Analysis of Differences Between Western-Caucasian and East-Asian Basic Expressive Faces for Automatic Facial Expression Recognition

Benitez Garcia Gibran de Jesus

Graduate School of Informatics and Engineering
The University of Electro-Communications

A thesis submitted for the degree of
Doctor of Philosophy in Engineering

The University of Electro-Communications

September 2017

Analysis of Differences Between Western-Caucasian and East-Asian Basic Expressive Faces for Automatic Facial Expression Recognition

Approved by Supervisory Committee:

Chairperson: Prof. Masahide Kaneko
Member : Prof. Kazuo Tanaka
Member : Prof. Hiroshi Yokoi
Member : Prof. Masafumi Uchida
Member : Prof. Takayuki Nagai

顔表情自動認識における西洋人と東洋人の基本的表情の違いに対する分析

BENITEZ GARCIA Gibran de Jesus

概　要

表情認識（FER : Facial Expression Recognition）は、ヒューマン・コンピュータ・インタラクション（HCI）分野における重要な研究トピックの一つになっている。近年における表情の自動認識に関する研究の進展により、統制された環境下及び実環境下において高い認識率が達成されている。ここでは、表情認識システムにおいて問題となる、照明変化、個人差、部分的隠れ等の克服が図られている。これらの研究では、筆者が知る限り、いずれも基本的な顔表情についての文化的な普遍性（表情は人種によらず人類に共通）を前提としている。しかし、この普遍性に対しては近年心理学分野の専門家の一部から疑問が持たれ、反論が唱えられるようになっている。

本論文では、HCIの観点から顔表情の文化的普遍性を評価するために、西洋人と東洋人の典型的顔表情間での違いについて分析を行う。さらに、この分析を行うために顔表情の自動認識システムを提案する。本システムでは、顔を額、眉－目、口、鼻の４つの領域に分割し、各顔領域におけるアピアランス特徴と幾何学的特徴から個別に算出したフーリエ係数によって記述されるハイブリッド特徴を用いている。個々の顔における静的構造を考慮し、最終的にSVM（Support Vector Machines）により分類する。複数の標準的なデータベースに基づいて２つの異なる文化的地域に分けられた顔表情画像を用意する。これらの顔表情画像に対して表情の自動認識と表情顔に対する視覚的評価を行うことによって文化の違いに着目した分析を行う。基準となる表情認識として、西洋人、東洋人双方からの 40 人の被検者による表情認識実験について述べる。評価結果より、個々の顔領域及びそれらの組合せに基づく文化に特有な顔表情の違いを特定する。最後に、これらの違いに対処するための２つの可能な方法を提案する。一つは、各々の文化における典型的な色、形状、テクスチャ特徴の抽出に基づいて予め民族性を検出しておく方法である。もう一つは、最終的な分類プロセスにおいて文化に特有な基本表情を個別に考慮する方法である。

本論文における主たる成果は、以下の通りである。

１）西洋人と東洋人の顔表情におけるアピアランス特徴と幾何学的特徴の違いに対する定性的及び定量的分析
２）顔の部分領域への分割とハイブリッド特徴に基づく表情の自動認識システム
３）表情の自動認識において、多文化圏に関わる顔表情データベースを用いる場合の留意点
４）多文化環境における表情の自動認識における２つの対処方法

本論文の構成は次の通りである。

第１章では、本論文における研究の動機、目的、貢献について述べる。

第２章では、表情認識分野における研究の背景を述べると共に、心理学的観点から関連研究について詳しく述べる。特に、HCIの観点から多文化圏顔表情データベースを対象とした表情認識研究に着目する。

第３章では、顔の領域分割に基づく表情の自動認識手法を提案する。顔を４つの領域に自動分割する。本提案手法では、顔の一つの領域を使うだけでも６つの基本表情を認識することができる。このため、部分的な隠れの問題への対処としても有用である。最後に、同じ顔画像の複数の領域から得られた結果を統合するために、モーダルバリュー手法を提案する。

第４章では、ハイブリッド特徴のフーリエ係数に基づく自動表情認識手法を提案する。本手法では、異なる３つの顔領域の画素値（アピアランス特徴）と形状（幾何学的特徴）から抽出された情報を利用する。部分的隠れの問題にも対処できる。本手法は、アピアランス情報及び幾何学的情報の各々に対する局所フーリエ係数（LFC）及び顔フーリエ記述子（FFD）の組合せに基づいている。更に、特徴抽出時に表情顔から無表情顔を差し引くことによって、各個人の顔の静的構造による影響を考慮する。

第５章では、アピアランス特徴、幾何学的特徴、及びハイブリッド特徴に基づく表情の自動認識及び視覚的分析によって構成される西洋人と東洋人の基本表情の分析について述べる。表情認識に対する分析では、グループ内、グループ外での性能、及び、多文化環境での認識に着目する。人間の被験者を対象とした、基本的な顔表情の認識における文化的な違いを示す実験についても本章で述べる。最後に、多文化環境を対象とした表情認識における２つの可能な対処策を提案する。一つは、予め民族性を検出しておく方法、もう一つは文化に特有な表情に関する既知情報を考慮する方法である。

第６章では、結論及び今後の課題について述べる。

Analysis of Differences Between Western-Caucasian and East-Asian Basic Expressive Faces for Automatic Facial Expression Recognition

Benitez Garcia Gibran de Jesus

ABSTRACT

Facial Expression Recognition (FER) has been one of the main targets of the well-known Human Computer Interaction (HCI) research field. Recent developments on this topic have attained high recognition rates under controlled and “in-the-wild” environments overcoming some of the main problems attached to FER systems, such as illumination changes, individual differences, partial occlusion, and so on. However, to the best of the author’s knowledge, all of those proposals have taken for granted the cultural universality of basic facial expressions of emotion. This hypothesis recently has been questioned and in some degree refuted by certain part of the research community from the psychological viewpoint.

In this dissertation, an analysis of the differences between Western-Caucasian (WSN) and East-Asian (ASN) prototypic facial expressions is presented in order to assess the cultural universality from an HCI viewpoint. In addition, a full automated FER system is proposed for this analysis. This system is based on hybrid features of specific facial regions of forehead, eyes-eyebrows, mouth and nose, which are described by Fourier coefficients calculated individually from appearance and geometric features. The proposal takes advantage of the static structure of individual faces to be finally classified by Support Vector Machines. The culture-specific analysis is composed by automatic facial expression recognition and visual analysis of facial expression images from different standard databases divided into two different cultural datasets. Additionally, a human study applied to 40 subjects from both ethnic races is presented as a baseline. Evaluation results aid in identifying culture-specific facial expression differences based on individual and combined facial regions. Finally, two possible solutions for solving these differences are proposed. The first one builds on an early ethnicity detection which is based on the extraction of color, shape and texture representative features from each culture. The second approach independently considers the culture-specific basic expressions for the final classification process.

In summary, the main contributions of this dissertation are:

1) Qualitative and quantitative analysis of appearance and geometric feature differences between Western-Caucasian and East-Asian facial expressions.

2) A fully automated FER system based on facial region segmentation and hybrid features.

3) The prior considerations for working with multicultural databases on FER.

4) Two possible solutions for FER with multicultural environments.

This dissertation is organized as follows.

Chapter 1 introduced the motivation, objectives and contributions of this dissertation.

Chapter 2 presented, in detail, the background of FER and reviewed the related works from the psychological viewpoint along with the proposals which work with multicultural databases for FER from HCI.

Chapter 3 explained the proposed FER method based on facial region segmentation. The automatic segmentation is focused on four facial regions. This proposal is capable to recognize the six basic expression by using only one part of the face. Therefore, it is useful for dealing with the problem of partial occlusion. Finally a modal value approach is proposed for unifying the different results obtained by facial regions of the same face image.

Chapter 4 described the proposed fully automated FER method based on Fourier coefficients of hybrid features. This method takes advantage of information extracted from pixel intensities (appearance features) and facial shapes (geometric features) of three different facial regions. Hence, it also overcomes the problem of partial occlusion. This proposal is based on a combination of Local Fourier Coefficients (LFC) and Facial Fourier Descriptors (FFD) of appearance and geometric information, respectively. In addition, this method takes into account the effect of the static structure of the faces by subtracting the neutral face from the expressive face at the feature extraction level.

Chapter 5 introduced the proposed analysis of differences between Western-Caucasian (WSN) and East-Asian (ASN) basic facial expressions, it is composed by FER and visual analysis which are divided by appearance, geometric and hybrid features. The FER analysis is focused on in- and out-group performance as well as multicultural tests. The proposed human study which shows cultural differences in perceiving the basic facial expressions, is also described in this chapter. Finally, the two possible solutions for working with multicultural environments are detailed, which are based on an early ethnicity detection and the consideration of previously found culture-specific expressions, respectively.

Chapter 6 drew the conclusion and the future works of this research.

CONTENTS

Abstract i
Contents iii

1. INTRODUCTION 1

1.1 Motivation 2
1.2 Objectives..3
1.3 Contributions 4
1.3.1 FER System based on Facial Region Segmentation................4
1.3.2 FER System based on Hybrid Features 5
1.3.3 Facial Expression Analysis of Differences between Western-Caucasian and East-Asian Faces 5
1.4 Organization of the Thesis 6

2. BACKGROUND 9

2.1 Universality of Facial Expressions of Emotion.........................10
2.2 Facial Expression Recognition..12
2.2.1 Main Problems of FER Systems..12
2.3 FER Methods..17
2.3.1 Face Image Acquisition 17
2.3.2 Facial Feature Extraction...19
2.3.3 Facial Expression Classification..22
2.4 Related Works 23
2.4.1 Psychological Approaches...24
2.4.2 HCI Approaches 24

3. FER SYSTEM BASED ON FACIAL REGION SEGMENTATION 27

3.1 Introduction 28
3.2 Facial Region Segmentation..29
3.3 Feature Extraction 31
3.3.1 Feature Vector Estimation..31
3.4 Modal Value Approach 32
3.5 Database 34

3.6 Experimental Results 36
3.6.1 Individual Facial Regions and its Combinations 37
3.6.2 Despite Partial Occlusion 40
3.7 Conclusion 42

4. FER SYSTEM BASED ON HYBRID FEATURES 45

4.1 Introduction 45
4.2 Appearance Features 48
4.2.1 Local Fourier Coefficients 48
4.2.2 Feature Vector Estimation 49
4.3 Geometric Features 50
4.3.1 Facial Fourier Descriptors 51
4.3.2 Feature Vector Estimation 52
4.4 Hybrid Feature Vector Estimation 52
4.4.1 Consideration of Static Structure of Faces 55
4.5 Experimental Results 56
4.5.1 Databases 57
4.5.2 Preliminary Results 57
4.5.3 Results with Full Databases 63
4.6 Conclusion 66

5. ANALYSIS OF WSN AND ASN FACIAL EXPRESSIONS 69

5.1 Introduction 70
5.2 Datasets 71
5.3 FER Analysis 72
5.4 Visual Analysis 73
5.4.1 Appearance-Based 74
5.4.2 Geometric-Based 75
5.5 Human Study 77
5.6 Solutions for Multicultural FER 78
5.6.1 Early Ethnicity Detection 78
5.6.2 Consideration of Culture-specific Expressions 81
5.7 Experimental Results 82
5.8 Conclusion 89

6. CONCLUSIONS 91

6.1 Conclusion 92
6.2 Future Works 94

REFERENCES 95

LIST OF PUBLICATIONS 103

LIST OF FIGURES

Figure 2.1. Basic structure of FER systems. 12

Figure 2.2. Concept map of main problems facing FER systems. 13

Figure 2.3. Relationships between problems and each step of FER systems. 14

Figure 2.4. Proposed algorithms related to FIA step. 16

Figure 2.5. Proposed algorithms related to FFE step. 20

Figure 2.6 Proposed algorithms related to FEC step. 22

Figure 3.1. Block diagram of proposed system 29

Figure 3.2. Example of facial region segmentation. 30

Figure 3.3. Example of feature vector conformation. 32

Figure 3.4. Diagram of modal value approach. 33

Figure 3.5. Example of partial occlusion simulation of database. From top to bottom: no occlusion, half left, half right, eyes and mouth occlusion. 36

Figure 3.6. Example of half occluded and mirror images. From top to bottom: half left occlusion, mirror right, half right occlusion, and mirror left. 37

Figure 3.7. Comparison between approach using whole image (SBE) and proposed methods without and with 4 different partial occlusions. 41

Figure 3.8. Effect of four types of partial occlusion by each facial expression using the SBE method. 42

Figure 3.9. Effect of four types of partial occlusion by each facial expression using the proposed method. 42

Figure 4.1 General framework of the proposed system using hybrid features. 47

Figure 4.2. Example of the facial region segmentation based on appearance and geometric features. 47

Figure 4.3. Example of the process of facial landmark localization. 51

Figure 4.4. Framework of the fusion process of hybrid Fourier features. 53

Figure 4.5. Framework of the consideration of static structure of the face for feature extraction process. 55

Figure 4.6. Example of two difference images reconstructed from its feature vectors. (a) Expressive images from Y_{Exp}; (b) Neutral images from Y_{Neu}; (c) difference images from Z 56

Figure 4.7. Examples of sub-block segmentation of non-square regions. 58

Figure 4.8. Results of LFC with different sub-block sizes using Eyes-Eyebrows, Nose, Mouth and All facial regions for feature extraction process. 58
Figure 4.9. Examples of facial shapes represented by different number of fiducial points (*K*). 59
Figure 4.10. Results of FFD with different number of fiducial points using Eyes-Eyebrows, Nose, Mouth and All facial regions for feature extraction process. 60
Figure 4.11. Confusion matrices of hybrid features using all regions. (a) DFT proposed method. (b) PCA baseline. 61
Figure 4.12. Confusion matrices of the proposed method using all regions. (a) Appearance-based features. (b) Geometric-based features. 61
Figure 4.13. Example of mistakes of the proposed method using different kind of features. 62
Figure 4.14. Example of a mistake of the proposed method using hybrid features. Face labeled as sadness and misrecognized with anger. 62
Figure 5.1. Example of images included in WSN dataset. 71
Figure 5.2. Example of images included in ASN dataset. 72
Figure 5.3. Average projected image of happiness from ASN dataset represented by using appearance and geometric features. 73
Figure 5.4. Visual representation of average expressions from WSN datasets using appearance features. 75
Figure 5.5. Visual representation of average expressions from ASN datasets using appearance features. 75
Figure 5.6. Example of the facial landmark localization using FaceFit software..75
Figure 5.7. Framework of DrawFace tool [99] for developing facial caricatures. 76
Figure 5.8. Visual representation of average expressions from WSN datasets using geometric features. 77
Figure 5.9. Visual representation of average expressions from ASN datasets using geometric features. 77
Figure 5.10. Screenshots of the software application used for the human study...78
Figure 5.11. FER framework using an early ethnicity detection. 79
Figure 5.12. Ethnicity detection process. 79
Figure 5.13. FER framework based on multicultural training for taking into account the differences on specific facial expressions. 81
Figure 5.14. Two training methods based on cultural-specific groups (WSN and ASN). 81
Figure 5.15. Three possible training options for working with multicultural databases. 82

Figure 5.16. Average projected vectors of six expressions, (a) from ASN and (b) from WSN. 85
Figure 5.17. Confusion matrices of mouth region for out-group test. (a) ASNvsWSN (b) WSNvsASN. 86
Figure 5.18. Confusion matrices of eyes region for out-group test. (a) ASNvsWSN (b) WSNvsASN...86
Figure 5.19. Examples of misclassification. (a) sample from ASN dataset showing anger misrecognized with sadness. (b) sample from WSN dataset showing fear misrecognized with disgust..86

LIST OF TABLES

Table 2.1. Databases more commonly used in FER systems. .... **Error! Bookmark not defined.**

Table 3.1. Example of modal value approach. 34

Table 3.2. Special case of modal value approach when two or more classes have the same decision. 34

Table 3.3. Frame numbers of each expression. 35

Table 3.4. Average recognition rate of individual facial regions and its possible combinations compared with methods from the literature...........................38

Table 3.5. Average recognition rate of the best results using modal value approach. 39

Table 3.6. Confusion matrix of the best result...39

Table 3.7. Average recognition rate of additional regions. 40

Table 3.8. Average recognition rate of the best results from each type of partial occlusion...41

Table 4.1. Number of frames and subject of each database. 57

Table 4.2. Average recognition rate of individual facial regions and its combinations obtained using the proposal and baseline methods...............60

Table 4.3. Average recognition rate of individual facial regions and its combinations by each database. 63

Table 4.4 Comparison with different approaches with CK+................................64

Table 4.5. Comparison with different approaches with MUG. 65

Table 4.6. Comparison with different approaches with TFEID. 65

Table 4.7. Comparison with different approaches under partial occlusion of specific facial regions..66

Table 4.8. Comparison with different approaches which present results with only one facial region...66

Table 5.1. Number of frames and subject of each cultural dataset.......................72

Table 5.2. Average recognition rate of individual facial regions and its combinations obtained by each feature extraction method using sub-sets of WSN and ASN. 83

Table 5.3. Average recognition rate of individual facial regions and its combinations obtained using the proposal and baseline methods...............84

Table 5.4. Classification accuracy per expression using the region of Mouth. 84
Table 5.5. Classification accuracy per expression using the “all” combination. ...85
Table 5.6. Average recognition rate of individual facial regions and its combinations obtained using multicultural datasets only. 87
Table 5.7. Classification accuracy of the proposed method and previous works with cross-cultural tests. 88
Table 5.8. Datasets included in each cultural group of previous cross-cultural studies. 88

CHAPTER I

1. INTRODUCTION

This chapter presents the motivation and main objectives of the research work presented in this thesis. The specific contributions and the organization of the thesis are also listed in this chapter.

