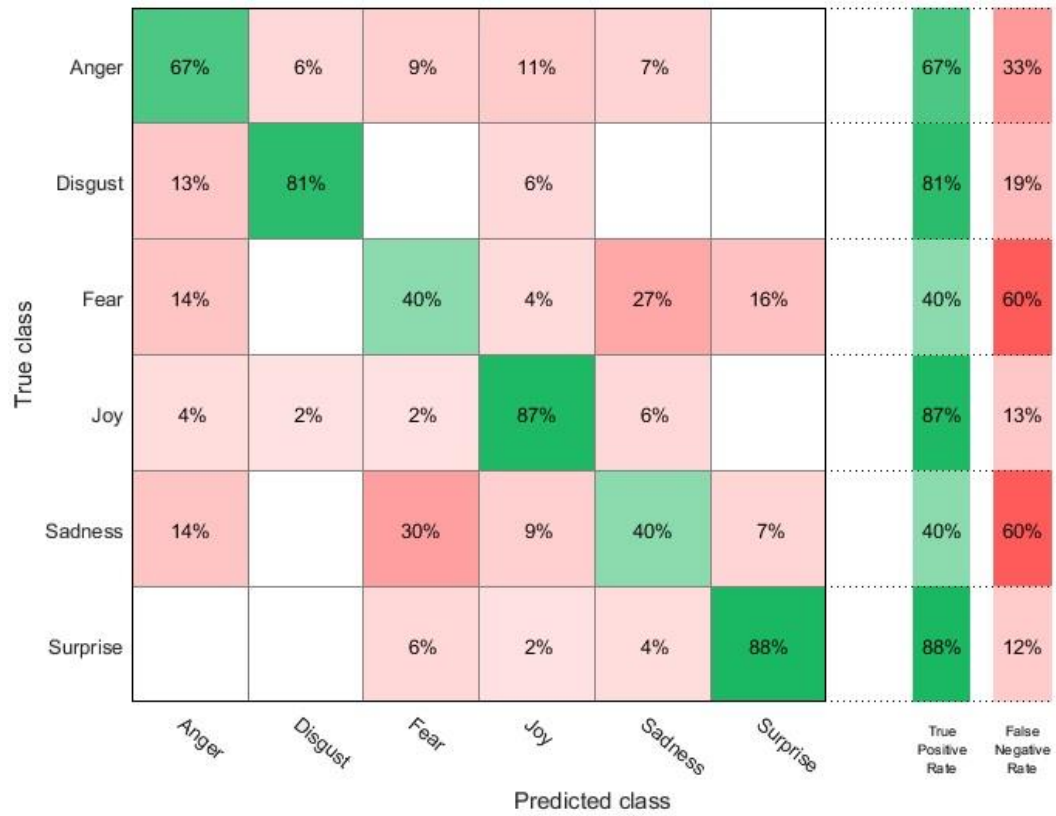


Figure 3.30: EG-classifier confusion matrix.

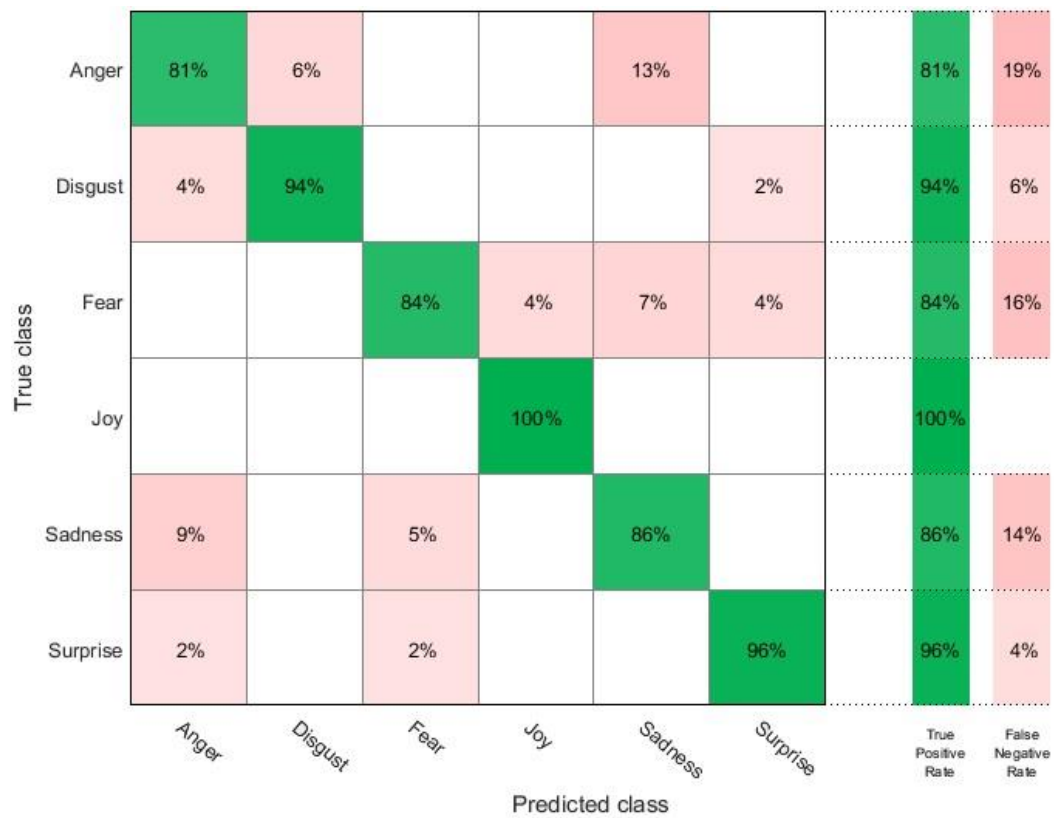


Figure 3.31: MEG-classifier confusion matrix.

3.5.3 ADFES and SelfieDat

In this section, two experiments are introduced. Initially, the proposed algorithm is tested using the ADFES database. Subsequently, the proposed models are tested using the SelfieDat. There are several reasons two different databases were utilized for testing purposes:

- 1) The equipment and candidates guidance for the ADFES database was professional while in SelfieDat, various cell phones were used and no guidance was offered.
- 2) Whereas the environmental conditions (light, weather, background) in ADFES were identical, in SelfieDat were random.
- 3) The images' size in ADFES was fixed while in SelfieDat was variable.

Essentially, the ADFES is similar to CK+ since both were created in laboratory conditions. On the other hand, testing the proposed models using SelfieDat images constitutes a challenge that reveals the advantages and disadvantages of the algorithm when is running under random circumstances. More information regarding the two databases is given in Sections 3.4.4 and 3.4.5.

Tables 3.20-3.21, demonstrate the results obtained by testing ADFES and trained CK+ and vice versa. The results obtained using the MEG-classifier model. Moreover, the second column shows the accuracy achieved by the training database and testing it with its own images, while the third column represents the accuracy obtained by the new database. In the first scenario where the algorithm trained with CK+ database and tested with ADFES, the overall performance was acceptable with 78% accuracy. Only fear emotion presented low accuracy (i.e. 63.6%). In the second scenario, the two databases switched roles. The overall performance was slightly lower at 73.2%. The performance affected significantly by the recognition rate of fear emotion (i.e. 37.8%). The comparison with the corresponding experiment for the HOG descriptor (76.6% and 83.6) defines the similarity of these two methods again. The only difference between them appears in the second scenario where the fear emotion had considerably lower accuracy.

Emotions	Intra-training (%)	Train: CK+ Test: ADFES (%)
<i>Anger</i>	81.5	100
<i>Disgust</i>	94.2	72.7
<i>Fear</i>	84.7	63.6
<i>Joy</i>	100	72.7
<i>Sadness</i>	86.7	81.8
<i>Surprise</i>	96.6	77.3
Average	90.6	78.0

Table 3.20: Average recognition accuracy for each emotion training CK+ and testing ADFES.

Emotions	Intra-training (%)	Train: ADFES Test: CK+ (%)
<i>Anger</i>	92.6	66.7
<i>Disgust</i>	96.4	81.5
<i>Fear</i>	85	37.8
<i>Joy</i>	100	100
<i>Sadness</i>	88.4	76.7
<i>Surprise</i>	100	76.9
Average	93.7	73.2

Table 3.21: Average recognition accuracy for each emotion training ADFES and testing CK+.

In Table 3.22-3.23, the results obtained from the three proposed models are presented. The most encouraging aspect in MEG-classifier's results is that the emotion of anger and surprise were demonstrated high accuracy over 75%. However, the recognition rates for the rest of the emotions were quite low around 50%. The performance of the other two models (M and EG) was poor with 55.6% and 48.6% respectively. As explained earlier the SelfieDat is referred to a real life's scenario. Therefore, when is tested with models were trained with a laboratory database's images is not expected to demonstrate exceptional results. A good solution to this issue is to create a robust and massive database with real life images.

At this point, it is worthy to mention that experiments show HOG models provide consistently better performance than LBP for a real life's scenario. The flexibility of HOG descriptor regardless of the image's illumination and resolution are greater than LBP.

Emotions	Intra-training (%)	Train: CK+ Test: SelfieDat (%)
<i>Anger</i>	81.5	100
<i>Disgust</i>	94.2	41.7
<i>Fear</i>	84.7	50
<i>Joy</i>	100	58.3
<i>Sadness</i>	86.7	41.7
<i>Surprise</i>	96.6	75
Average	90.6	61.1

Table 3.22: Average recognition accuracy for each emotion. The results obtained using MEG-classifier model.

Emotions	M-classifier (%)	EG-classifier (%)
<i>Anger</i>	91.7	83.3
<i>Disgust</i>	41.7	33.3
<i>Fear</i>	25	41.7
<i>Joy</i>	66.7	33.3
<i>Sadness</i>	41.7	25
<i>Surprise</i>	66.7	75
Average	55.6	48.6

Table 3.23: Average recognition accuracy for each emotion. The results obtained using M-classifier and EG-classifier models. The second column shows the accuracy for M-classifier, while the third column presents the accuracy for EG-classifier.

3.5.4 Comparison of the performance

In this section, training and testing was carried out using images from CK+ database. CK+ includes sequences of images for each emotion (from neutral to peak) and as in section 3.4 the last image of each sequence was used for training and testing purposes. The proposed framework is compared with other state-of-art frameworks using the same database.

In Table 3.24 is noticed, that ERS achieved exceptional results comparable to other FER research. In particular, one of the most recent works introduced by Happy et al. [10] where the same method (i.e. LBP) was adopted. The performance of ERS was better in two emotions (disgust and Joy); three of the emotions were slightly lower and only one emotion (fear) is considerable lower by 10%. The overall accuracy of [10] is higher than in the proposed algorithm. However, ERS exploits less areas on the human face and consequently reduces the size of the feature vector (in all the possible combinations of the parameters). In ERS, only two facial characteristics are detected and data is extracted from three facial patches whereas in [10] four facial characteristics have to be detected and data is extracted from 19 different patches. Therefore, in terms of computational complexity, ERS offers a more efficient solution. In the study of Zhong et al., initially the most effective facial patches during the emotions' arousal are discovered and then based on these findings the LBP feature are extracted. The overall accuracy of their system was 88.3%, slightly lower than ERS's. Exceptional results, by utilizing the HOG descriptor were achieved by [9] but as explained in Section 3.4.6 the feature vector produced from their algorithm is large since data are extracted from the whole face. Many similar studies of emotion recognition (such as [12, 13]) through facial expression analysis have shown remarkable results in one or more emotions prediction.