CONTENTS OF THE CHAPTER

1.1 Motivation 2

1.2 Objectives 3

1.3 Contributions 4

1.3.1 FER System based on Facial Region Segmentation 4

1.3.2 FER System based on Hybrid Features 5

1.3.3 Facial Expression Analysis of Differences between Western-Caucasian and East-Asian Faces 5

1.4 Organization of the Thesis 6

1.1 Motivation

The face presents essential information about any human being, such as sex, age, race, emotional state, and more. Facial expressions are a set of facial muscle movements which can directly express emotional states. Since the studies of Charles Darwin **[1]**, facial expressions have been considered as a universal language, which can be recognized across different races and cultures around the world. Following the study of Darwin, Paul Ekman et al. **[2]** established the universality of basic facial expressions of emotions which are consistent among cultures. Thus, the prototypic expressions of anger, disgust, fear, happiness, sadness and surprise have known to be universally recognized and expressed.

Since the appearance of the first computers and robots, one of the main targets of HCI (Human Computer Interaction) is to attain a complex interface which can understand and replicate the human emotions. For that reason, in the past two decades many research efforts have been proposed for automatic facial expression recognition (FER) **[3]**. The general approach of any FER system is based on three steps: face detection, feature extraction and expression classification **[4]**. In order to make an accurate analysis of facial expressions, the feature extraction process is crucial. Therefore, FER systems can be divided in three groups based on its feature extraction process: appearance-, geometric- and hybrid-based methods **[5]**. Appearance features represent the skin texture of the face and its changes, like wrinkles and creases. Meanwhile, geometric features represent the shape of the face by using specific feature points from different facial parts. Finally, hybrid features describe the facial characteristics by merging the attributes of appearance- and geometric-based methods **[6]**.

FER systems can be applied to virtual reality, smart environments, user profiling, customer satisfaction analysis, and more **[3]**. Indeed, with the rapid globalization and cultural integration, cross-cultural communication is now fast becoming essential. In addition, wireless communication has merged the needs of understanding the complexities of emotions with remote technology for distance communication **[7]**. Related to this issue, from the psychological field a central debate has emerged about the reliability of the six basic expressions across different cultures. Hence, a certain part of the research community has promoted an opposite theory of the universality hypothesis. They proposed that the facial expressions are based on cultural learning and different races may have different ways to express emotions. Therefore, the six basic expressions cannot cover the requirements of all different cultures **[8]**.

Recent cross-cultural studies have to some degree refuted the assumed universality of prototypic expressions by finding differences in perceiving facial expressions of Western-Caucasian and East-Asian people **[9, 10]**, which are related to the mental representations of six basic facial expressions of emotion. In spite of these findings, from the viewpoint of HCI the cultural universality is taken for granted. Therefore, FER systems do not consider the cultural differences of facial expressions **[11]**. On the other hand, there are still some questions that cannot be precisely answered. For example, should companion robots and digital avatars be designed to display a set of facial expressions that are universally recognized, or should they be adapted to express cultural-specific emotions? In order to try to answer these kind of questions, the differences of facial expressions among cultures have to be taken into account by FER systems. Furthermore, a deep analysis of recognition and especially representation of those differences is needed.

1.2 Objectives

This thesis builds on the knowledge that from a viewpoint of HCI the cultural universality of the six basic expressions is taken for granted. Hence the aim of this research is to analyze the prototypic facial expressions of Western-Caucasians and East-Asians for FER from a composite perspective of HCI and psychology. Following objectives are set as guidelines to fulfil the aim.

- To analyze related works from a psychological viewpoint in detail.
- To study methods from HCI viewpoint that employ multicultural datasets for evaluating FER systems.
- To study the main problems that affect the FER systems.
- To develop an approach to recognize facial expressions from specific facial regions.
- To analyze the effect of the static structure of the face for FER.
- To develop a FER system based on integral facial features, which incorporates features of appearance and geometric facial regions.
- To define a methodology for analyzing all possible scenarios for cross-cultural facial expression recognition.
- To determine automatic tools for analyzing the visual differences of faces while showing facial expressions.

- To compare the findings obtained by proposed FER systems with performance reached by human beings of diverse culture origin.
- To establish the prior considerations when working with multicultural datasets for FER systems.
- To propose possible solutions for multicultural environments of FER taking into account the considerations and differences previously found.

1.3 Contributions

The main contributions of this research work are resumed in the proposals of FER system based on facial region segmentation, FER system based on hybrid features, and facial expression analysis of differences between Western-Caucasian and East-Asian faces.

1.3.1 FER System based on Facial Region Segmentation

The proposed FER system based on facial region segmentation tries to overcome the problem of partial occlusion. The proposal instead of working with a whole image, automatically segments the face into the four facial regions of eyes-eyebrows, forehead, mouth and nose. Thus, in order to solve the partial occlusion problem, the facial regions that are not occluded are used for the classification. Indeed, the use of facial region segmentation also helps to improve the performance of FER for non-occluded faces because it allows the possibility to get several decisions from one facial image. The system employs the sub-block Eigenphases algorithm in the feature extraction process. The main contributions of this system are listed below.

- An automatic method for facial region segmentation based on the eyes distance.
- A method for merging feature vectors of individual facial regions based on PCA.
- Particular solutions of different partial occlusions using non-occluded facial regions.
- A classification approach which employs the most frequent decision of different SVMs for unifying results of different facial regions from the same expressive face.

1.3.2 FER System based on Hybrid Features

The fully automated FER system based on hybrid features reduces the problem of individual differences by taking into account the effect of the static structure of individual faces and improve the recognition rate by merging appearance and geometric features. The system is based on Fourier coefficients and it extracts appearance features by using 2-D Discrete Fourier Transform (DFT) called Local Fourier Coefficients and geometric features by using Fourier Descriptors (FD) called Facial Fourier Descriptors, finally both features are merged using PCA in the feature extraction step. In order to overcome the problem of individual differences, neutral faces are subtracted from the expressive at the level of feature conformation. The main contributions of this system are:

- An appearance-based feature extraction method using Local Fourier Coefficients (LFC).
- A geometric-based feature extraction method using Facial Fourier Descriptors (FFD).
- An approach for fusing feature vectors from different facial regions and types of features.
- A practical solution for the problem of individual differences by employing neutral faces.

1.3.3 Facial Expression Analysis of Differences between Western-Caucasian and East-Asian Faces

The methodical analysis of Western-Caucasian and East-Asian prototypic facial expressions intends to assess the cultural universality of the basic expressions of emotions. This analysis is composed by facial expression recognition and visual analysis. Additionally, a human study supports the culture-specific facial expression differences based on individual and combined facial regions. Based on these findings, two possible solutions for multicultural FER are proposed. In summary, the main contributions of this analysis includes:

- A methodical process for cultural analysis of FER systems based on six different cross-cultural classification modalities supported by psychological studies.
- Supplementary findings to the theory of cultural specificity for FER based on four individual facial regions and its combinations.

- Qualitative and quantitative analysis of appearance and geometric feature differences between Western-Caucasian and East-Asian facial expressions.
- A detailed comparison among cultural differences shown in different datasets and the mental representations reported in the psychological literature.
- The prior considerations when working with multicultural datasets on FER systems.
- Two possible solutions for working with multicultural environments. Based on an early ethnicity detection and independent culture-specific basic expressions for the classification process.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows.

- **Chapter 2**
  Presents, in detail, a description of the long debate about the universality of facial expressions of emotions, the background of FER and related works from the psychological viewpoint along with the proposals which work with multicultural databases for FER from HCI.
- **Chapter 3**
  Explains the proposed FER method based on facial region segmentation which is focused on four facial regions. This proposal is capable to recognize the six basic expression by using only one part of the face. Therefore, a modal value approach is proposed for unifying the different results obtained by facial regions of the same face image.
- **Chapter 4**
  Describes the proposed fully automated FER method based on Fourier coefficients of hybrid features. This method takes advantage of information extracted from pixel intensities (appearance features) and facial shapes (geometric features) of three different facial regions. Hence, it also overcomes the problem of partial occlusion. This proposal is based on a combination of Local Fourier Coefficients (LFC) and Facial Fourier Descriptors (FFD) of appearance and geometric information, respectively. In addition, this method takes into account the effect of the static structure of the faces by subtracting the neutral face from the expressive at the feature extraction level.

- **Chapter 5**
  Introduces the proposed analysis of differences between Western-Caucasian (WSN) and East-Asian (ASN) basic facial expressions, it is composed by FER and visual analysis which are divided by appearance, geometric and hybrid features. The FER analysis is focused on in- and out-group performance of independent (WSN and ASN) cultural and multicultural (MUL) datasets. On the other hand, the visual analysis is based on reconstructed images from Eigenfaces for appearance features and caricature faces for geometric features. The proposed human study which shows cultural differences in perceiving the basic facial expressions, is also described in this chapter. Finally, the two possible solutions for working with multicultural environments are detailed, which are based on an early ethnicity detection and the consideration of previously found culture-specific expressions, respectively.
- **Chapter 6**
  Draws the conclusion and the future works of this research

CHAPTER II

2. BACKGROUND

Facial expressions are the straight link to showing human emotions. However, there is an active debate about the universality of the basic facial expressions of emotion. In order to detail the starting point of the main topic presented in the thesis, this chapter presents details about the long debate about the universality hypothesis, and the principles of automated facial expression recognition systems. In addition, related works of the thesis are also presented here.

CONTENTS OF THE CHAPTER

2.1 Universality of Facial Expressions of Emotion 10

2.2 Facial Expression Recognition 12

2.2.1 Main Problems of FER Systems 12

2.3 FER Methods 17

2.3.1 Face Image Acquisition 17

2.3.2 Facial Feature Extraction 19

2.3.3 Facial Expression Classification 22

2.4 Related Works 23

2.4.1 Psychological Approaches 24

2.4.2 HCI Approaches 24

2.1 Universality of Facial Expressions of Emotion

Facial expressions are a set of facial muscle movements which can directly express human emotions. Charles Darwin was the first one to try to reveal the origins of facial expressions **[1].** He claimed that facial expressions are innate and evolved human behaviors, which can be recognized across different races and cultures around the world. On the other hand, anthropological studies found cultural differences in behaviors expected to be biological, instinctual and therefore universal. For example, the opposite gestures for indicating "yes" and "no" **[12]**, and the masking of negative emotions with smiles in Asian countries **[13]**. Therefore, anthropologists proposed that facial expressions are socially learned and not biologically innate **[14]**.

It is worth noting that facial expressions are part of the communication process among humans, which involves the signaling and decoding of information, in this case, any facial expression. In order to measure both sides of the process, Paul Ekman et al. **[15]** proposed the Facial Action Coding System (FACS) which is focused on anatomical facial muscle movements called Action Units (AUs), with this system they standardized the prototypic expressions of anger, disgust, fear, happiness, sadness and surprise **[2]**. Thus, the universality of six basic facial expressions of emotions was stablished. Studies that support this hypothesis are based on decoding process which involves the perception of emotions based on top-down information. Usually for testing the universal recognition of facial expressions, the *n*-Alternative Forced Choice task where $n = 6$ referred to the six basic expressions. However, the conclusions of those studies do not consider the misclassification errors which can be affected by cultural differences **[8]**. Moreover, Jack R.E. **[8]** summarized several studies of recognition research of "universal" facial expressions and showed that they do not obtain similar levels of recognition across cultures, highlighting that the six prototypic expressions are not universally recognized at the same level and those results required more careful interpretation.

Furthermore, Elfenbein et al. **[16]** defined the in-group advantage hypothesis which establishes that members of different cultural groups have different ways of developing facial expressions, and each person tends to appreciate other people's facial expressions based on their own cultural knowledge. Once more, this hypothesis is based on the missing considerations of the studies that support the cultural universality of six basic expressions. Specifically mentioning that some of these studies do not provide the statistical tests that could show whether any cultural differences emerged in accuracy. Even some researchers assert that, in order to highlight the universal recognition, some works deliberately hid the mentioned information, then the cultural differences could be totally avoided **[17, 18]**. In summary, the in-group advantage affirms that people are more accurate at recognizing facial expressions shown by members of their same cultural group, hypothesis which has been supported by different experiments **[10, 19-21]**.

A different reason for the variation in accuracy of the basic facial expressions of emotion is related to culture-specific decoding rules **[22]**, which may restrain the previous knowledge of certain facial expression, for example, recognizing anger in someone's face but reporting sadness for being a more socially acceptable emotion. Recently, different psychological studies have been proposed in order to address the origins of cultural differences in facial expression recognition **[20, 23, 24]**, and for the first time, they show that some cultures have systematic confusions due to an inadequate decoding strategy which restricts distinguishing certain expressions. The problem lies into a repetitively process of sampling information from certain facial regions that dismisses the rest of the face. For instance, Jack et al. **[25]** demonstrated that Westerners distribute eye fixations across the face, whereas East Asians mainly fixate the eye region. Thus, they conclude that Westerns recognized the six basic expressions with high accuracy, whereas East Asians have several problems to recognize the expressions of fear and disgust, confusing them with surprise and anger, respectively. Summarizing, these studies show that culture builds the expectations of facial expression signals (where to focus when recognizing emotions) and challenge the universality of the six basic facial expressions of emotion.

Certain facial expressions have proved to have biological origins, this is the case for fear and disgust. The opened eyes of fear expression increase visual field for reacting against danger, whereas wrinkled nose of disgust blocks the nasal passage protecting from noxious contaminants **[26]**. However, as mentioned before, these primitive facial expressions presents the lowest recognition rates across cultures **[25]**. Thus, some facial elements that represent facial expressions are shaped by cultural learning and social interactions. On the other hand, a part of the psychological research community believe that only a subset of the six basic facial expressions of emotion is universally recognized, i.e. the expressions of happiness, surprise, anger and sadness **[27, 28]**. Therefore, a new formulation of the cultural universality of facial expressions should be proposed with fewer than six facial expression signals.

As a summary, facial expressions may have biological origins, may have evolved from primitive human beings, and certainly have been shaped by cultural factors and social interactions. On the other hand, the debate about the universality of facial expressions of emotion started just after the proposal of the universality hypothesis more than one-hundred years ago. However, with the new cross-disciplinary approaches, this debate has reached a new juncture where many research works can contribute to revealing some of the still unanswered questions.

2.2 Facial Expression Recognition

Automatic Facial Expression Recognition (FER) is a computer system that attempts to automatically detect and recognize facial motions which describe an specific facial expression **[3]**. The general approach of any automatic FER system consists of three steps: face image acquisition, facial feature extraction and facial expression classification (**Figure 2.1**). Face image acquisition (FIA) is a stage for getting the face image and automatically find the facial region from the input image or sequences of frames. After the face is found, the next step is to extract and describe the facial changes caused by the facial expressions, this task is done by facial feature extraction stage (FFE). Finally, the extracted facial features can be recognized as a group of AUs or prototypic expressions in the facial expression classification stage (FEC) **[4]**.

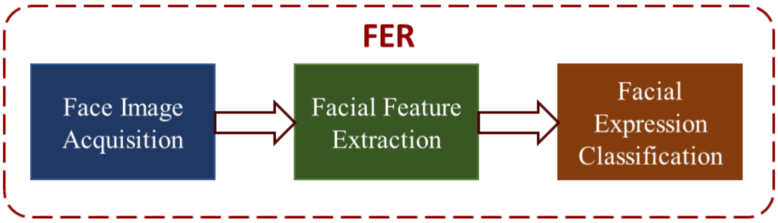


Figure 2.1. Basic structure of FER systems.

2.2.1 Main Problems of FER Systems

There are many problems which affect any FER system, the must recurrent and its relations are illustrated in **Figure 2.2**. A few problems can be easily solved by limiting the database or improving the algorithm used for one of the three steps for the basic structure before mentioned. On the other hand, some problems as those related with the intensity of facial expressions on testing data are not easy to cover.