From the above comparison and the unique capability of ERS to perform emotion recognition with partial occlusion the proposed method constitutes one of the most robust solutions in FER analysis.

Emotion	Carcagni et al. [9] (%)	Happy et al. [10] (%)	Zhong et al. [177] (%)	Poursaberi et al. [12] (%)	Yang et al. [13] (%)	ERS (proposed) (%)
<i>Anger</i>	88.6	87.8	71.4	87.1	81.18	81.5
<i>Disgust</i>	89.0	93.3	95.3	91.6	55.29	94.2
<i>Fear</i>	100	94.3	81.1	91.0	81.39	84.7
<i>Joy</i>	100	94.2	95.4	96.9	94.90	100
<i>Sadness</i>	100	96.4	88.0	84.6	74.36	86.7
<i>Surprise</i>	97.4	98.5	98.3	91.2	91.08	96.6
<i>Average</i>	95.8	94.1	88.3	90.4	80.00	90.6

Table 3.24: Comparison with the state-of-art methods for six basic emotions. Bold indicates the best results.

3.6 Summary

In this chapter, an automatic system that combines face and facial feature detection, facial texture analysis and assessment for the six basic emotions are described. Two robust face detection algorithms [8, 103] were combined, achieving 100% accuracy in all of the three face databases. Moreover, an effective computational algorithm that exploits the eyes' location to perform face alignment is presented. Two well-known texture analysis approaches (HOG and LBP) were investigated to extract the feature of three facial patches. Finally, two classification techniques (SVM and LDA) were examined in order to recognize six basic human emotions. Two emotion recognition systems were proposed based on these techniques while three classifiers were proposed for each system.

The overall system is designed to confront the significant issue of partial occlusion. Despite its simplicity, the proposed algorithm performs well and can be characterized particularly competitive in regards to recent research.

Chapter 4

Conclusion and Future Work

4.1 Conclusions

This thesis has presented novel image processing algorithms for emotion recognition in images that can include multiple faces.

An integrated approach of automatic emotion recognition assessment system was developed, including the localization of three facial feature regions, the face alignment, the extraction of information regarding the facial texture, the classification of the data and the quantitative assessment of an artificial intelligence-based solution.

The aim of this thesis was to study the facial expressions of emotion, in terms of texture analysis. Two popular approaches, histograms of oriented gradients and local binary patterns, were exploited in the feature extraction process to derive information from three facial patches. Based on this work, an emotion recognition algorithm, called Emotion Recognition System (ERS) was developed in Matlab. It provides exceptional performance as well as low computational cost, especially when all the facial regions are utilized.

To overcome issues presented in face detection, such as false detection and misdetection, a novel face detection algorithm that combines two significant approaches, has been proposed. Experimental results show that the proposed algorithm can effectively detect multiple faces within an image. In particular, based on the three databases utilized during this research, the proposed algorithm was 100% accurate.

In literature, numerous approaches have been proposed for the emotion recognition, although none of them is designed to confront the partial occlusion issue. In order to solve this common problem, a novel algorithm that includes three pre-trained classifiers has been presented. Therefore based on the detected regions the following models are enabled:

- 1) MEG-classifier: All the regions are available. The classifier was trained and tested with CK+ database, demonstrating strong recognition rate, 90.3% and 90.7% with HOG and LBP respectively.
- 2) M-classifier: Only the mouth region is considered. The classifier was trained and tested with CK+ database, demonstrating strong recognition rate, 84.7% and 83.4% with HOG and LBP respectively.
- 3) EG-classifier: Only the eyes and glabella regions are considered. The classifier was trained and tested with CK+ database, demonstrating strong recognition rate, 67.0% and 68.5% with HOG and LBP respectively.

Finally, in this thesis a new novel database with 12 images is developed for the evaluation of the proposed system. Although there are various databases available in

the literature, the SelfieDat dataset is useful for two reasons: all the images were captured by various cell phones and the conditions (i.e. light, pose) were variable. Usually, in related work only professional datasets are utilized. However, by training and testing algorithms with flawless images the results are unrealistic. In order to understand, the capability of the proposed system, experiments performed using the SelfieDat, with accuracy 84.7% and 61.1% for HOG and LBP techniques respectively.

To conclude, a number of facial expression analysis methods have presented, and have proved that they can effectively be used in facial emotion analysis.

4.2 Future Work

Although the proposed facial emotion recognition system was demonstrated to perform well, several issues remain to be investigated in order to enhance its capability. As discussed in section Section 2.2.3 six primary emotions were characterized universal [58] and combinations of them produce new emotions. In our methods, only the primary emotions are assessed by examining their expressions independently. Although, secondary emotions are not taken into account in this work, it would be useful to extend the existing system to recognize emotions such as, interest, distraction or boredom. In recent research, Yang et al. [134] considered the intensity of emotion as a crucial factor of emotion activation. The latter in conjunction with Plutchik [55] studies (see Section 2.2.3) can lead to a more realistic system, considering more aspects of human psychological structure.