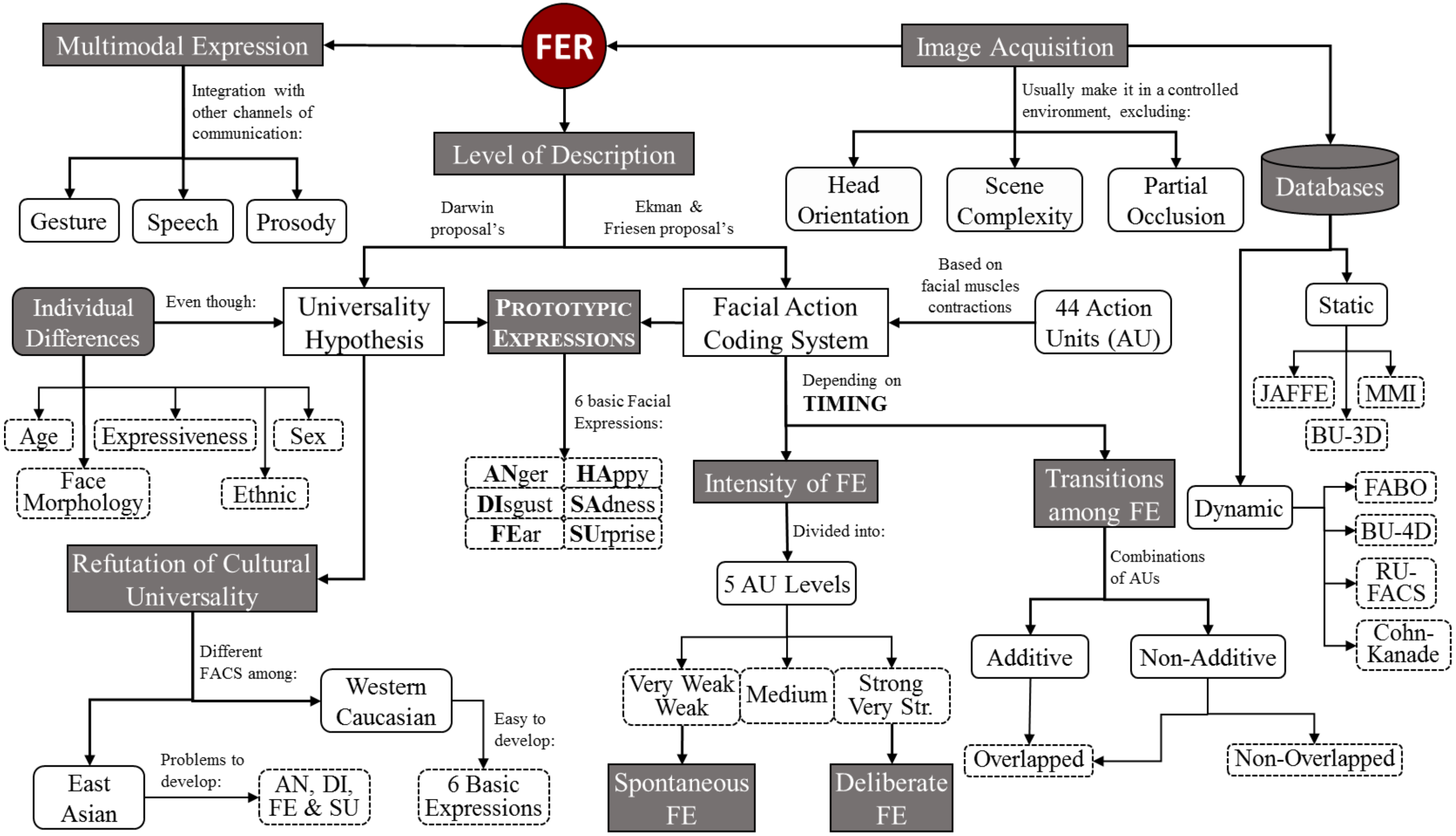


Figure 2.2. Concept map of main problems facing FER systems.

From **Figure 2.2** we can detect 11 main problems (highlighted in grey) which directly affect the performance of FER systems **[3, 4, 6]**. In addition, these problems can be specifically related with one of the three stages of FER structure, these relations are illustrated in **Figure 2.3**. Shown below are listed the mentioned problems with a briefly explanation on their relations.

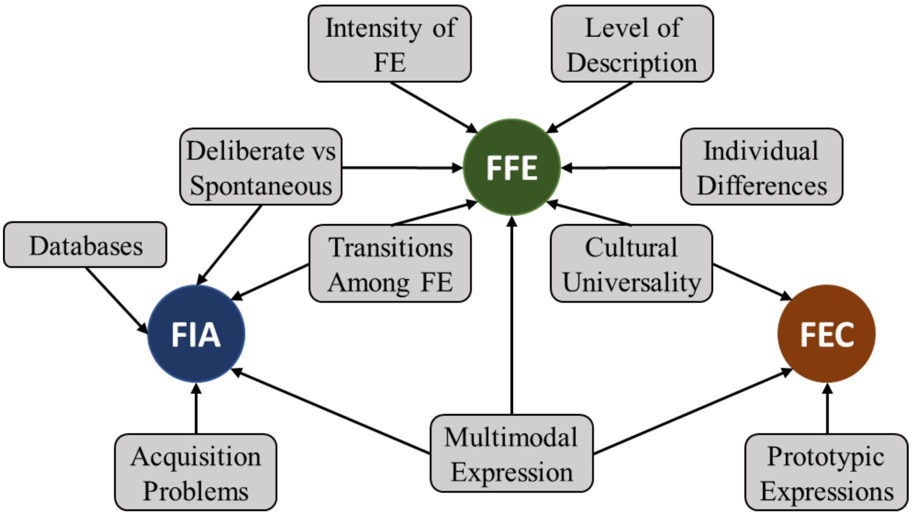


Figure 2.3. Relationships between problems and each step of FER systems.

- ***Acquisition Problems.*** Image acquisition problems are related to the properties of video cameras, the size of the facial region relative to the dimensions of the input image, and all the issues of lighting and background. These factors plus head orientation and partial occlusion may influence face detection and in general, the performance of any FER system. This problem affects only the stage of face image acquisition (FIA).
- ***Databases***. This problem lies on the limited data sets which in some cases lack of diversity with respect to age, gender and ethnic background. Same as previous, this problem is related only with FIA stage.
- ***Deliberate vs Spontaneous Expression.*** Most of the facial expression databases are collected by asking professional actors to perform series of facial expressions. Thus, these representations are not straight linked to the true feelings of the subjects and may differ in appearance and timing from spontaneously occurring behavior. On the other hand, an ideal FER system may properly perform for both, deliberate and spontaneous expressions. This problem faces the FIA stage and it is related with the data set and the way to obtain it. In addition, it is also related with the stage of facial feature extraction (FFE) because the way to describe spontaneous and deliberate facial expressions may be different one from the other.

- ***Transitions among Expressions.*** The assumption of singular expressions which are the last state which started from neutral position is not always feasible because in real life facial expressions are more complex. For example, the variation level of AUs within the same expression could be misclassified. In this case, a database should include combinations of AUs, especially for those that involve co-articulation effects. Thus, this problem is related with FIA and FFE stages because timing should be taken into account for the way to extract its features.
- ***Intensity of Facial Expressions.*** Expressions can vary in intensity, and of course a low intensity level is more difficult to recognize than a peak one. This problem is strongly related with FIA stage because in most cases it has to be handled in a sequence of frames instead of only static pictures.
- ***Level of Description.*** This problem is based on the 6 prototypic expressions (anger, disgust, fear, happiness, sadness and surprise) and the way to recognize all of the possible human facial expressions. In this case, some individual facial actions like wink could be considered as an expression. This problem is strongly related with FFE stage.
- ***Individual Differences.*** Differences in appearance like facial shape, texture, color and age are a latent problem in FER systems. In addition, differences in the way of expressing emotions, related to frequency of peak expressions, degree of facial plasticity, individual morphology and neutral expression misrecognition, are problems that have to be avoided on the FFE stage.
- ***Cultural Universality.*** This problem lies with some experimental proofs that can be found in psychological studies which argued that in general, Asian subjects have difficulties to express some of the six prototypic basic expressions. Therefore, a group of researchers found that Western-Caucasian and East-Asian people cannot develop in same degree the same facial expressions. Basically this problem faces either the FFE and FEC stages.
- ***Prototypic Expressions.*** The prototypic expressions tend to be insufficient to categorize all of the emotional states of human beings. Therefore some researchers propose to use more than just the 6 basic facial expressions, or vice versa. In this case, the problem is related just with the stage of facial expression classification (FEC).
- ***Multimodal Expression.*** Facial expressions are just one part of a complex system which includes different channels of nonverbal communication for describing human behavior. For example, the expression of happiness is often associated with an increase of vocal fundamental frequency. This problem intends to include more than only facial aspects to recognize the human emotions. Hence, it is related with the 3 stages of FER.

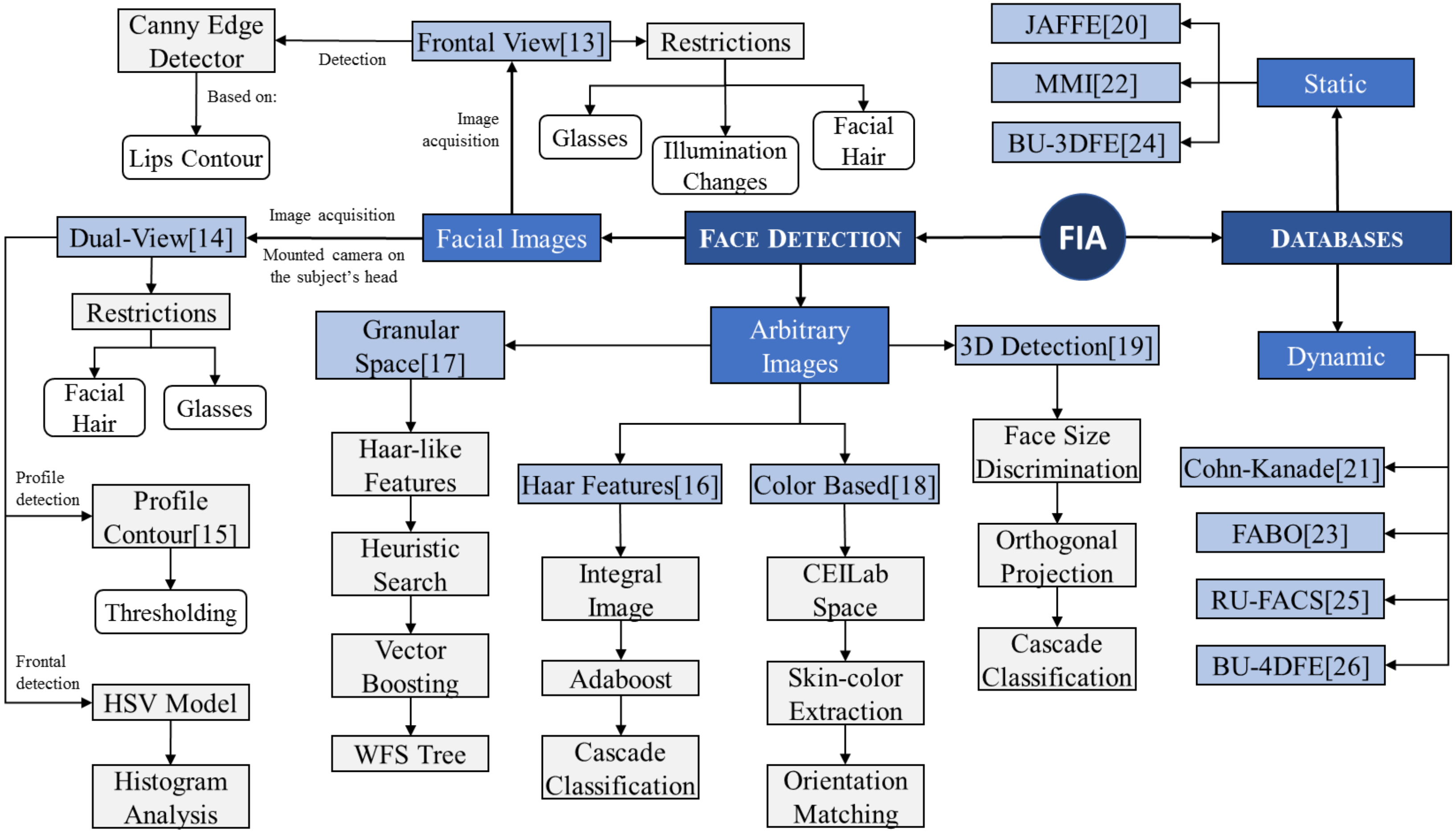


Figure 2.4. Proposed algorithms related to FIA step.

2.3 FER Methods

As mentioned before, all of the problems presented in the previous section are related with the three steps of FER basic structure. Therefore, many of the proposed FER methods are focused in one of those stages. The following sub-sections present all surveyed methods divided on the basic structure steps of: face image acquisition, facial feature extraction and facial expression classification.

2.3.1 Face Image Acquisition

FIA stage is in charge to acquire a face image and properly detect the face on that frame. This step is extremely important because without a good face detection the system can't provide acceptable results. This stage is divided into two sections, first one focused on proposed methods for face detection and second on databases available for FER evaluation. The surveyed methods are presented in the concept map shown in **Figure 2.4**.

Face Detection. Most FER research assumes that face images are properly detected and aligned, otherwise some works include one step of face detection at the beginning of their systems. Face detection algorithms may be performed with photographs taken in controlled environments (facial images) or in arbitrary scenes (arbitrary images). The decision to develop a method which includes what kind of face detection algorithm depends on the data set to be used or even the application of the system.

To detect the face using facial images, Huang and Huang **[29]** obtain an estimation of the facial region location within the image by employing a Canny edge detector. This location is estimated based on the valley of pixel intensities that is generated between the lips and the two vertical boundaries that represent the outline of the face, drawn as two symmetrical edges. The main problems of this approach are related to the partial occlusion, thus the face doesn’t have to present external elements like glasses, scarf, facial hair and medical masks. Pantic and Rothkrantz **[30]** use dual-view facial images, obtained by mounting a camera on a helmet. Their approach is based on a method which employs the HSV color model in order to localize the contour of the face. This algorithm is similar to that based on the relative RGB model **[31]**. Finally, the profile view image is processed by applying a profile-detection algorithm, which defines the profile contour from a thresholded image.

On the other hand, for the algorithms which use images under uncontrolled environments, Viola-Jones **[32]** is the most widely used. Based on a set of rectangle Haar features, they developed a robust real-time face detector. Thus, the features that could be a face are discriminated by using Adaboost in a cascade architecture. Some disadvantages of this method are the problems to handle illumination changes and face rotation. Huang et al. **[33]** developed a rotation invariant multi-view face detector based on Viola-Jones algorithm. This method detects faces with random rotation in-plane and off-plane based on a novel Width-First-Search (WFS) tree detector structure. They employed the Vector Boosting algorithm for learning vector-output strong classifiers, the domain-partition-based weak learning method and the sparse feature in granular space. Deligiannidis and Arabnia **[34]** use the CIELab color space to extracts skin color regions and employs a correlation-based method (Orientation Matching) for the detection of faces as elliptic regions. Finally, Lyons et al. **[35]** based on 3D images performed the face detection using face size discrimination, orthogonal projection and cascade architecture classification.

Table 2.1. Databases more commonly used in FER systems.

Database	*Subjects*	*Expressions*	*Race*	*Format*	*Data*
JAFFE **[36]**	10	6 + Neu	East-Asian	Static	2D
CK+ **[37]**	120	6 + Neu	Multicultural	Dynamic	2D
MMI **[38]**	75	6 + Neu	Caucasian	Static	2D
FABO **[39]**	23	6 + Neu	Caucasian	Dynamic	2D
BU-3DFE **[40]**	100	6 + Neu	Caucasian	Static	3D
RU-FACS **[41]**	100	6 + Neu	Caucasian	Dynamic	3D
MUG **[81]**	123	6 + Neu	Caucasian	Dynamic	2D
TFEID **[82]**	40	6 + Neu	East-Asian	Static	2D
JACFEE **[96]**	56	6 + Neu	Multicultural	Static	2D
GEMEP **[97]**	10	6 + Neu	Caucasian	Dynamic	2D
RU-FACS **[105]**	100	5	Caucasian	Dynamic	3D
Belfast **[106]**	256	6 + Neu	Caucasian	Dynamic	2D
SMIC **[107]**	16	3 + Neu	Multicultural	Dynamic	2D
DISFA **[108]**	27	6 + Neu	Multicultural	Dynamic	2D
SEMAINE **[109]**	150	5 + Neu	Multicultural	Dynamic	2D

Databases. Every FER system needs a database for training and evaluating its performance. Therefore, standard databases are required for comparing different methods and systems. Some of the most used public available databases for FER are listed in Error! Reference source not found..

It is worth noted that not all the databases include the six basic expressions of emotions and most of them are captured in two dimensions format. On the other hand, several databases include only images of Caucasian subjects and just a few includes a large number of East-Asian samples. Finally, recent databases used to include video sequences instead of only static images.

2.3.2 Facial Feature Extraction

After detecting the facial region in the observed scene, the next step is to extract the most relevant information of the showing facial expression. FFE stage is the one in charge of this important task. In general, the way of describing the face can be divided into three types based on the features obtained. Appearance features represent the changes of the face by its texture; geometric features employs the shape and locations of facial components; and hybrid features combine the characteristics of both. The surveyed methods related with FFE stage are presented in the concept map of **Figure 2.5**. As shown in this Figure, the approaches are divided into the three types of representations, which are listed below.

Appearance-based Approaches. Some algorithms are widely used to extract facial features in this modality, such as Gabor filters, PCA and LBP. One of the first algorithms to perform this task was Gabor wavelets. Bartlett et al. **[41]** proposed to use the Gabor energy filters which square and then sum the outputs of two Gabor filters in quadrature. In this way, the authors tried to provide robustness to lighting conditions and to image shifting. In a more recent approach, Gu et al. **[42]** propose to use multi-scaled Gabor filters for local patches. Then, using radial grids the resulted Gabor decompositions are encoded. Finally, the facial expressions are represented by global features obtained by using local classifiers fed by the encoded local features.