The current system has shown impressive performance for the emotion recognition in images. To increase and consolidate the present work, additional research which could be pursued is reported below:

- 1) In the experiments, performed using the ADFES and SelfieDat the recognition rate of specific emotions (i.e. disgust) are not associated with the intra-trained performance of the system. Therefore, more data with spontaneous expressions are required. It would be beneficial for the system's analysis if the SelfieDat could be updated with more images being added so it can be used for training purposes too. That could bring new information about the performance of the algorithm under real life's conditions.
- 2) A natural extension for the Emotion Recognition System described in Section 4, will be to apply the current techniques in video sequences. In this case, the classifier will be trained with 3-5 sequences for each emotion and lead to a system that will operate in a real-time environment.
- 3) Further extensions and adjustments in feature extraction techniques, will increase the system's processing time. Especially, in continuous input case, the system has to handle the data very rapidly.
- 4) Improving system's performance by applying rules during the feature extraction process could be a useful addition. For example, the glabella region will be taken into account only if wrinkles are detected.

- 5) Concerning the face detection, the rotation of more than 15 degrees issue can lead to a misdetection. For example, in fear emotion people may turn their faces in act of self-protection. A solution could be to perform face recognition through more than one angles but the system would suffer in terms of processing time.

Appendix A: The SelfieDat database

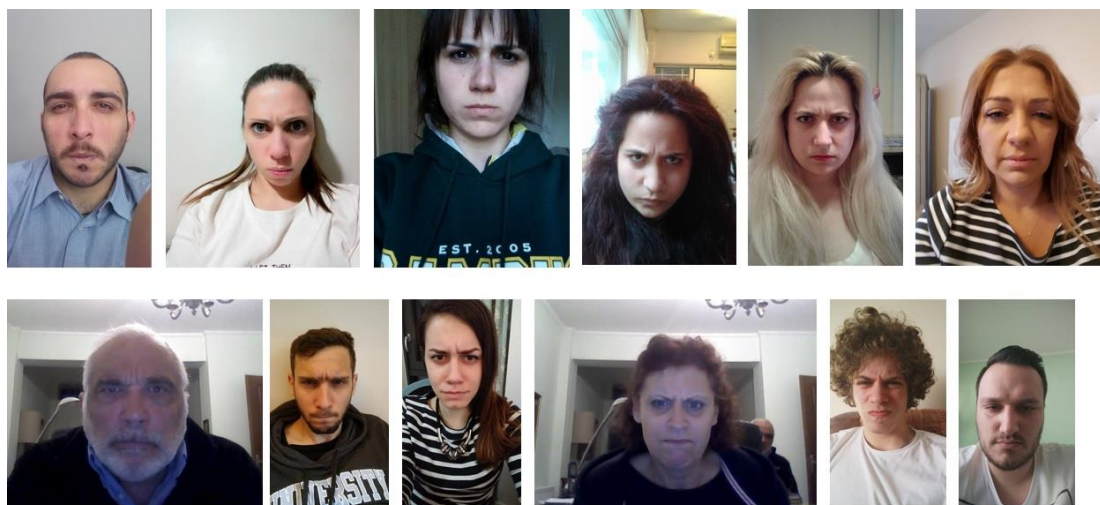


Figure A1: The anger emotion in SelfieDat.

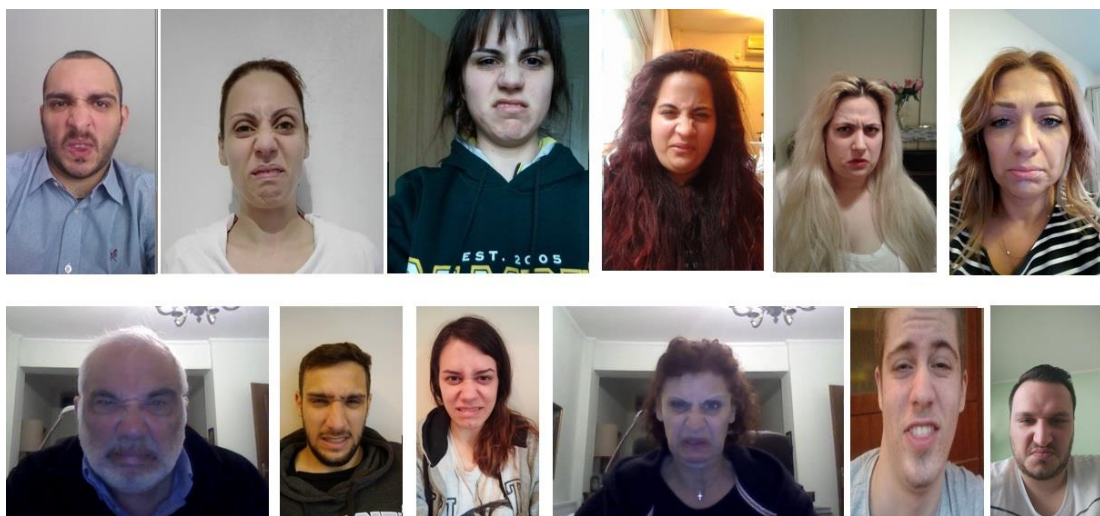


Figure A2: The disgust emotion in SelfieDat.

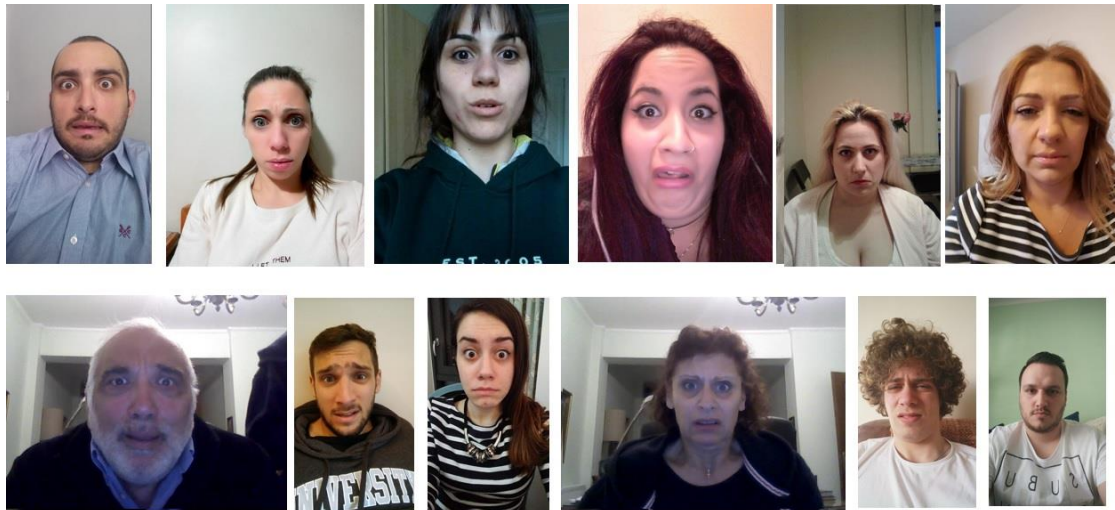


Figure A3: The fear emotion in SelfieDat.



Figure A4: The joy emotion in SelfieDat.



Figure A5: The sadness emotion in SelfieDat.



Figure A6: The surprise emotion in SelfieDat.

References

- [1] P. Ekman and R. J. Davidson, *The nature of emotion : fundamental questions*. New York: New York : Oxford University Press, 1994.
- [2] K. R. Scherer, "What are emotions? And how can they be measured?," *Social science information*, vol. 44, no. 4, pp. 695-729, 2005.
- [3] J. A. Russell, "Facial expressions of emotion: What lies beyond minimal universality?," *Psychological bulletin*, vol. 118, no. 3, p. 379, 1995.
- [4] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.
- [5] R. Socher, B. Huval, B. P. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-Recursive Deep Learning for 3D Object Classification," in *NIPS*, 2012, vol. 3, no. 7, p. 8.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
- [7] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137-154, 2004.
- [8] M. M. Ibrahim, "Video processing analysis for non-invasive fatigue detection and quantification," University of Strathclyde, 2014.
- [9] P. Carcagnì, M. Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 1, 2015.
- [10] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1-12, 2015.
- [11] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2562-2569: IEEE.

- [12] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich, "Gauss–Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, pp. 1-13, 2012.
- [13] P. Yang, Q. Liu, and D. N. Metaxas, "Exploring facial expressions with compositional features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2638-2644: IEEE.
- [14] F. De la Torre and J. F. Cohn, "Facial expression analysis," in *Visual analysis of humans*: Springer, 2011, pp. 377-409.
- [15] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775-1787, 2009.
- [16] NSF. Available: <https://www.nsf.gov>
- [17] IEEE. Available: <http://technav.ieee.org/tag/5835/emotion-recognition>
- [18] J. H. Turner, *Human emotions: A sociological theory*. Taylor & Francis, 2007.
- [19] P. R. Kleinginna Jr and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and emotion*, vol. 5, no. 4, pp. 345-379, 1981.
- [20] K. Marx, "Capital, vol. 1, tr. Ben Fowkes," ed: New York: Penguin Books, 1990.
- [21] E. Durkheim, "The elementary forms of the religious life: A study in religious sociology, trans. Joseph Ward Swain; reprint," ed: New York: The Free Press, 1965.
- [22] D. R. Heise, *Understanding events: Affect and the construction of social action*. CUP Archive, 1979.
- [23] A. R. Hochschild, "There's no place like work," *The New York Times Magazine*, vol. 20, 1997.
- [24] T. D. Kemper, *A social interactional theory of emotions*. Wiley New York, 1978.
- [25] J. H. Turner and J. E. Stets, *The sociology of emotions*. Cambridge University Press, 2005.
- [26] S. L. Gordon, "The sociology of sentiments and emotion," ed: New York: Basic Books, 1981, pp. 562-592.
- [27] G. Peterson, "Cultural theory and emotions," in *Handbook of the Sociology of Emotions*: Springer, 2006, pp. 114-134.
- [28] J. A. Houser and M. J. Lovaglia, "Status, emotion, and the development of solidarity in stratified task groups," *Advances in group processes*, vol. 19, pp. 109-137, 2002.
- [29] J. E. Stets, "Identity theory and emotions," in *Handbook of the sociology of emotions*: Springer, 2006, pp. 203-223.
- [30] E. J. Lawler, "An affect theory of social exchange1," *American Journal of Sociology*, vol. 107, no. 2, pp. 321-352, 2001.
- [31] L. D. Molm, *Coercive power in social exchange*. Cambridge University Press, 1997.
- [32] N. H. Frijda, "The psychologists' point of view," *Handbook of emotions*, vol. 2, pp. 59-74, 2000.
- [33] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, p. 145, 2003.