The PCA approach called Eigenface was recently used by Mohammadi et al **[43]** where each learned principal component is used as the atom of the dictionary for sparse representation and classification of universal facial expressions. The Fisherface approach is a modification from Eigenface method which use PCA + linear discriminant analysis (LDA). Wang et al. **[44]** use this method to recognize facial expression with infrared images. A relatively recent and powerful texture describing method is Local Binary Patterns (LBP), Zhao and Zhang **[45]** propose to use LBP features compressed and correlated by PCA instead of the traditional LBP histograms. Finally, they used discriminant kernel locally linear embedding.

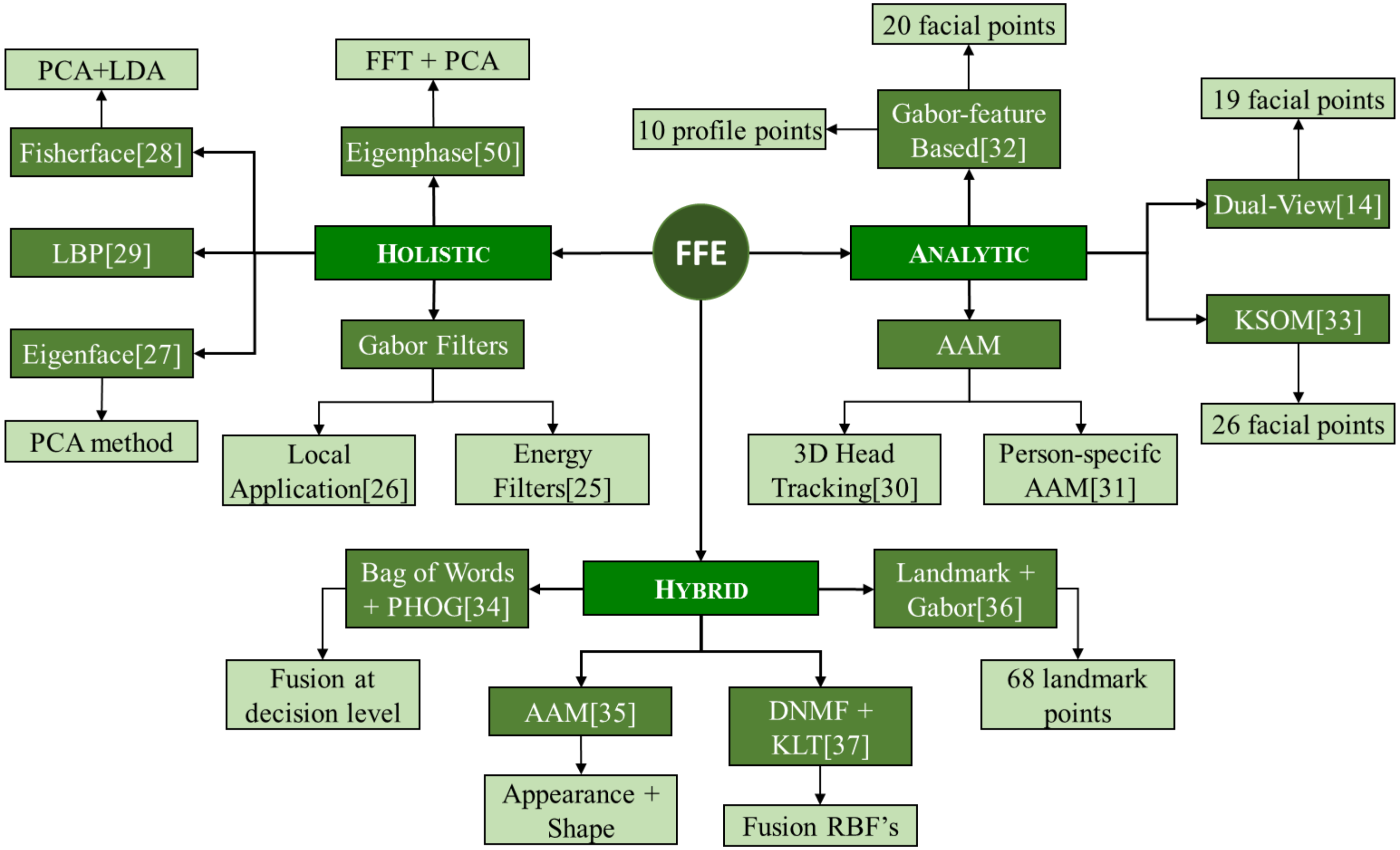


Figure 2.5. Proposed algorithms related to FFE step.

Geometric-based Approaches. The automatic active appearance model (AAM) is the most famous approach used to avoid the manual process of the initialization of geometric-based approaches. Xiao et al. **[46]** used AAM for tracking the head on a 3D environment. This approach also tracks nonrigid features in order to recover the head pose. Finally, the expressions are recognized by using the stabilized facial region which is matched with a common expression orientation. Otherwise, Cohn et al. **[47]** proposed person-specific AMMs to apply gradient descent search using facial shapes in order to detect depression from AUs.

The approach of Pantic and Rothkrantz **[30]** employs a point-based model composed of 20 facial feature points for the frontal-view and 10 profile points for the side-view model. Both points are tracked to define the showing facial expression. Valstar and Pantic **[48]** proposed a landmark detector based on Gabor feature-based boosted classifiers. Using particle filtering with factorized likelihoods, 20 facial feature points are automatically localized and tracked in a sequence of frames. On the other hand, Majumder et al. **[49]** used 26 landmarks for describing the regions of eyes, lips and eyebrows. This, proposal employs extended Kohonen self-organizing map (KSOM) for defining the geometric features.

Hybrid-based Approaches. Li et al. **[50]** used a hybrid method which employs facial component-based bag of words and Pyramid HOG (Histogram of Orientated Gradient) for texture and shape extraction, respectively. They obtained independent results from both feature-based algorithms and combined them on a decision level using SVM. Zhang et al. **[51]** proposed a method to generate the initial model for AAM fitting by using geometric features. This approach detects the facial features more accurately because generates the model in adaptively way. Finally, they employed appearance parameters optimized by adopting quadratic mutual information (QMI) in order to form hybrid feature vectors for describing the whole set of facial features.

Wan and Aggarwal **[52]** proposed a fusion of a face shape generated by 68 facial points, and the texture information obtained by using Gabor filters. In this way, they intend to describe the facial expressions based on local pixel intensity variations in combination with the facial shapes at a global level. Moreover, Kotsia et al **[53]** presented a FER system for video sequences based on Discriminant Non-negative Matrix Factorization (DNMF) and deformed Candide facial grid for extracting appearance and geometric features respectively. Pyramidal variant of Kanade–Lucas–Tomasi (KLT) algorithm is used for tracing the facial points. Finally the fusion stage is carried out by Median Radial Basis Functions (MRBFs). In this way, the facial expressions and a set of AUs can be detected.

2.3.3 Facial Expression Classification

The classification of facial expressions or AUs is carried out in FEC stage, which is the last step of conventional FER systems. In order to achieve this task, different classification methods have been applied. Artificial neural network (ANN), support vector machines (SVM), hidden Markov models (HMM) and Random Forest are some of the most widely used approaches. Any FER system can be divided by its FEC modality, which can be based on static or sequence frames of facial images. Static approaches use only one frame for the recognition process (in some cases the neutral face image can be used as baseline in the FFE step). On the other hand, sequence FEC employs the temporal information obtained by using several frames from a sequence of frames for recognizing the showing facial expression. The concept map of **Figure 2.6** shows the methods surveyed for both FEC modalities.

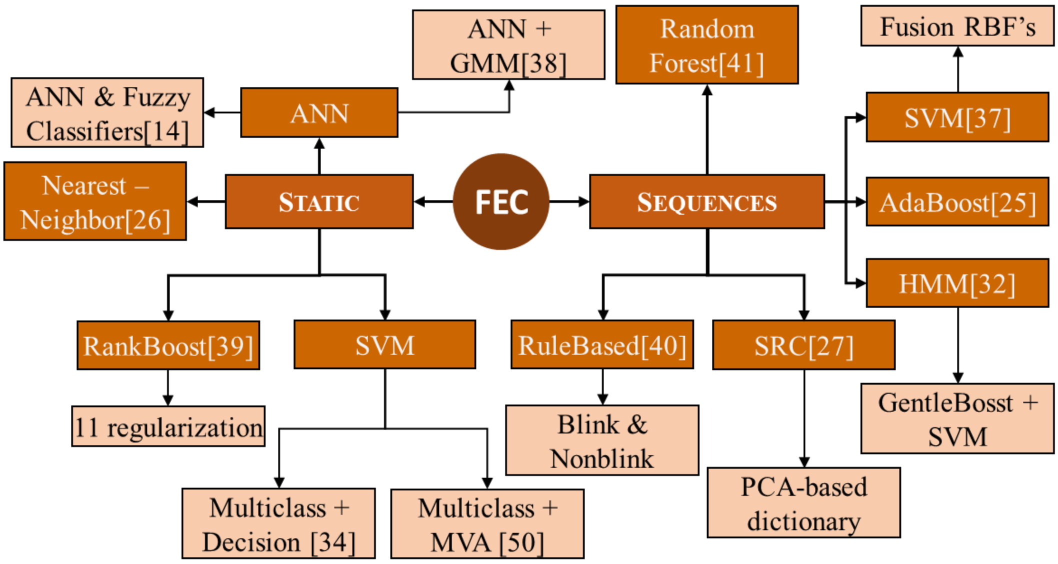


Figure 2.6 Proposed algorithms related to FEC step.

Static FEC. Gu et al. **[42]** used one of the simplest classification algorithms for FER, they used k-nearest-neighbor with Euclidean distance for FEC. ANN which is one of the most widely used algorithm in pattern recognition is employed by Pantic and Rothkrantz **[30]** in combination with Fuzzy Classifiers. On the other hand, Wen and Huang **[54]** proposed a two stage classifier which first recognizes the neutral expression and then one of the six specified expressions. They employed ANN to classify neutral and non-neutral expressions. Otherwise, Gaussian mixture models (GMMs) is employed for classifying the remaining expressions.

SVM is well-known in face recognition and FER, that's why Li et al. **[50]** used this classifier in a multiclass mode to recognize the 6 prototypical expressions independently from holistic and analytic approaches to finally fuse the results by using again SVM in a last decision stage. The algorithm of RankBoost using l1 regularization is employed by Yang et al. in **[55]**. In addition, the output ranking scores of the six specified expressions are used to estimate the intensity of the showing facial expression.

Sequences FEC. Moriyama et al. **[56]** presented a system which recognizes AUs of eyes and eyebrows in spontaneously occurring behavior. This method employs a rule-based classifier powered by several frames of the sequence. Even this proposal is not focused on FER but it recognizes blinks and non-blinks in a spontaneous environment. Otherwise, Bartlett et al. **[41]** also work in the same environment but they performed FER based on FACS using AdaBoost and SVM.

Valstar and Pantic **[48]** applied a combination of GentleBoost, SVM, and hidden Markov models (HMM) to classify AUs and their temporal activation models based on the tracking data. On the other hand, Kotsia et al. **[53]** classify AUs from geometric- and appearance-based features using SVM. In order to detect the facial expression based on the set of showing AUs, many approaches were proposed for fusing both type of features, such as SVM (decision level) and Median Radial Basis Functions (MRBFs).

Based on the Sparse Representation Classification (SRC), Mohammadi et al. **[43]** proposed an space combination of different atoms of a previously defined dictionary. In this way, they assumed that a facial expressive image can be linearly modeled. Recently, Fang et al. **[57]** compared Random Forest against other 5 classification methods based on a framework which explores a parametric space of over 300 dimensions.

2.4 Related Works

As mentioned before, in spite of the proposals which support the universality of basic facial expressions of emotions, a certain part of the research community has promoted an opposite theory. Those proposals started from psychological viewpoint. Thus, related works are clearly divided by psychological and HCI viewpoints. Psychological studies try to prove the refutation of the universality hypothesis of facial expressions, meanwhile, those of HCI attempt to explain the low accuracy performance obtained by cross-cultural tests, taking the universality hypothesis as granted.

2.4.1 Psychological Approaches

From the psychological viewpoint, Dailey et al. **[10]** evaluated the effect of culture-specific facial expression interpretation by analyzing the recognition capability of U.S. and Japanese participants. Their work is based on a human study using a cross-cultural dataset as stimuli. In order to explore the interaction of the assumed universal expressions with cultural learning, the authors proposed to reproduce the previously studied human behavior by using a computational model based on Gabor filtering, PCA and artificial neural networks. Dailey's experiment helps to demonstrate how the interaction with other people in a cultural context defines the way of recognizing a culture-specific facial expression dialect. In summary, they found in-group advantages for recognizing facial expressions, since each racial group was better than the other at classifying facial expressions posed by members of the same culture.

In a more recent study, Jack et al. **[9]** claimed to refute the universal hypothesis of facial expressions by using generative grammars and visual perception for analyzing the mental representations of Western and Eastern cultural individuals. In this proposal, facial expression representations per culture based on the 6 basic emotions were modeled and they found that each emotion is not expressed using a combination of facial movements common to both racial groups. Finally, the authors concluded that the 6 basic emotions can clearly represent the Western facial expressions, but those are inadequate to accurately represent the conceptual space of emotions for East Asians, demonstrating a culture-specific based representation of the basic emotions.

As a summary, the mentioned cross-cultural studies have found differences on perceiving (decoding) facial expressions between cultures, as well as on the mental states related to each basic expression, concluding that facial expressions of emotion could be defined as culture-specific instead of universal. However, these findings are approached from a psychological viewpoint only, thereby did not consider the differences that can be found from automatic FER systems, nor the facial differences produced by expressing the emotions among cultures.

2.4.2 HCI Approaches

From the HCI viewpoint, Da Silva and Pedrini **[58]** proposed an analysis of recognition performance when a restricted out-group scenario occurs. They trained different FER systems using facial expression static images from one specific culture and tested a set of different culture, using only occidental and oriental face databases. This analysis was based on 3 different standard feature extraction methods and 3 machine learning algorithms. For experiments they used the databases of CK+ and MUG as occidental dataset and only JAFFE as oriental. The best results obtained were achieved by in-group test, followed by those of the multicultural test. However, when out-group test was applied, the accuracy dramatically decreased. Finally, they concluded that multicultural training should be considered when an efficient recognition performance is needed, in addition, they pointed out that the six basic expressions are universal with subtle differences which could be influenced by lighting changes or other image problems.

Ali et al. **[59]** performed a similar study, where ensemble classifier construction was intended to find how the classifiers will be trained to accurately classify multicultural databases. This proposal utilized boosted NNE (Neural Network Ensemble) trained with HOG filtered images and combined with Naïve Bayes method for cross-classifying primarily from Moroccan and Caucasian datasets with those of Japanese and Taiwanese. Their 3 datasets were composed by RAFD database for Caucasian and Moroccan; JAFFE for Japanese; TFEID for Taiwanese. Experimental results shown that for out-group test when RAFD or TFEID are used for training the JAFFE performance are noticeable low, nevertheless, when JAFFE is trained the TFEID test performance is slightly better than that of RAFD. However, the best results reported were achieved again by in-group followed by multicultural test. Therefore, Ali et al. concluded that promising results are obtained when multicultural databases are used for training, even increasing the accuracy achieved by the cross-database experiments without regard of decreasing the accuracy in the same database experiments. In addition, they attached the problems of out-group performance not only because of the difference of culture but also factors such as differences in the number of samples per expression, facial structure and visual representation.

In general, many FER studies have included cross-database tests, nevertheless most of them strongly supports the universality hypothesis **[42, 58-61]**. Even when considerable differences among performance accuracy of multicultural and out-group tests are found, they attribute those problems to external factors such as algorithm robustness or image quality rather than question the universality of facial expressions itself.

Based on the literature of cross-cultural studies, it is possible to summarize some important points which must be considered in order to develop a robust cross-cultural analysis of FER:

- Consistency in the size of the datasets and in the number of samples per expression.
- Variety of geographic region, culture, ethnicity and race of the participants.
- Robust facial feature extraction method (especially to illumination changes).
- Review of in-group, out-group, multicultural and out-group multicultural cross-classification.
- Consideration of static structure of individuals faces (neutral face treatment).
- Visual representation of results and human study validation.
- Independent facial region treatment per expression.

CHAPTER III

3. FER SYSTEM BASED ON FACIAL REGION SEGMENTATION

This chapter presents a facial expression recognition system based on segmentation of a face image into four facial regions (eyes-eyebrows, forehead, mouth and nose). In order to unify the different results obtained from facial region combinations, a modal value approach that employs the most frequent decision of the classifiers is proposed.

CONTENTS OF THE CHAPTER

3.1 Introduction 28
3.2 Facial Region Segmentation 29
3.3 Feature Extraction 31
3.3.1 Feature Vector Estimation 31
3.4 Modal Value Approach 32
3.5 Database 34
3.6 Experimental Results 36
3.6.1 Individual Facial Regions and its Combinations 37
3.6.2 Despite Partial Occlusion 40
3.7 Conclusion 42

3.1 Introduction

As mentioned in **Section 2.2.1**, many problems concern FER systems, such as illumination changes, pose, angle of the input camera, partial occlusion, and so on. Partial occlusion can be seen as a noise and could disturb the facial expression feature extraction or it would cause information loss in FER. There are two types of partial occlusion: temporary and systematic. Temporary occlusion is when a part of the face is obscured momentarily as a result of a person moving or an external factor. On the other hand, systematic occlusion results when the person wearing something which covers part of one's face **[62]**. Therefore, to develop an algorithm of robust FER under occlusion conditions has become an important research topic **[63-65]**.

The proposal presented in this chapter introduces a robust FER algorithm which takes into account the problem of partial occlusion. The proposal is based on the segmentation of a face image into several regions using the sub-block eigenphases in each of them, instead of working with a whole image. It uses, specifically, the facial regions that are not occluded. In addition, the use of facial region segmentation also helps to improve the performance of FER for non-occluded faces because it allows the possibility to get several decisions from one facial image. Therefore, a method to unify these results is also proposed. The modal value approach employs the most frequent decision of the classifiers for taking the final decision, in this way all of the different results obtained from one image are unified.

The proposed algorithm was evaluated using leave-one-subject-out method with 300 frames of the Cohn-Kanade database that includes face images of 97 subjects, where each one was instructed to display the six basic facial expressions. In order to evaluate the effectiveness of the proposed method for facial expression recognition under partial occlusion four types of occlusion were adopted: half left, half right, mouth and eyes occlusion. The performance of the proposed method was tested with non-occluded faces as well as partially occluded faces and compared with two recent approaches which use sub-block eigenphases **[66]** and linear binary pattern (LBP) **[67]** respectively.

The block diagram of the proposed system is shown in **Figure 3.1**. In the stage of facial region segmentation, the face images are segmented into 4 facial regions: eye-eyebrow, forehead, mouth and nose. The feature extraction stage is based on sub-block eigenphases algorithm, performing the algorithm independently for each facial region. To apply this algorithm firstly the phase spectrum of the facial region is obtained, then the Principal Component Analysis (PCA) is applied in order to conform an individual feature vector of each facial region and finally conform the final feature vector concatenating the *N* individual feature vectors of facial regions. *N* depends on the number of the facial regions that are used in the process. In the classification stage SVM is applied so as to make the recognition of facial expressions, using the multi-class mode specifically employing 6 classes, one for each expression (anger, disgust, fear, happiness, sadness, and surprise) afterwards based on the decision from different classifiers modal value approach is applied.

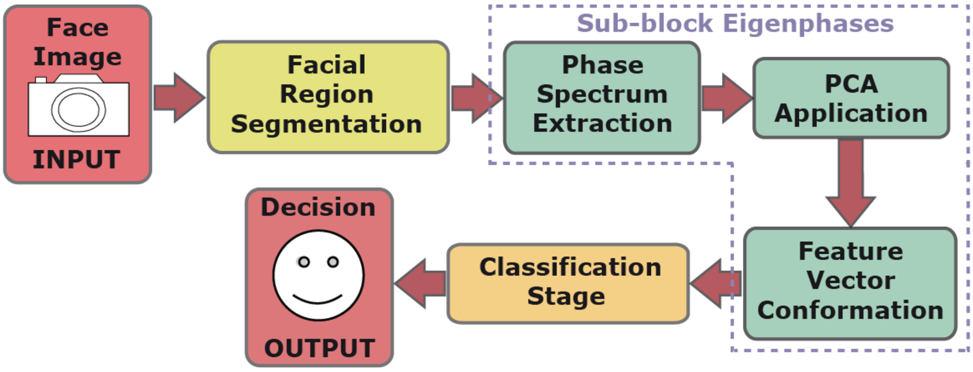


Figure 3.1. Block diagram of proposed system

The baseline algorithm of sub-block eigenphase method has already been proposed in **[66]**, but the original algorithm works with the whole facial image. The proposed method in the present chapter handles each of the facial regions individually and the main focus is how to combine the available facial regions to overcome partial occlusions as well as to achieve better FER performance for non-occluded faces.

3.2 Facial Region Segmentation

As the main proposal of this system, face images are segmented into four regions that contain information of eyes-eyebrows, forehead, mouth and nose. This segmentation enables not only to exclude some facial parts in the case of partial occlusion, but also to evaluate the contribution that each facial region has on the FER, which results in a robust, modal value approach.

The original idea for this segmentation process was proposed in **[50]**, where the distance between the irises (ED) is taken as a baseline for the cropping process. For detecting the eyes position of a face image, they used an algorithm proposed by Vukadinovic et al. **[68]**. However, in this thesis, the Viola-Jones algorithm is used for detecting the face region as well as the eyes position. Thus, we obtain a detected face region (defined as ***DFR***) of size N^2, the eyes position of the face can be defined as E_L, E_R for left and right eye respectively, where $E_L, E_R \in \boldsymbol{DFR}$. Then, in order to segment the facial regions for appearance features, we used the distance between eyes, which is defined as $ED = |E_L - E_R|$ and experimentally we found the relation between *ED* and the three specific facial regions. For instance, consider O as the origin of the plane ***DFR***, where $O = (x_L - x_R, y_L - y_R)/2$. Thus, the upper left vertex of each facial region is defined as follows,

$$\begin{aligned} P_{Eye} &= (x_O - ED\ ,\ y_O + 2/5ED)\,, \\ P_{Nos} &= (x_O - 4/5ED\ ,\ y_O - 2/5ED)\,, \\ P_{Mou} &= (x_O - 3/5ED\ ,\ y_O - 4/5ED)\,, \end{aligned} \tag{3-3-1}$$

where $P_{Eye}, P_{Nos}, P_{Mou}$ represent the initial positions of eyes-eyebrows, nose and mouth regions respectively. Finally, the size of each facial region is defined as,

$$\begin{aligned} A_{Eye} &= 2ED \cdot 4/5ED\,, \\ A_{Nos} &= 8/5ED \cdot 3/5ED\,, \\ A_{Mou} &= 6/5ED \cdot 4/5ED\,, \end{aligned} \tag{3-2}$$

being $A_{Eye}, A_{Nos}, A_{Mou}$ the area of the respective $\boldsymbol{FR}_{Eye}, \boldsymbol{FR}_{Nos}, \boldsymbol{FR}_{Mou}$ facial regions.

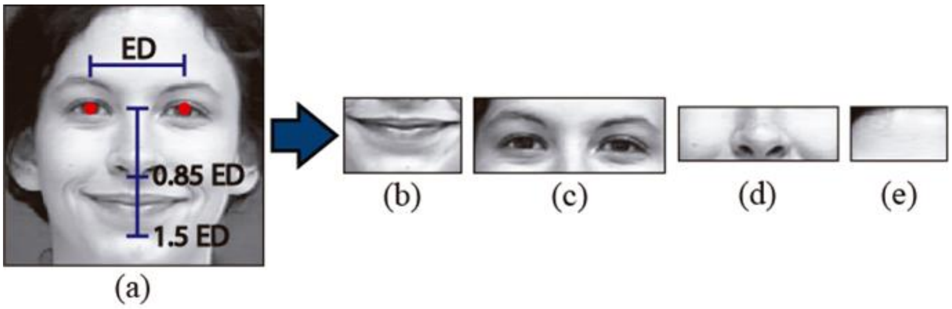


Figure 3.2. Example of facial region segmentation.

This relation proposes that the top of the mouth region is 0.85ED and the bottom 1.5ED from the irises position. This approach is shown in Error! Reference ource not found. (a). Based on this issue, the bottom of the nose region is 0.85ED and the top 0.35ED, the bottom of the eyes-eyebrows region is 0.35ED and the top -0.4ED, finally the bottom of the forehead region is -0.4ED and the top -ED, where “-” means the distance in the upwards direction from irises position.

Subsequently the mouth, eyes-eyebrows, nose and forehead region are segmented which have size of 1.2ED(width)x0.65ED(height), 2EDx0.75ED, 1.7EDx 0.5ED and EDx0.6ED respectively as shown in Error! Reference source ot found. (b), (c), (d) and (e). For this proposal, it is assumed that the segmentation described above is correctly achieved.

3.3 Feature Extraction

Each of the facial regions is further divided into sub-blocks of 2x2 pixels, and the phase spectrum is extracted for each segment by using the fast Fourier transform. The phase spectrum is employed in Eigenphases algorithm **[69]** because the Oppenheim's study **[70]** proved that the most important information of an image is contained in the phase instead of the magnitude. In summary, the process of this step is: to divide the facial region in several sub-blocks, then to obtain the phase spectrum of each sub-block to finally get a phase spectrum matrix of the complete facial region **[66]**.

For applying the PCA first the phase spectrum matrix has to be converted into a column vector, and subsequently the column vectors of all training images will form a matrix in order to calculate the covariance matrix to finally get the matrix of principal characteristics (PM). It is important to notice that in contrast to the original Sub-block Eigenphases algorithm **[66]** which yields only one principal matrix from one face image, in this thesis four principal matrices from one face image were calculated, one matrix from each of the 4 facial regions, respectively.

3.3.1 Feature Vector Estimation

The feature vectors are the product of the principal matrix by each column vector of the training images. Similarly to the previous stage this process is applied to each facial region independently. Thus, the final step is to concatenate the individual feature vectors in order to create the final feature vector that represents the entire face.

As shown in **Figure 3.3**, (a), (b), (c) and (d) represent the process to get the individual feature vector related to eyes-eyebrows, forehead, mouth and nose respectively, finally (e) represents the concatenation of the four individual feature vectors which results in the final feature vector. The final feature vector depends on the number of facial regions used; up to four feature vectors from one facial image are concatenated, although in the case of occlusions, the number could be less than four. To show the contribution of each facial region, all possible combinations of the four facial regions are presented in the experimental results section

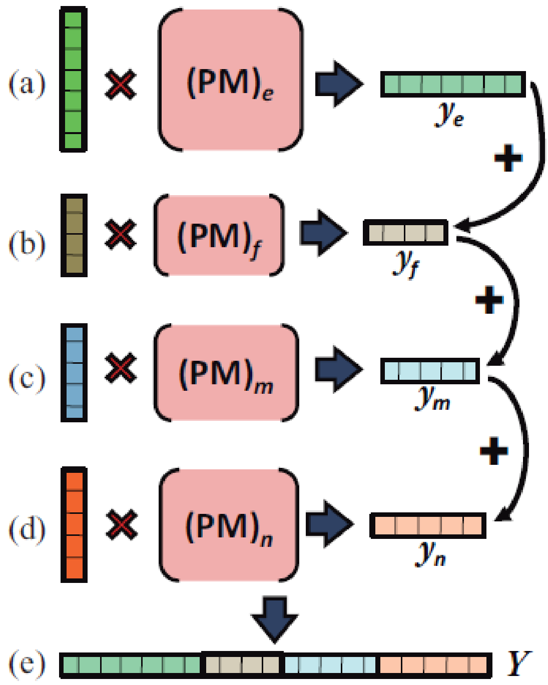


Figure 3.3. Example of feature vector conformation.

3.4 Modal Value Approach

For the classification stage, a multi-class Support Vector Machine (SVM) **[71]** employing RBF kernels were used in order to classify the six basic facial expressions. In this work the library LIBSVM **[72]** is employed to achieve this task. SVM has to be used in two different modalities: training and testing. For training, the equal number of feature vectors should be introduced to the SVM as training images. Accordingly, six templates are obtained, which are linked to the facial expressions of anger, disgust, fear, happiness, sadness, and surprise (6 classes). Afterwards on the testing mode, in broad outlines the SVM compares the test feature vector with all templates to decide from which class it belongs to. It is important to mention that this decision depends on the facial region combination gotten by previous stage. Therefore if many combinations are used, more than one decision can be taken from the same facial image. In addition, the SVM output values provide similarity scores for each class whose range is 0.0 and 1.0, denoting 1.0 the exact match.

The proposed feature extraction method is capable of producing several feature representations from one facial image, because several combinations are possible by concatenating up to four facial regions. For each feature representation a SVM was trained individually, so that several independent decisions from one face image can be obtained. Thus, in order to improve the average recognition rate from each facial region and based on the fact that more than one result can be obtained from each frame tested, a modal value approach to unify these results is proposed, which helps to improve the recognition performance.

Modal value approach consists in selecting as a final decision the most frequently (modal value) class gotten from the classifiers associated with the final feature vector which depends on the combinations of facial regions as described in **Section 3.3.1**.

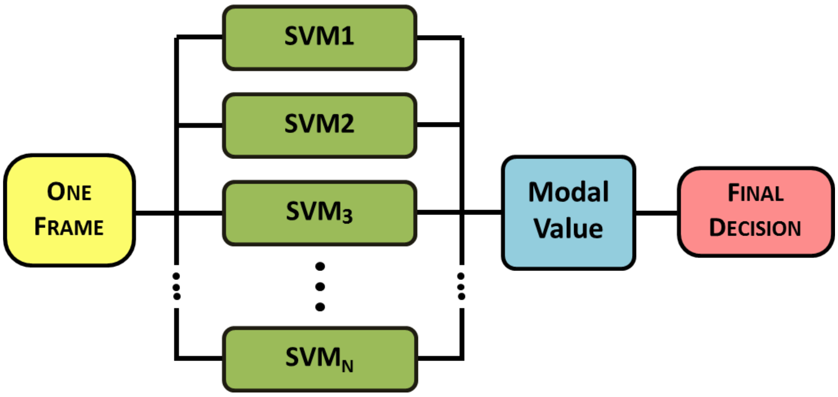


Figure 3.4. Diagram of modal value approach.

Figure 3.4 shows the procedure to apply the modal value approach, where it is possible to see that more than one SVM is associated by one frame and *N* depends on the number of classifiers used in the approach (it depends on combinations of facial regions used in feature vector conformation), modal value is selected from the decision of these classifiers in order to take the final decision. It is important to mention that in order to apply this approach at least 3 classifiers are necessary, because with only two decisions it is not possible to calculate correctly the modal value.

Error! Reference source not found. shows an example of modal value pproach using 4 sample faces. In this example two frames are displaying the expression of anger and the other two of fear. None of the 3 classifiers used here, SVM_1, SVM_2, and SVM_3 produced the perfect results; the recognition accuracies are 3/4, 3/4 and 2/4 respectively. However, unifying the decisions from these classifiers by taking the modal values leads to 100% of accuracy, as shown in the bottom row of this Table.

Table 3.2 presents a special situation that can be occurred when applying the modal value approach. When two or more classes have the same number of positive decisions, the output values of such classifiers are averaged and taking as the final decision the result with the highest numerical value. In this example the decision of the classifiers SVM$_1$ and SVM$_2$ was anger while the decision of SVM$_3$ and SVM$_4$ was fear. However taking into account the average among them the highest value was provided by the average obtained from SVM$_1$ and SVM$_2$ therefore the final decision was anger. It is important to mention that this procedure can also be applied when it is working with only two classifiers. Thus, this process represents an alternative of the modal value approach when only two classifiers are employed.

Table 3.1. Example of modal value approach.

Sample	***1***	***2***	***3***	***4***	***Result***
Exp.	Anger	Anger	Fear	Fear	--
Frame					--
SVM$_1$	Anger	Disg	Fear	Fear	3/4
SVM$_2$	Anger	Anger	Fear	Disg	3/4
SVM$_3$	Fear	Anger	Happy	Fear	2/4
Modal	***Anger***	***Anger***	***Fear***	***Fear***	***4/4***

Table 3.2. Special case of modal value approach when two or more classes have the same decision.

Expression	***Anger***		
Frame			
SVM$_1$	Anger	0.82	**mean**
SVM$_2$	Anger	0.89	**0.855**
SVM$_3$	Fear	0.75	mean
SVM$_4$	Fear	0.80	0.775
Final Decision	***Anger***		

3.5 Database

300 peak expressive frames of the Cohn-Kanade database **[37]** were used for the experiments presented in this chapter. The database contains face images of 97 subjects ranging in age from 18 to 30 years old, 65 percent were female and 45

male. The images were taken under a controlled environment and digitized into 640x480 pixels in grayscale values. For the experiments the face part was cropped at 280x280 pixels.

Table 3.3 shows the number of frames by each expression used in this work, based on the six basic expressions: anger, disgust, fear, happiness, sadness and surprise.

Table 3.3. Frame numbers of each expression.

Expression	Ang	Disg	Fear	Happ	Sad	Surp
No. of images	30	34	47	70	54	65

In order to evaluate the effectiveness of the proposed method under partial occlusion four different types of occlusion were used, accordingly this section is divided into following two subsections.

There is not public available facial expression database that contains different types, position or size of partial occlusion. Therefore, four different types of partial occlusion were simulated in this work: occluded half left, occluded half right, occluded eyes and occluded mouth. The motivation for applying partial occlusions on these regions comes from real situations of daily life. For example, sunglasses often occlude the two eyes and in some cases also the eyebrows, scarves and medical masks often occlude the mouth, and when some people is smiling they put one hand or an object to cover one’s mouth. **Figure 3.5** shows the four different occlusions applied to one subject who displays the 6 basic expressions.

As shown in **Figure 3.5** the four different types of partial occlusion were simulated by superimposing graphically black mask regions on the non-occluded 300 frames selected from Cohn-Kanade database. It is worth nothing that the occlusions introduced in the face images are more critical than real life occlusions. For example in the left and right side occlusions, which emulate occlusion due to hair styles and shadows, half of the face is completely occluded with a black mask which does not happen in the situations mention above. Also to emulate the use of sunglasses, a black mask completely occludes the eyes-eyebrows part of the face which is larger than the real life occlusion. Finally the mouth occlusion used is similar to the occlusion produced by scarves or medical masks. These kinds of occlusions, shown in **Figure 3.5**, which are more critical than real life ones were used because several real life occlusions due to sunglasses and shadows are efficiently solved by sub-block eigenphases algorithm **[66]**. Then if the proposed algorithm is able to solve those occlusions, it can expected that it will be able to perform fairly well with real life occlusions.