- [34] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169-200, 1992.
- [35] W. James, "II.—What is an emotion?," *Mind*, no. 34, pp. 188-205, 1884.
- [36] W. Canon, "Bodily Changes in Pain, Hunger, Fear, and Rage. Researches into the Function of Emotional Excitement," ed: Harper and Row, New York, 1929.
- [37] J. LeDoux, "Emotional networks and motor control: a fearful view," *Progress in brain research*, vol. 107, pp. 437-446, 1996.
- [38] M. B. Arnold, "Emotion and personality," 1960.
- [39] J. R. Averill, E. M. Opton Jr, and R. S. Lazarus, "Cross-cultural studies of psychophysiological responses during stress and emotion," *International Journal of Psychology*, vol. 4, no. 2, pp. 83-102, 1969.
- [40] E. Aronson, T. Wilson, and R. Akert, "Social psychology 7th Ed," *New Jersey: Upper Saddle River*, 2010.
- [41] K. R. Scherer, "Toward a dynamic theory of emotion," *Geneva studies in Emotion*, vol. 1, pp. 1-96, 1987.
- [42] C. E. Izard, *Human emotions*. Springer Science & Business Media, 2013.
- [43] S. S. Tomkins, "Affect, imagery, consciousness: Vol. I. The positive affects," 1962.
- [44] K. Oatley, *Best laid schemes: The psychology of the emotions*. Cambridge University Press, 1992.
- [45] M. Argyle, "Bodily Communication. 2nd," *London: Methuen*, 1988.
- [46] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
- [47] P. Ekman, "Should we call it expression or communication?," *Innovation: The European Journal of Social Science Research*, vol. 10, no. 4, pp. 333-344, 1997.
- [48] P. Ekman, "Emotions revealed: Recognizing faces and feelings to improve communication and emotional life," 2003.
- [49] J. Morton and M. H. Johnson, "CONSPEC and CONLERN: a two-process theory of infant face recognition," *Psychological review*, vol. 98, no. 2, p. 164, 1991.
- [50] C. L. Masten *et al.*, "Recognition of facial emotions among maltreated children with high rates of post-traumatic stress disorder," *Child abuse & neglect*, vol. 32, no. 1, pp. 139-153, 2008.
- [51] S. H. Yoo, D. Matsumoto, and J. A. LeRoux, "The influence of emotion recognition and emotion regulation on intercultural adjustment," *International Journal of Intercultural Relations*, vol. 30, pp. 345-363, 2006.
- [52] J. F. Cohn and P. Ekman, "Measuring facial action," *The new handbook of methods in nonverbal behavior research*, pp. 9-64, 2005.
- [53] C. E. Izard, R. R. Huebner, D. Risser, and L. Dougherty, "The young infant's ability to produce discrete emotion expressions," *Developmental psychology*, vol. 16, no. 2, p. 132, 1980.
- [54] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system (FACS)," *A technique for the measurement of facial action. Consulting, Palo Alto*, vol. 22, 1978.
- [55] R. Plutchik, "Emotion: A psychoevolutionary synthesis Harper & Row New York," 1980.
- [56] "Plutchik's emotion wheel ", ed.
- [57] "Plutchik's emotion cone," ed.

- [58] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.
- [59] G. Mandler, *Mind and body: Psychology of emotion and stress*. WW Norton, 1984.
- [60] J. T. Cacioppo and G. G. Berntson, "Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates," *Psychological bulletin*, vol. 115, no. 3, p. 401, 1994.
- [61] A. Mehrabian and J. A. Russell, "A measure of arousal seeking tendency," *Environment and Behavior*, vol. 5, no. 3, p. 315, 1973.
- [62] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.
- [63] J. Van Der Schalk, S. T. Hawk, A. H. Fischer, and B. Doosje, "Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES)," *Emotion*, vol. 11, no. 4, p. 907, 2011.
- [64] J. F. Cohn, Z. Ambadar, and P. Ekman, "Observer-based measurement of facial expression with the Facial Action Coding System," *The handbook of emotion elicitation and assessment*, pp. 203-221, 2007.
- [65] K. R. Scherer and P. Ekman, *Handbook of methods in nonverbal behavior research*. Cambridge University Press Cambridge, 1982.
- [66] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010, pp. 94-101: IEEE.
- [67] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (JAFFE) database," ed, 1998.
- [68] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151-160, 2013.
- [69] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.
- [70] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*, 2005, p. 5 pp.: IEEE.
- [71] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 149-161, 2008.
- [72] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.
- [73] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32-41, 2012.
- [74] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 1, pp. 34-58, 2002.

- [75] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern recognition*, vol. 27, no. 1, pp. 53-63, 1994.
- [76] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, vol. 4, pp. 2537-2540: IEEE.
- [77] H. Mekami and S. Benabderrahmane, "Towards a new approach for real time face detection and normalization," in *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, 2010, pp. 455-459: IEEE.
- [78] S. A. Sirohey, "Human face segmentation and identification," 1998.
- [79] C.-C. Han, H.-Y. M. Liao, G.-J. Yu, and L.-H. Chen, "Fast face detection via morphology-based pre-processing," in *International Conference on Image Analysis and Processing, 1997*, pp. 469-476: Springer.
- [80] Y. Huang, X. Ao, and Y. Li, "Real time face detection based on skin tone detector," *International Journal of Computer Science and Network Security*, vol. 9, no. 7, pp. 71-77, 2009.
- [81] J. Zhang, Q. Zhang, and J. Hu, "RGB color centroids segmentation (CCS) for face detection," *International Journal on Graphics, Vision and Image Processing*, vol. 9, no. II, pp. 1-9, 2009.
- [82] Y. Dai and Y. Nakano, "Face-texture model based on SGLD and its application in face detection in a color scene," *Pattern recognition*, vol. 29, no. 6, pp. 1007-1017, 1996.
- [83] T. Sakai, M. Nagao, and S. Fujibayashi, "Line extraction and pattern detection in a photograph," *Pattern recognition*, vol. 1, no. 3, pp. 233-248, 1969.
- [84] P. Sinha, "Perceiving and recognizing three-dimensional forms," Massachusetts Institute of Technology, 1995.
- [85] B. Scassellati, "Eye finding via face detection for a foveated active vision system," in *AAAI/IAAI*, 1998, pp. 969-976.
- [86] K. Anderson and P. W. McOwan, "Robust real-time face tracker for cluttered environments," *Computer Vision and Image Understanding*, vol. 95, no. 2, pp. 184-200, 2004.
- [87] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38-59, 1995.
- [88] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *European conference on computer vision*, 1998, pp. 484-498: Springer.
- [89] T. F. Cootes and C. J. Taylor, "Locating faces using statistical feature detectors," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, 1996, pp. 204-209: IEEE.
- [90] Y. M. Lui, J. R. Beveridge, A. E. Howe, and L. D. Whitley, "Evolution strategies for matching active appearance models to human faces," in *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, 2007, pp. 1-7: IEEE.
- [91] L. Wolf, "Face recognition, geometric vs. appearance-based," *Encyclopedia of Biometrics*, pp. 495-500, 2015.
- [92] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.
- [93] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, 1994, pp. 84-91: IEEE.

- [94] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Transactions on Pattern analysis and Machine intelligence*, vol. 12, no. 1, pp. 103-108, 1990.
- [95] **M. Dusenberry**. (Jan 2015). *On Eigenfaces: Creating ghost-like images from a set of faces*. Available: <http://mikedusenberry.com/on-eigenfaces/>
- [96] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst2007.
- [97] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 1, pp. 23-38, 1998.
- [98] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*, 1997, pp. 130-136: IEEE.
- [99] A. V. Nefian and M. H. Hayes, "Face detection and recognition using hidden Markov models," in *Image Processing, 1998. ICIIP 98. Proceedings. 1998 International Conference on*, 1998, vol. 1, pp. 141-145: IEEE.
- [100] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, 1995, pp. 23-37: Springer.
- [101] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Computer vision, 1998. sixth international conference on*, 1998, pp. 555-562: IEEE.
- [102] A. Haar, "On the theory of orthogonal function systems," *Math. Ann*, vol. 69, no. 3, pp. 331-371, 1910.
- [103] P. Viola and M. Jones, "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*, 2001, vol. 1, pp. I-I: IEEE.
- [104] M. Da'san, A. Alqudah, and O. Debeir, "Face detection using Viola and Jones method and neural networks," in *Information and Communication Technology Research (ICTRC), 2015 International Conference on*, 2015, pp. 40-43: IEEE.
- [105] A. Gupta and D. R. Tiwari, "Face Detection Using Modified Viola Jones Algorithm," *International Journal of Recent Research in Mathematics Computer Science and Information Technology*, vol. 1, no. 2, pp. 59-66, 2014.
- [106] K. Benhallou, M. Kech, A. Ouamri, and K. Benhallou, "An Efficient face detection based on improved Viola & Jones."
- [107] M. Pantic, "Facial expression recognition," in *Encyclopedia of biometrics*: Springer, 2009, pp. 400-406.
- [108] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 699-714, 2005.
- [109] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, 2006, pp. 149-149: IEEE.
- [110] R. Shbib and S. Zhou, "Facial expression analysis using active shape model," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 1, pp. 9-22, 2015.

- [111] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 132-137.
- [112] D. Datcu and L. Rothkrantz, "Facial expression recognition in still pictures and videos using active appearance models: a comparison approach," in *Proceedings of the 2007 international conference on Computer systems and technologies*, 2007, p. 112: ACM.
- [113] I. Matthews and S. Baker, "Active appearance models revisited," *International journal of computer vision*, vol. 60, no. 2, pp. 135-164, 2004.
- [114] D. Gabor, "Theory of communication. Part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429-441, 1946.
- [115] T. Aach, A. Kaup, and R. Mester, "On texture analysis: Local energy transforms versus quadrature filters," *Signal processing*, vol. 45, no. 2, pp. 173-181, 1995.
- [116] D. J. Fleet, A. D. Jepson, and M. R. Jenkin, "Phase-based disparity measurement," *CVGIP: Image understanding*, vol. 53, no. 2, pp. 198-210, 1991.
- [117] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169-1179, 1988.
- [118] D. Zheng, Y. Zhao, and J. Wang, "Features extraction using a gabor filter family," in *Proceedings of the sixth Lusted International conference, Signal and Image processing, Hawaii*, 2004.
- [119] H.-B. Deng, L.-W. Jin, L.-X. Zhen, and J.-C. Huang, "A new facial expression recognition method based on local gabor filter bank and pca plus lda," *International Journal of Information Technology*, vol. 11, no. 11, pp. 86-96, 2005.
- [120] J. Ou, X.-B. Bai, Y. Pei, L. Ma, and W. Liu, "Automatic facial expression recognition using Gabor filter and expression analysis," in *Computer Modeling and Simulation, 2010. ICCMS'10. Second International Conference on*, 2010, vol. 2, pp. 215-218: IEEE.
- [121] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, no. 6, pp. 615-625, 2006.
- [122] D.-C. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509-512, 1990.
- [123] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51-59, 1996.
- [124] P. Matti, H. Abdenour, and Z. Guoying, "Computer vision using local binary patterns," ed: Berlin: Springer Verlag, 2011.
- [125] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005.
- [126] S. Liao, W. Fan, A. C. Chung, and D.-Y. Yeung, "Facial expression recognition using advanced local binary patterns, tsallis entropies and global