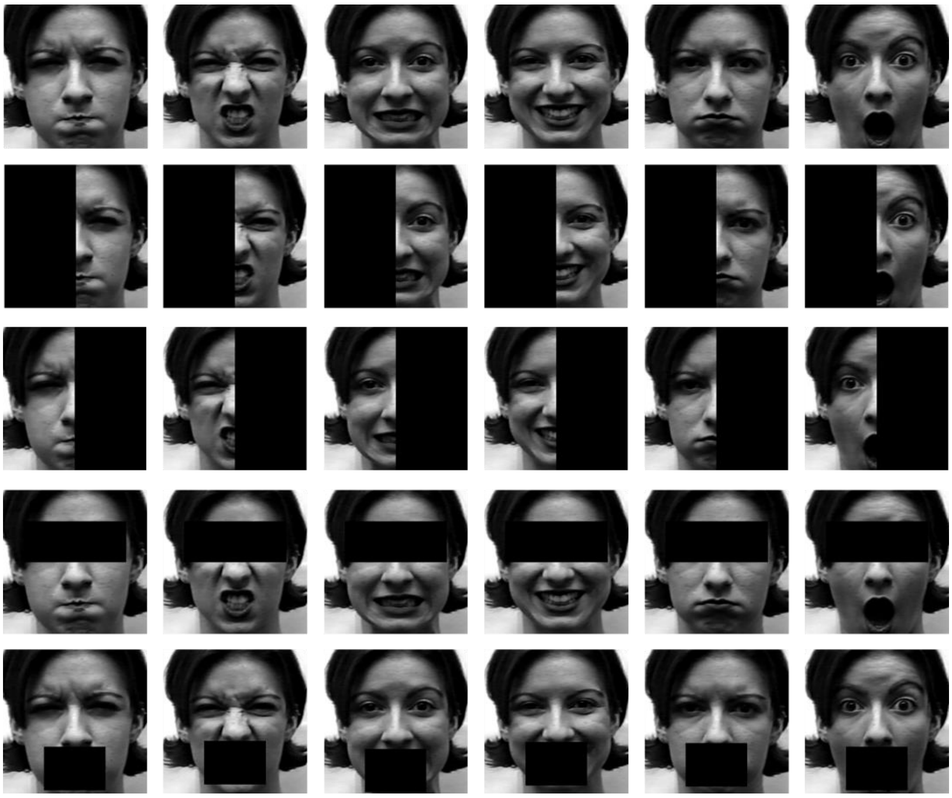


Figure 3.5. Example of partial occlusion simulation of database. From top to bottom: no occlusion, half left, half right, eyes and mouth occlusion.

For the specific case of half left/right face occlusions, the proposed method is not applicable directly because facial regions used in the method depend on both sides of the face. Therefore, the proposal to solve this problem is to generate a mirror image based on the half side not occluded in order to work with a whole face instead of a half. Hence, it is possible to obtain a better recognition rate against to holistic methods.

The mirror images are possible to obtain due to the occlusion is exactly half of the face image. Thus, a mirror image is composed by half not occluded side concatenated with its inverted image (mirror). Error! Reference source not found. hows two subjects of the database with half left/right occlusion and its consequent mirror images. It is important to mention that also this process is applied to all database used in this work.

3.6 Experimental Results

For the experimental results the cropped images are automatically resized to 80x80 pixels. The recognition accuracy was measured using a derivation of leave-one-subject-out which consists in to exclude the test sample of the training stage. The average recognition rates and the confusion matrices have been presented to show the accuracy of facial expression recognition. Confusion matrices for FER show a 6x6 matrix with information about the correct basic expressions in its rows and the predictive classification results in its columns. Thus, diagonal entries represent the correct classification of the system, meanwhile the off-diagonal entries correspond to misclassification problems.

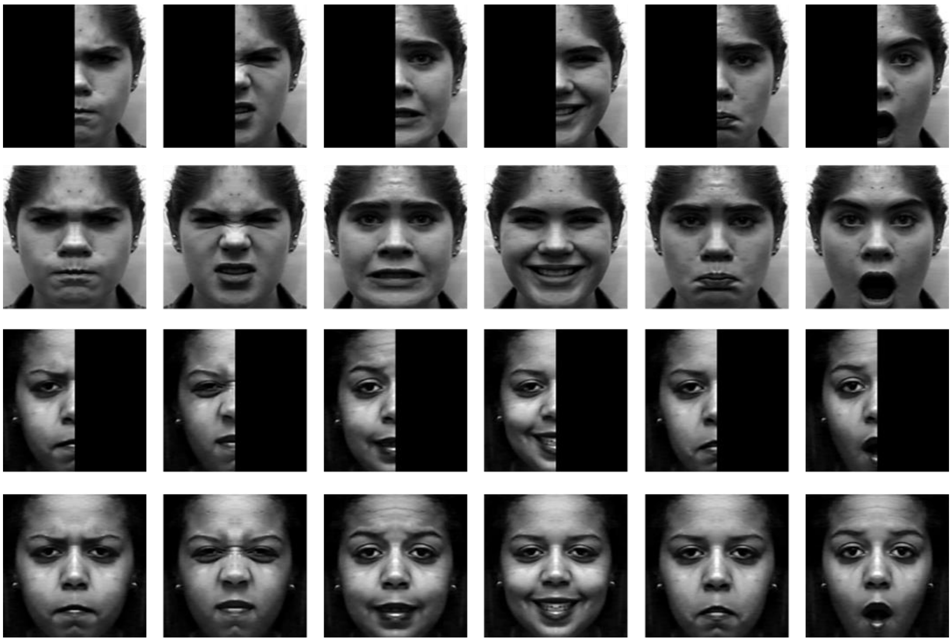


Figure 3.6. Example of half occluded and mirror images. From top to bottom: half left occlusion, mirror right, half right occlusion, and mirror left.

It is important to mention that for all experiments the training was performed with non-occluded database. This section is divided in 3 subsections: experiments with different number of facial regions, experiments despite partial occlusion and effect of partial occlusion on each facial expression.

3.6.1 Individual Facial Regions and its Combinations

Since the proposed method is based on facial region segmentation into 4 regions and the combination of the features individually obtained from those regions, all possible combinations of the 4 facial regions had been analyzed. The proposed system was also compared with two basic methods for reference that uses the whole face image at once to extract a feature vector, without facial region segmentation. For the case of reference methods, the whole image is used for calculating one principal matrix (PM), and that matrix was used to obtain a feature vector that describes the entire face. In this case, recognition was performed with only one classifier, since there are not multiple facial regions.

Table 3.4 presents the average recognition rate of the combinations divided in one, two, three, and all facial regions, compared with the result of reference methods, sub-block eigenphases (SBE) **[66]** and LBP method (LBP) **[67]**. The best result is achieved using 3 regions, by Eyes-Mouth-Nose (EMN) with 87.7% which is better than employing the 4 facial regions (All) with 86.7%. EMN case also outperformes by about 9% the recognition rate obtained using the SBE (with 78.3%) and by about 13% the result provided by the LBP (with 74.3%). Moreover, it can be noticed that Mouth (M) with 79.3% and Eyes-Mouth (EM) with 86% are the best result achieved by using 1 and 2 regions respectively.

Another important point that it is possible to see from **Table 3.4** is that the combinations which use mouths region provide the highest average recognition rate independently of the number of facial regions used in the process. Meanwhile, when the mouth region is not employed, the combinations of other regions do not provide competing recognition rates.

Table 3.4. Average recognition rate of individual facial regions and its possible combinations compared with methods from the literature.

Abbreviation	*Combinations*	*Result (%)*
E	Eyes-eyebrows	53.33
F	Forehead	28.67
M	Mouth	79.33
N	Nose	61.00
EF	Eyes-Forehead	56.67
EM	Eyes-Mouth	86.00
EN	Eyes-Nose	69.33
FM	Forehead-Mouth	79.33
FN	Forehead-Nose	61.67
MN	Mouth-Nose	83.00
EFM	Eyes-Forehead-Mouth	85.00
EFN	Eyes- Forehead-Nose	70.33
EMN	**Eyes-Mouth-Nose**	**87.67**
FMN	Forehead-Mouth-Nose	82.67
All	All regions	86.67
SBE	Sub-block eigenphases **[50]**	78.33
LBP	LBP method **[51]**	74.33

Next, all possible combinations using modal value approach described in **Section 3.4** were tested. In this approach, several SVMs associated with different combinations of facial regions were executed in parallel, obtaining a recognition result from each classifier, and taking the modal value to unify the recognition results. The number of classifiers used in this test was 16, 15 from the combinations of 4 sub-regions and one from the whole image (which is SBE). It is important to mention that for the results obtained, at least 3 classifiers were required to find the modal value.

Table 3.5. Average recognition rate of the best results using modal value approach.

No. SVMs	***Combinations***	***Result (%)***
4	EM – FM – All – SBE	92.00
4	M – FM – EMN – All	91.67
6	M – N – EM – FM – EMN – All	90.00
4	M – EM – EMN – All	89.33
3	FM – EMN – SBE	88.00

Average recognition rates from the best results using modal value approach are shown in **Table 3.5**. It is possible to see that the best result obtained by modal value approach achieves 92% of average recognition rate using 4 classifiers, the combinations used in this case were: Eyes-Mouth (EM), Forehead-Mouth (FM), the four regions (All) and without region segmentation (SBE). This result provides the highest recognition rate obtained in this chapter, outperforming by almost 15% the average recognition rate of the SBE and by almost 19% when LBP is used. On the other hand around 5% is the improvement compared with the best result obtained when only one classifier is used (EMN).

Table 3.6. Confusion matrix of the best result.

	Ang	Disg	Fear	Happ	Sad	Surp
Ang	**83.3**	0.0	3.3	3.3	10.0	0.0
Disg	0.0	**88.2**	5.9	2.9	0.0	2.9
Fear	0.0	0.0	**76.6**	17.0	6.4	0.0
Happ	0.0	0.0	2.9	**97.1**	0.0	0.0
Sad	0.0	0.0	0.0	1.9	**98.1**	0.0
Surp	0.0	0.0	1.5	0.0	0.0	**98.5**

Table 3.6 shows the confusion matrix of the best result obtained by the proposed system (EM-FM-All-SBE). It can be noticed that for the proposed method surprise, sadness, happiness and disgust expressions are easy to recognize, while anger and fear expressions are not. Also, it is possible to see that the average recognition rate of surprise is the highest and fear is the lowest with 98.5% and 76.6% respectively. The problem to recognize fear is related with happiness, because in 17% of the cases the proposed system misrecognizes fear expression with happiness.

3.6.2 Despite Partial Occlusion

The proposed algorithm was evaluated using four types of partial occlusion: occluded half left, occluded half right occluded eyes-eyebrows and occluded mouth, which are described in **Section 3.5** and shown in **Figure 3.5**. To solve these occlusions, in addition to the 4 regions shown in Error! Reference source not found., our additional regions were included in order to have more possible combinations when modal value approach is used, such as: left eye, right eye, half left face and half right face. Here left/right eye is the half part of eyes-eyebrows region shown in Figure 2(c) and half left/right face is the half part of the whole image shown in Figure 2(a). To determine the contribution of these additional regions to the facial expression recognition each of them are used independently and the evaluation results are shown in **Table 3.7**.

Table 3.7. Average recognition rate of additional regions.

Region	***Result (%)***
Left Eye (LE)	61.00
Right Eye (RE)	50.33
Half Left Face (LF)	75.00
Half Right Face (RF)	79.33

For each type of partial occlusion a different solution is proposed. For the specific case of occluded half face a solution using mirror images is employed in which the reconstructed image is segmented into 8 regions (E, F, M, N, LE, RE, LF and RF) described above, which are used together with the whole mirror image in the modal value approach. Otherwise the problem of eyes-eyebrows occlusion is overcome using the facial regions which are not occluded: forehead, mouth and nose. The same solution is employed for mouth occlusion using only eyes-eyebrows, left eye, right eye, forehead and nose regions.

Table 3.8. Average recognition rate of the best results from each type of partial occlusion.

Occluded Region	*Best Solution*	*Result (%)*
Half Left	EMN – All – LF – RF	87.00
Half Right	EMN – All – LF – RF	83.33
Eyes-Eyebrows	M – FM – MN – FMN	87.67
Mouth	N – EN – FN – EFN	75.33

The recognition rates of the best results obtained from each type of partial occlusion are shown in **Table 3.8**. Note that hyphens “-” indicates that the modal value approach was adopted to unify the outputs of multiple classifiers. Therefore, the best solutions for the partial occlusion problems were obtained using 4 classifiers.

In order to compare the results of the proposed system despite occlusion, recognition was performed without mirroring nor facial region segmentation, using only whole image (SBE). The comparison is shown in **Figure 3.7**, where none occluded recognition serving as baseline is also presented. In all cases the proposed method improves the results of the approach using whole image. The maximum improvement is obtained for half left occlusion, thus proposed method using modal value approach improves around 55% the average recognition rate of SBE approach that provides only 33%.

Moreover it is possible to see that the results among half left/right occlusion provide almost the same recognition rate, and the recognition rate with occluded eyes is higher than the result when the mouth is occluded.

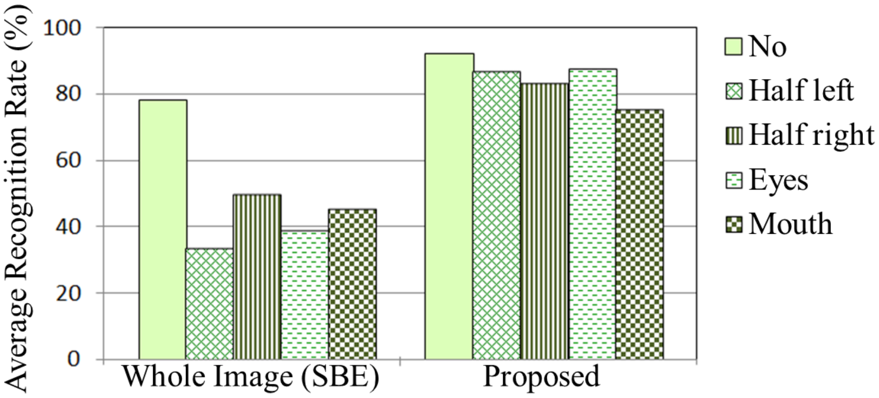


Figure 3.7. Comparison between approach using whole image (SBE) and proposed methods without and with 4 different partial occlusions.

Figure 3.8 and **Figure 3.9** show, by each facial expression, the recognition performance despite the four types of partial occlusion employing SBE and the proposed method respectively. From both images it can be clearly seen that the effect of partial occlusion differs for different expressions.

Figure 3.8 presents interesting results for half left/right occlusion. For example, when half left is occluded, the expressions of disgust, fear and surprise dramatically decrease. On the other hand, when half right is occluded, anger and sadness present several problems. **Figure 3.9** instead presents higher recognition rates and different response of the partial occlusions. For example, the performance of the system for fear degrades when mouth is occluded while for happiness the system performs fairly well with any kind of occlusion.

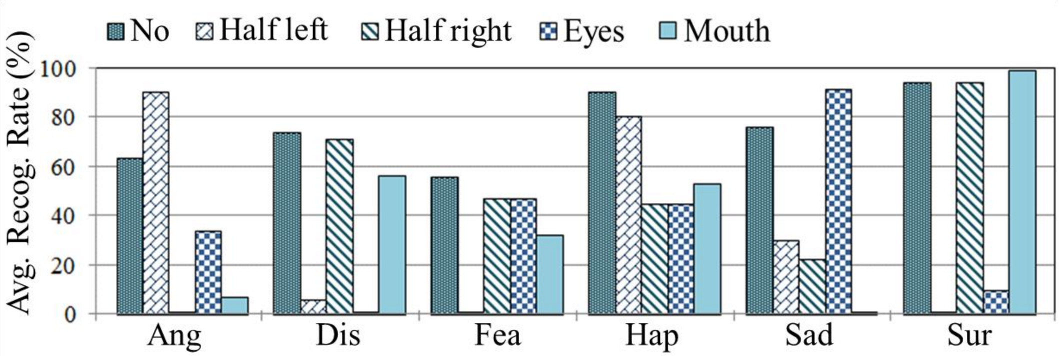


Figure 3.8. Effect of four types of partial occlusion by each facial expression using the SBE method.

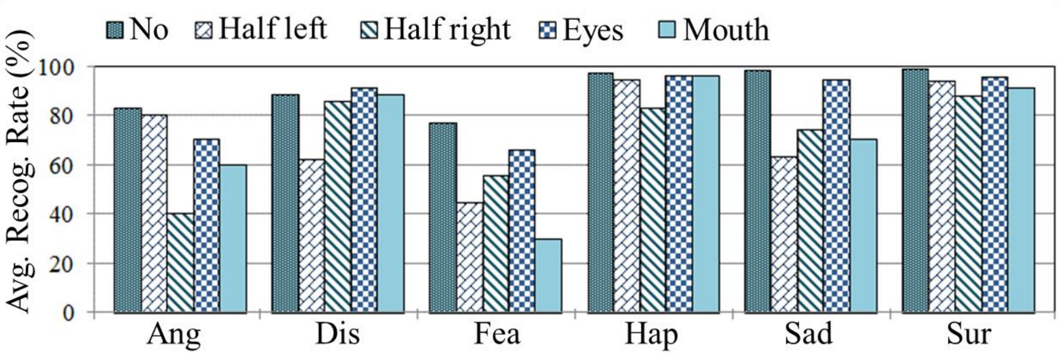


Figure 3.9. Effect of four types of partial occlusion by each facial expression using the proposed method.