- appearance features," in *Image Processing, 2006 IEEE International Conference on*, 2006, pp. 665-668: IEEE.
- [127] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, 2007.
- [128] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803-816, 2009.
- [129] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893: IEEE.
- [130] G. A. Korn and T. M. Korn, *Mathematical handbook for scientists and engineers: Definitions, theorems, and formulas for reference and review*. Courier Corporation, 2000.
- [131] C. McCormick. (2013). *HOG Person Detector Tutorial*. Available: <http://mccormickml.com/2013/05/09/hog-person-detector-tutorial/>
- [132] C. Orrite, A. Gañán, and G. Rogez, "Hog-based decision tree for facial expression classification," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2009, pp. 176-183: Springer.
- [133] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 884-888: IEEE.
- [134] P. Yang, "Facial expression recognition and expression intensity estimation," Rutgers University-Graduate School-New Brunswick, 2011.
- [135] I. A. Essa and A. P. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, 1995, pp. 360-367: IEEE.
- [136] S. Kimura and M. Yachida, "Facial expression recognition and its degree estimation," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 295-300: IEEE.
- [137] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.
- [138] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1588-1595, 2004.
- [139] M. Gargsha, P. Kuchi, and I. Torkkola, "Facial expression recognition using artificial neural networks," *Artif. Neural Comput. Syst*, pp. 1-6, 2002.
- [140] R. Katratwar and P. Ghonge, "Emotion Analysis by Facial Feature Detection."
- [141] C. L. Lisetti and D. E. Rumelhart, "Facial Expression Recognition Using a Neural Network," in *FLAIRS Conference*, 1998, pp. 328-332.
- [142] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 454-459: IEEE.
- [143] J. Zhao and G. Kearney, "Classifying facial emotions by backpropagation neural networks with fuzzy inputs," in *International Conference on Neural Information Processing*, 1996, vol. 1, pp. 454-457.

- [144] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms.," ed. Proc. 23rd International Conference on Machine Learning., 2006.
- [145] M. Mahadevi and C. Sumathi, "Facial Expression Recognition for Color Images Using Genetic Algorithm."
- [146] H. Nomiya, S. Sakaue, and T. Hochin, "Recognition and intensity estimation of facial expression using ensemble classifiers," in *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*, 2016, pp. 1-6: IEEE.
- [147] N. Sebe, M. S. Lew, I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition using a cauchy naive bayes classifier," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, vol. 1, pp. 17-20: IEEE.
- [148] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and image understanding*, vol. 91, no. 1, pp. 160-187, 2003.
- [149] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162-1171, 2011.
- [150] A. M. Ashir and A. Eleyan, "Facial expression recognition based on image pyramid and single-branch decision tree," *Signal, Image and Video Processing*, pp. 1-8, 2017.
- [151] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856-1863, 2007.
- [152] J. Jia, Y. Xu, S. Zhang, and X. Xue, "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*, 2016, pp. 1-5: IEEE.
- [153] V. Vapnik and A. Chervonenkis, "A note on one class of perceptrons," *Automation and remote control*, vol. 25, no. 1, p. 103, 1964.
- [154] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152: ACM.
- [155] Y. Dou, M. Zhou, J. Wang, and J. Qiang, "Facial Expression Recognition Based-On Saliency Guided Support Vector Machine," in *Computational Intelligence and Design (ISCID), 2016 9th International Symposium on*, 2016, vol. 2, pp. 389-393: IEEE.
- [156] J. de Andrade Fernandes, L. N. Matos, and M. G. dos Santos Aragão, "Geometrical Approaches for Facial Expression Recognition Using Support Vector Machines," in *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*, 2016, pp. 347-354: IEEE.
- [157] Y. Sun and J. Yu, "Facial Expression Recognition by Fusing Gabor and Local Binary Pattern Features," in *International Conference on Multimedia Modeling*, 2017, pp. 209-220: Springer.
- [158] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain," in *Proceedings of the 9th international conference on Multimodal interfaces*, 2007, pp. 15-21: ACM.

- [159] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 2005, vol. 2, pp. II-370: IEEE.
- [160] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 833-851, 2008.
- [161] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 878-883: IEEE.
- [162] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179-188, 1936.
- [163] S. S., "Linear Discriminant Analysis," Available: <http://www.saedsayad.com/lda.htm>
- [164] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 2, pp. 568-573: IEEE.
- [165] Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 83-92, 2013.
- [166] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, 1999.
- [167] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Exploring human visual system: study to aid the development of automatic facial expression recognition framework," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 49-54: IEEE.
- [168] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130-140, 2007.
- [169] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *Joint Pattern Recognition Symposium*, 2003, pp. 297-304: Springer.
- [170] J. Kovac, P. Peer, and F. Solina, *Human skin color clustering for face detection*. IEEE, 2003.
- [171] MATLAB, "CascadeObjectDetector," in *Computer Vision System Toolbox*, 2016b ed: The MathWorks, Inc., 2012.
- [172] K. Delac, M. Grgic, and T. Kos, "Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions," in *International Conference on Systems, Signals and Image Processing*, 2006, vol. 1, pp. 21-23.
- [173] R. Fries and J. Modestino, "Image enhancement by stochastic homomorphic filtering," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 6, pp. 625-637, 1979.

- [174] V. I. Ponomarev and O. B. Pogrebnyak, "Image enhancement by homomorphic filters," in *SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, 1995, pp. 153-159: International Society for Optics and Photonics.
- [175] X.-m. Zhang and L.-s. Shen, "Image contrast enhancement by wavelet based homomorphic filtering," *Acta Electronica Sinica*, vol. 29, no. 4, pp. 531-533, 2001.
- [176] MATLAB2, "extractHOGFeatures," in *Computer Vision System Toolbox*, 2016b ed: The MathWorks, Inc., 2013.
- [177] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE transactions on cybernetics*, vol. 45, no. 8, pp. 1499-1510, 2015.
- [178] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1960-1967.