3.7 Conclusion

The proposal presented in this chapter introduces a facial expression recognition algorithm based on face image segmentation into four facial regions. Several combinations of facial regions are possible in this approach, resulting in different classifiers corresponding to the combination. In order to unify the results obtained from different classifiers, a modal value approach was also proposed in this chapter. Based on the experimental results, it is possible to conclude that the use of facial region segmentation improves the average recognition rate compared to the approaches which uses the whole image all together (SBE and LBP). Being the best result that obtained by the combination of by Eyes-Mouth-Nose (EMN). In addition, analyzing the performance of individual facial regions and its combinations, it can be concluded that the mouth is the most important part of the face for developing facial expression recognition. This is because all of the results that include this region perform better than the rest.

As a conclusion, it is possible to notice that the best result obtained in the experimental results of this chapter was provided by EM-FM-All-SBE combination in the modal value approach which achieves 92% of average recognition rate. Therefore, the use of modal value approach improves the performance of FER systems when those are built on different facial regions of the same facial image.

Another advantage of the proposed method is that even with only one part of the face it is possible to make the facial expression recognition, achieving almost 80% of average recognition rate if the mouth region is available. This fact becomes very important when several regions of the face are invisible in the case of partial occlusion. If the left or right half of the face is occluded, mirroring images of the non-occluded part can be used. Thus, it is possible to improve the recognition result using facial region segmentation.

Finally, based on the analysis of the effect of partial occlusions on each facial expression, it is clearly noted that fear, anger and disgust are the most difficult expressions to recognize in almost all types of partial occlusion, sadness is not very difficult to recognize for the system especially when the eyes-eyebrows region is occluded. Moreover, happiness and surprise are the expressions that the system can easily recognize despite all types of partial occlusion presented in the **Section 3.5**.

CHAPTER IV

4. FER SYSTEM BASED ON HYBRID FEATURES

This chapter presents a facial expression recognition system based on hybrid features which considers the effect of the static structure of the faces. Appearance and geometric information of specific facial regions are extracted by applying the Discrete Fourier Transform (DFT). In order to decrease the individual differences, this proposal subtracts the static structure from the expressive faces.

CONTENTS OF THE CHAPTER

4.1 Introduction 45
4.2 Appearance Features 48
4.2.1 Local Fourier Coefficients 48
4.2.2 Feature Vector Estimation 49
4.3 Geometric Features 50
4.3.1 Facial Fourier Descriptors 51
4.3.2 Feature Vector Estimation 52
4.4 Hybrid Feature Vector Estimation 52
4.4.1 Consideration of Static Structure of Faces 55
4.5 Experimental Results 56
4.5.1 Databases 57
4.5.2 Preliminary Results 57
4.5.3 Results with Full Databases 63
4.6 Conclusion 66

4.1 Introduction

As mentioned before, FER systems based on its feature extraction process can be divided in two groups: appearance- and geometric-based methods. Appearance features represent the skin texture of the face and its changes (wrinkles and creases), meanwhile geometric features represent the shape of the face by using specific feature points from different facial parts. In addition, there is a special type of feature extraction method called hybrid-based that merges both kind of facial features **[5]**. Hybrid approach represents a FER system which employs appearance and geometric features for describing facial expressions. The process of features fusion can be applied in the step of feature extraction as well as in that of classification **[50, 73-75]**.

Among the algorithms used for feature extraction, the Discrete Fourier Transform (DFT) has been successfully applied for facial recognition. For example, in **[76]** three different Fourier feature domains were fused for face recognition, and in **[66]** phase spectrums of non-overlapped sub-blocks using PCA. However, those approaches were applied using only appearance features.

The proposal presented in this chapter introduces a fully automated FER system based local Fourier coefficients (appearance features) and facial Fourier descriptors (geometric features), finally both features are merged using PCA (Principal Component Analysis) in the feature extraction step. In order to overcome the problem of individual differences presented in any FER system, the static structure of faces was considered. To this end, the neutral faces are subtracted from expressive faces. It is important to mention that this subtraction is also included in the feature extraction process. In order to locally analyze the changes that may appear on the face while showing facial expressions, this proposal is based on 51 facial points (geometric features) and three specific facial regions: eyes-eyebrows, nose and mouth (appearance features). Furthermore, this proposal is extended to the combinations that can be formed by each independent region. Finally, facial expressions are recognized using Support Vector Machines (SVMs).

The proposed system was evaluated using leave-one-subject-out method with four standard databases: the extended Cohn-Kanade (CK+), MUG and TFEID database. In order to evaluate the effectiveness of the fusion using Fourier coefficients, the system was tested without applying the Fourier Transform and obtaining individual feature vectors for appearance and geometric features, this preliminary evaluation was carried out using a subset of the CK+ database. In addition, studies of local Fourier coefficients with different size of sub-blocks and facial Fourier descriptors with a different number of fiducial points are also presented. Finally, the performance of the proposed method is compared with different methods from the literature, which use hybrid features and the same databases.

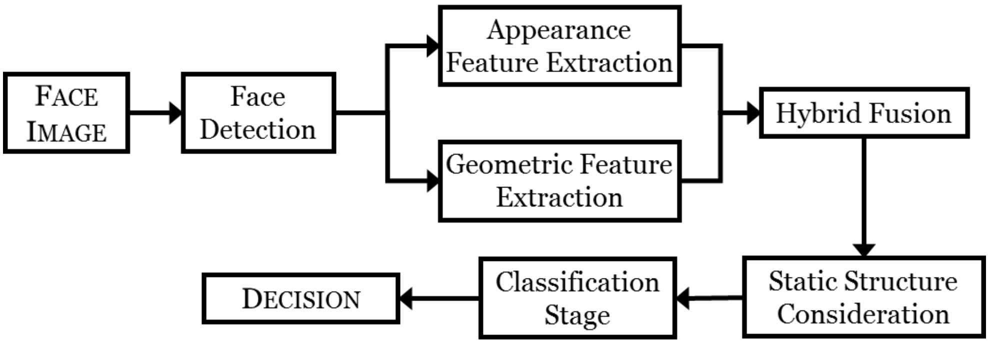

FACE IMAGE
Face Detection
Appearance Feature Extraction
Geometric Feature Extraction
Hybrid Fusion
Static Structure Consideration
Classification Stage
DECISION

Figure 4.1 General framework of the proposed system using hybrid features.

Figure 4.1 presents the general framework of the proposed method. The system is based on the face detection using Viola-Jones algorithm, following by the feature extraction process which is divided into four phases: appearance features extraction, geometric features extraction, hybrid fusion and the consideration of the static structure of faces. Finally in the classification stage SVM is applied so as to make the recognition of facial expressions, using the multi-class mode specifically employing 6 classes, one for each expression (anger, disgust, fear, happiness, sadness, and surprise).

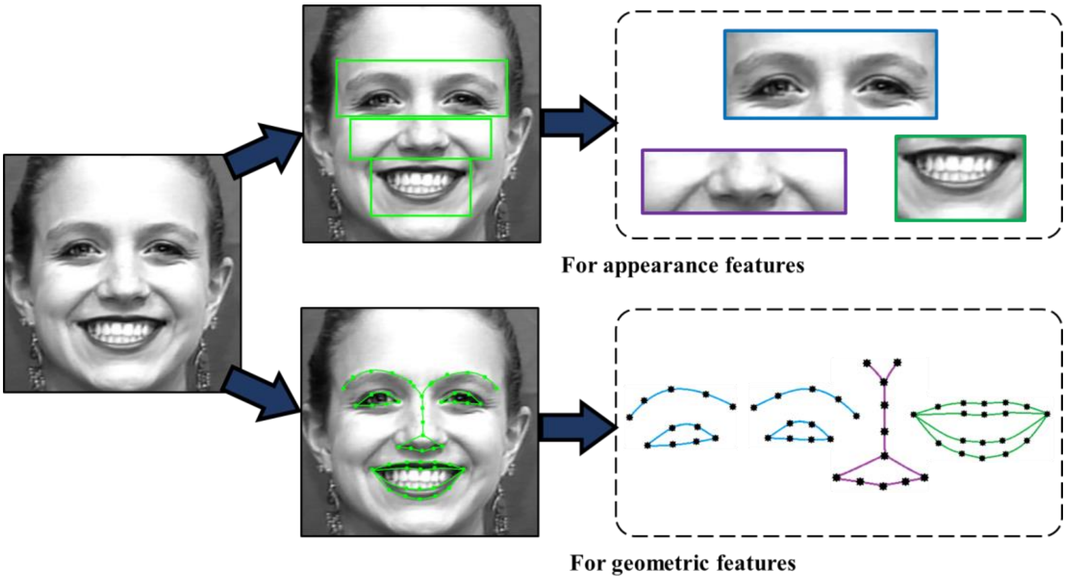


Figure 4.2. Example of the facial region segmentation based on appearance and geometric features.

As mentioned before, the system is based on three different facial regions. Thus, the facial region segmentation is included in each feature-based extraction individually. An example of this region segmentation of both feature-based methods is shown in **Figure 4.2**.

4.2 Appearance Features

Appearance features are individually obtained from three independent facial regions. Thus, for each facial region the proposed feature extraction method was independently applied. The first step of this process is facial region segmentation which is based on the distance between irises and its relation with the rest facial regions as explained in **Section 3.2**.

It is important to mention that the proposed method is based on the sub-block analysis of Eigenphases algorithm presented in **[66]** which concludes that for face recognition, the ideal size of a sub-block for facing illumination problems is the smallest possible. However since the previous analysis of the ideal sub-block size of eigenphases algorithm found that it is equal to 2x2 pixels, and due to the complex part of the Fourier transform being equal to zero in this particular case. Rather than using the phase spectrum for the feature extraction process as in **[66]**, in this thesis the use Local Fourier Coefficients (LFC) was proposed.

4.2.1 Local Fourier Coefficients

Appearance feature extraction is carried out by using LFC which builds on the 2-D DFT. This process consists of dividing the input image into several sub-blocks to locally extract Fourier coefficients. For instance, the 2-D DFT is defined as:

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-2j\pi(ux/M+vy/N)}, \tag{4-1}$$

where $f(x,y)$ is a digital image of size $M\times N$ and it must be evaluated for values of the discrete variables u and v in the ranges $u=0,1,2,\ldots,M-1$ and $v=0,1,2,\ldots,N-1$.

For instance, consider $\boldsymbol{FR}_{roi}$ as the *roi*-th facial region image of size $M\times N$, and for convenience, $\boldsymbol{FR}$ represents any of the three facial regions which have to be divided into sub-blocks of size $L\times L$. Then, the local 2-D DFT of the current facial region is given by a modification of the **Equation (4-1)**:

$$f_{p,q}(u,v)=\sum_{x=0}^{L-1}\sum_{y=0}^{L-1}\boldsymbol{FR}_{p,q}(x,y)\,e^{-j2\pi(ux/L+vy/L)}, \tag{4-2}$$

where $0 \le u, v < L$, and $\boldsymbol{FR}_{p,q}(x,y)$ represents the (p,q)-th sub-block of the facial region $\boldsymbol{FR}$. Since the minimum sub-block size is $L = 2$, the imaginary component of complex Fourier coefficients is equal to zero so that

$$f_{p,q}(u,v) = \boldsymbol{Re}(u,v) + j \times 0 \times \boldsymbol{Im}(u,v)\,, \tag{4-3}$$

where $\boldsymbol{Re}(u,v)$ and $\boldsymbol{Im}(u,v)$ are the real and imaginary components of $f_{p,q}(u,v)$ respectively. The ideal size of L has been analyzed in **[66]**, however this analysis is focused only on the phase component of the Fourier transform. Therefore, an analysis of the ideal sub-block size for real components of LFC is presented in **Section 4.5.2**.

Thus, the local Fourier coefficient matrix is given by:

$$\boldsymbol{lfc} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,N/L} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,N/L} \\ \vdots & \vdots & \ddots & \vdots \\ f_{M/L,1} & f_{M/L,2} & \cdots & f_{M/L,N/L} \end{bmatrix}, \tag{4-4}$$

where $\boldsymbol{lfc}$ has the same dimensions as $\boldsymbol{FR}$. In summary, $\boldsymbol{lfc}$ matrix represents the real components of frequency features obtained locally by each sub-block of size $L \times L$.

4.2.2 Feature Vector Estimation

In order to have correlated information with the set of training images and for dimensionality reduction, the PCA was applied. To this end, the $\boldsymbol{lfc}$ matrix is converted into a column vector, so that

$$\boldsymbol{V}_{lfc} = \mathrm{vec}(\boldsymbol{lfc}(m,n)), \tag{4-5}$$

where $\boldsymbol{V}_{lfc}$ is the column vector of $\boldsymbol{lfc}$ for $0 \le m,n < M,N$. Next, LFC vectors of the training set have to be concatenated to form the matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi}_{lfc} = \left[\boldsymbol{V}_{lfc}{}^{0} - \boldsymbol{\mu}_{lfc} \;,\; \boldsymbol{V}_{lfc}{}^{1} - \boldsymbol{\mu}_{lfc} \;,\; \ldots \;,\; \boldsymbol{V}_{lfc}{}^{P-1} - \boldsymbol{\mu}_{lfc}\right], \tag{4-6}$$

where P is the total number of images used for training and $\boldsymbol{\mu}_{lfc}$ is given by:

$$\boldsymbol{\mu}_{lfc} = \frac{1}{P}\sum_{n=0}^{P-1} \boldsymbol{V}_{lfc}(n)\,, \tag{4-7}$$

Next, the eigenvalues of the covariance matrix:

$$\boldsymbol{\Omega}_{lfc} = \boldsymbol{\Phi}_{lfc}{}^{T}\boldsymbol{\Phi}_{lfc}\,, \tag{4-8}$$

are estimated which has up to P eigenvectors associated with non-zero eigenvalues, where $P < M \times N$. Those eigenvectors are then stored in a descendent order according to the corresponding eigenvalues. The sorted eigenvectors of the covariance matrix determine the subspace $\boldsymbol{\Psi}_{lfc}$ associated to the current facial region, which is defined by:

$$\boldsymbol{\Psi}_{lfc} = \left[\boldsymbol{V}_0^T, \boldsymbol{V}_1^T, \cdots, \ \boldsymbol{V}_{H-1}^T\right], \tag{4-9}$$

where V_0 is the eigenvector associated with the largest eigenvalue, V_1 is the eigenvector associated with the second largest eigenvalue and so on, and H is the number of eigenvectors used for the further projections. It is worth noting that this process is applied so that 90% of the variance of training Fourier coefficient vectors is retained. Finally, the reduced space feature vector of appearance features $\boldsymbol{Y}_{lfc}$ is given by:

$$\boldsymbol{Y}_{lfc} = \boldsymbol{\Psi}_{lfc}{}^T \left(\boldsymbol{V}_{lfc} - \boldsymbol{\mu}_{lfc}\right), \tag{4-10}$$

where $\boldsymbol{\Psi}_{lfc}$ is the facial region subspace and $\boldsymbol{\mu}_{lfc}$ is the mean vector of all training images..

4.3 Geometric Features

As mentioned before, the geometric features are also focused on three independent regions of the face. However, they define the shape of the specific facial regions instead of texture information. The process of facial landmark localization comes after the face detection applied by using Viola-Jones algorithm. For this approach, 51 facial points were used for describing the shapes of eyes-eyebrows, nose and lips. To this end, a deformable face tracking model **[77]** trained by employing a cascade of linear regression functions was employed, which was performed by detecting the face in the first frame and then applying facial landmark localization at each consecutive frame using fitting results of the previous frame as initialization. This algorithm has been tested for working under controlled environments as well as “in-the-wild” **[78]**. An example of the facial landmark localization process is illustrated in **Figure 4.3** which shows the face detected by Viola-Jones, the initialization of the deformable face model and the shape represented by the final fitting facial points.

As well as the previous method, the proposal of geometric features is based on the Fourier transform. However, the application of DFT in this particular case is known as Fourier Descriptor (FD), which is a contour-based shape descriptor widely used for content-based image retrieval (CBIR) **[79]**. Moreover, FD has been used a few times for face recognition **[80]**, and as far as the author knows, it has not been applied to FER. For that reason, the application Facial Fourier Descriptors (FFD) was proposed as a feature extraction method for these kind of features.

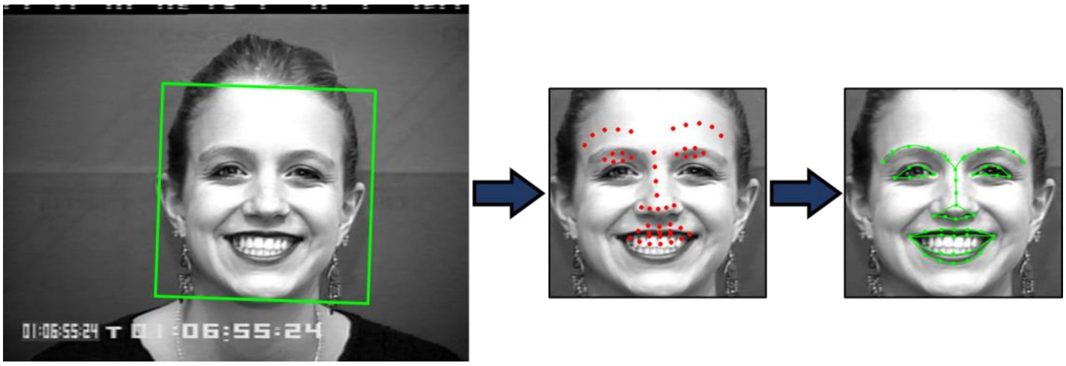


Figure 4.3. Example of the process of facial landmark localization.

4.3.1 Facial Fourier Descriptors

FFD represents a digital boundary of 1D Fourier coefficients estimated by a sequence of coordinate pairs transformed by applying the DFT. To this end, each facial region shape is considered as *K*-point coordinate pairs, *K* being the number of facial feature points of the shape. An analysis of the effect of different number of *K* is presented in **Section 4.5.2**.

In order to apply the FFD process, suppose that a specific shape of the ***FR***-th facial region is represented as a sequence of coordinates, so that:

$$s_{\mathrm{FR}}(k) = \left[x_{\mathrm{FR}}(k), y_{\mathrm{FR}}(k)\right], \qquad (4\text{-}11)$$

where $k = 0,1,2,\ldots,K-1$. From one of the four facial regions, complex numbers have to be generated from each coordinate pairs, as in

$$s(k) = \left[x(k) - x_c\right] + j\left[y(k) - y_c\right], \qquad (4\text{-}12)$$

where (x_c, y_c) represents the centroid of the shape, which is the average of the coordinate pairs so that

$$x_c = \frac{1}{K}\sum_{t=0}^{K-1} x(t), \quad y_c = \frac{1}{K}\sum_{t=0}^{K-1} y(t). \qquad (4\text{-}13)$$

Subsequently, the FFD of $s(k)$ is given by

$$ffd(u) = \sum_{k=0}^{K-1} s(k)\, e^{-j2\pi uk/K}, \qquad (4\text{-}14)$$

for $u = 0,1,2,\ldots,K-1$, where $ffd(u)$ represents the Fourier Descriptors of the facial region shape, which have to be projected into the current facial region subspace similarly to the process of LFC.

4.3.2 Feature Vector Estimation

Same as the appearance feature vector estimation, all FFD vectors of the training set have to be concatenated to form the matrix Φ:

$$\Phi_{Ge} = \left[F_{Ge}{}^{0} - \mu_{Ge} \;\; F_{Ge}{}^{1} - \mu_{Ge} \;\; \ldots \;\; F_{Ge}{}^{P-1} - \mu_{Ge} \right], \tag{4-15}$$

where P is the total number of images used for training and μ_{Ge} is given by:

$$\mu_{Ge} = \frac{1}{P} \sum_{n=0}^{P-1} F_{Ge}(n), \tag{4-16}$$

Next, the eigenvalues of the covariance matrix:

$$\Omega_{Ge} = \Phi_{Ge}{}^{T} \Phi_{Ge}, \tag{4-17}$$

are estimated which has up to P eigenvectors associated with non-zero eigenvalues, where $P < K - 1$. These eigenvectors are then stored in a descendent order according to the corresponding eigenvalues. The sorted eigenvectors are given by:

$$\Psi_{Ge} = \left[V_0^T \;\; V_1^T \;\; \cdots \;\; V_{K-1}^T \right]. \tag{4-18}$$

It is important to mention that rather than the appearance-based method, this process is applied so that 99% of the variance of training Fourier coefficient vectors is retained. Thus, the reduced space feature vector of appearance features Υ_{Ge} is defined by:

$$\Upsilon_{Ge} = \Psi_{Ge}^{T} \left(F_{Ge} - \mu_{Ge} \right), \tag{4-19}$$

where Ψ_{Ap} is the facial region subspace and μ_{Ap} is the mean vector of all training images obtained by using geometric features.

4.4 Hybrid Feature Vector Estimation

The combination of geometric and appearance features has been successfully applied for FER **[3, 4, 11]**. Some approaches perform the fusion at classification level but better results are obtained when the combination is done at the feature extraction phase **[11]**. From the literature it is possible to see that feature extraction methods based on hybrid features are focused on appearance features and improves the final feature vector by adding facial landmarks. In this way, the characteristics of both type of features are different and the final feature vector could be directly affected by the individual problems of one kind of features. Therefore, the fusion process of the proposed method is based on the application of PCA for correlating the information of both types of features. In addition, the proposed feature extraction method employs the same principle based on the DFT. Thus, the appearance features are extracted by LFC and the geometric features by FFD. It is worth noting that in order to efficiently correlate the features, PCA process has to be applied individually before the fusion.

The framework of the complete feature extraction process is shown in **Figure 4.4**. The first step is related to the face detection followed by the individual process of feature vector estimation of appearance- and geometric-based approaches. For appearance features, the region segmentation into three facial regions has to be applied. Next, the LFC is applied for each sub-block size of *L*x*L* pixels, and finally the feature vector is obtained by applying PCA to the whole training set. On the other hand, for geometric features the first step is to localize the *K* facial landmarks which define the three facial region shapes. Subsequently, the FFD are obtained, and finally the PCA is applied same as the appearance-based method. In this way, feature vectors from two different type of features were described using the same representation provided by DFT. Thus, the final step is the fusion of both individual feature vectors by correlating the training information and projected into a common Eigenspace.

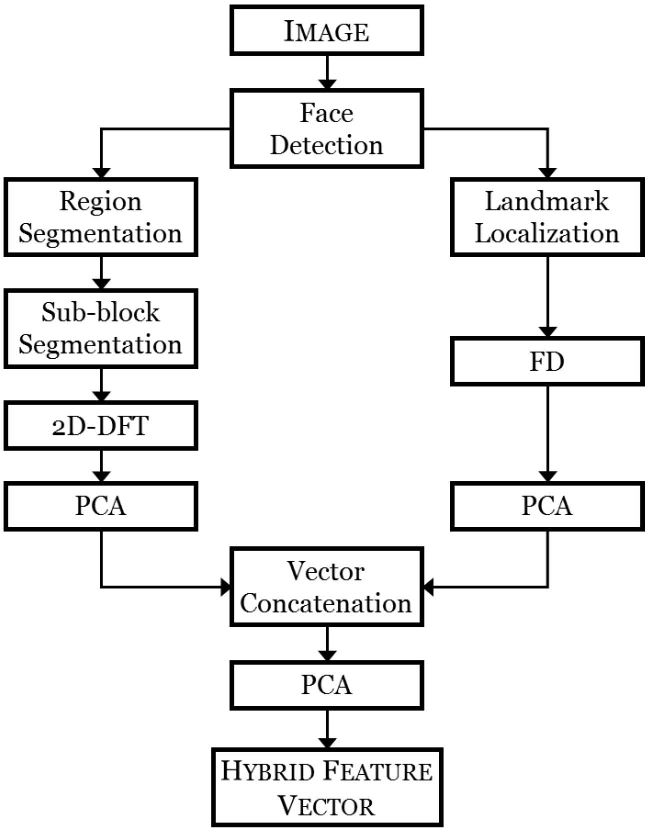


Figure 4.4. Framework of the fusion process of hybrid Fourier features.

After the calculation of individual feature vectors of appearance and geometric features, as defined in **Equations (4-10)** and **(4-19)**, respectively. Thus, the process begins with the concatenation of both feature vectors, so that

$$\mathrm{H} = \left[\mathrm{Y}_{Ge}^{T}, \mathrm{Y}_{Ap}^{T}\right]^{T}. \qquad (4\text{-}20)$$

Subsequently, the PCA process should be applied with these new hybrid vectors, so that:

$$\Phi_H = \left[H^0 - \mu_H \;\; H^1 - \mu_H \;\; \dots \;\; H^{P-1} - \mu_H\right], \qquad (4\text{-}21)$$

where P is the total number of training images and μ_H is given by:

$$\mu_H = \frac{1}{P}\sum_{n=0}^{P-1} H(n), \qquad (4\text{-}22)$$

Next, the eigenvalues of the covariance matrix:

$$\Omega_H = \Phi_H{}^T \Phi_H, \qquad (4\text{-}23)$$

are estimated which has up to P eigenvectors associated with non-zero eigenvalues, where P is smaller than the combined length of vectors F_{Ap} and F_{Ge}. These eigenvectors are then stored in a descendent order according to the corresponding eigenvalues. Sorted vectors are defined by the matrix Ψ_H which is given by:

$$\Psi_H = \left[V_0^T \;\; V_1^T \;\; \cdots \;\; V_{K-1}^T\right]. \qquad (4\text{-}24)$$

Similar the geometric-based method, this process is applied so that 99% of the variance of training vectors is retained. Thus, the final hybrid feature vector of a specific facial region is defined by:

$$\mathrm{Y}_H = \Psi_H^T\left(H - \mu_H\right), \qquad (4\text{-}25)$$

where Y_H can be seen as a projection of the hybrid vector H into the eigenspace Ψ_H of the current facial region, and μ_H refers to the mean hybrid vector of all training images. It is worth noting that Y_H only represents the feature vector of an individual facial region. Therefore, a feature vector based on all possible combinations of the three specific facial regions is defined as:

$$\mathrm{Y} = \bigcup_{l=1}^{FR} \mathrm{Y}_H(l), \;\; FR = 1,2,3, \qquad (4\text{-}26)$$

where Y_H represents an individual hybrid feature vector based on a specific facial region and Y the concatenation of them. Thus, Y can be conformed up to three individual facial regions.

4.4.1 Consideration of Static Structure of Faces

An important problem attainted to FER is the difference in appearance of individual's face, such as texture, color, and shape but must important are the differences in the way to express the facial expressions. These differences are related to frequency of peak expressions, degree of facial plasticity, individual morphology and neutral face state. In general, these characteristics can be refer to the static structure of individual faces.

In order to design a FER system robust to the static structure of individual faces, some works assumed that facial expressions can be represented as a linear combination of expressive and neutral face images of the same subject **[43, 80]**. It is well known that the structural characteristics and texture information which define a specific expression appear when the face images change from neutral to expressive, hence the difference image may represent those changes and it can reduce the dependency on the subject's identity as well. Therefore, difference images may also reduce the physical differences among faces from different races and be focused only in the way to constitute the facial expressions. For that reason, the subtraction of both feature vectors was proposed in contrast of using only expressive images. **Figure 4.5** illustrates the framework of this proposal which has to be applied after the PCA process. Thus, projections of expressive and neutral eigenspaces have to be calculated individually in order to obtain the definitive feature vector by subtracting neutral from expressive individual's projected vectors.

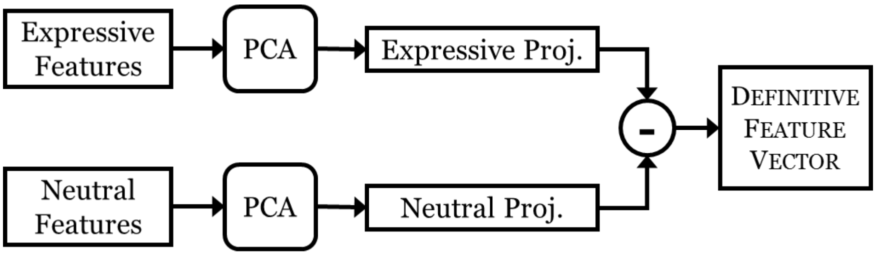


Figure 4.5. Framework of the consideration of static structure of the face for feature extraction process.

For the proposed method, the treatment of the static structure of the face is based on individual facial regions and its combinations as detailed in **Equation (4-26)**. Therefore, the definitive feature vector is defined by:

$$Z(l) = Y_{Exp}(l) - Y_{Neu}(l)\,, \tag{4-27}$$

for $l = 0,1,2,\ldots,Q-1$, where Q is the total number of expressive images in the dataset, Y_{Exp} and Y_{Neu} represents the final feature vectors of expressive and neutral facial image, and Z the difference vector which is the definitive feature vector used in this proposal. It is important to mention that for this process, the information of neutral images must be considered each time the PCA process is applied, so that each matrix Φ defined in **Equations (4-10)**, **(4-19)** and **(4-25)** must consider neutral images in the training set of images.

As a visual example of the effect of mentioned subtraction, **Figure 4.6** shows two difference images of subjects from different gender and races showing the same facial expression. As can be observed, the expressive images in the first column present physical changes based on the static structure of each individual's face. These noticeable differences between images of the same expression may affect the classification process. On the other hand, the difference images illustrated in the third column, present clear similarities based on the facial actions produced by the expression rather than physical differences.

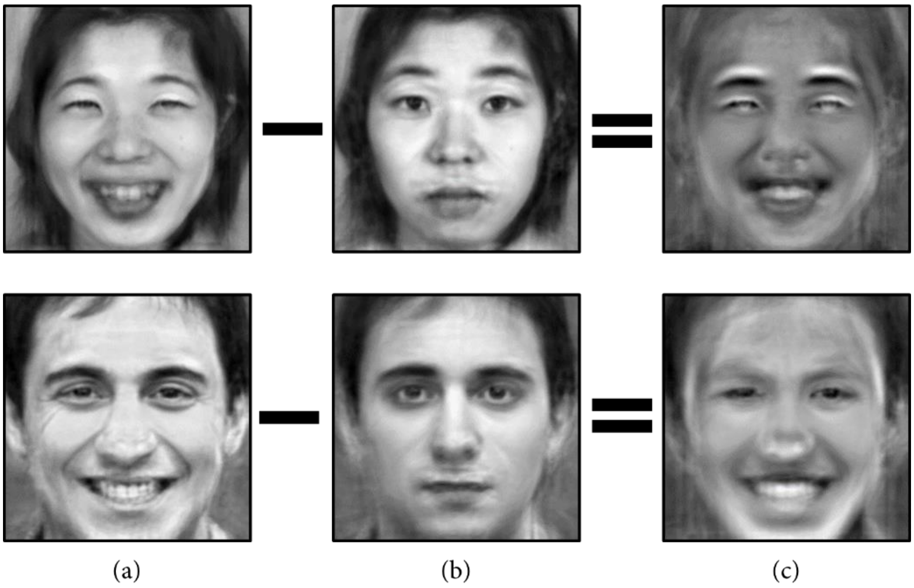


Figure 4.6. Example of two difference images reconstructed from its feature vectors. (a) Expressive images from Y_{Exp} ; (b) Neutral images from Y_{Neu} ; (c) difference images from Z .

4.5 Experimental Results

Similar to the proposed system presented in **Chapter 3**, feature vectors of the six basic expressions were classified by multi-class SVMs with RBF kernels **[72]**. The system was evaluated following a widely used protocol in FER, this is leave-one-subject-out (LOSO) cross-validation. This method consists of dividing the database according to the number of subjects, such as each sub-group consists of only images from the same subject. Then, one of these sub-groups has to be picked out for testing and the remaining are used for training. This procedure has to be repeated the same number of times as the number of subjects in the database. Finally, the recognition accuracy is averaged over all trials. In addition to the average recognition rate of LOSO, confusion matrices are also presented for evaluation results. The diagonal entries of the confusion matrices represent the accuracy of the facial expressions correctly classified, whereas the off-diagonal rates the misclassification problems

4.5.1 Databases

The database employed for the preliminary test is a subset selected from the Extended Cohn-Kanade database (CK+) **[37]**, which includes expressive and neutral faces of 90 subjects. 240 peak expressive frames (40 per basic expression) and 90 neutral frames (from each subject) form the total number of images of this subset. Essentially, this subset was selected taking the same number of samples per expression in order to avoid misinterpretations of the results for the preliminary test.

The fully automated system was evaluated using the complete version of the CK+ database **[37]**, the Multimedia Understanding Group Facial Expression Database (MUG) **[81]**, and the Taiwanese Facial Expression Image Database (TFEID) **[82]**. **Table 4.1** shows the number of subjects and frames per expression of each database, where 362 expressive frames comprise the CK+, 304 the MUG and 229 the TFEID. It is important to mention that for CK+ data set, the number of images from the expressions of fear and sadness was increased by selecting two expressive frames from each sequence (not only peak frames). Thus, the original number of sequences of these expressions are 25 and 28 respectively.

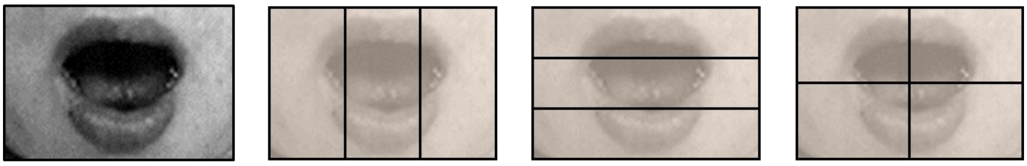
Table 4.1. Number of frames and subject of each database.

Database	*Subjects*	*Ang*	*Disg*	*Fear*	*Happ*	*Sad*	*Surp*
CK+	116	45	59	**50**	69	**56**	83
MUG	52	52	51	48	52	49	52
TFEID	40	34	40	40	40	39	36

4.5.2 Preliminary Results

The results presented in this section includes the analysis of sub-block sizes for LFC, the analysis of the number of fiducial points for FFD, and the classification performance of feature extraction vectors obtained without using the proposals based on DFT.

Thus, feature vectors based on appearance, geometric and hybrid features were obtained using only PCA on the segmented regions presented in **Sections 4.2** and **4.3.** In addition, it is important to mention that based on the results presented in **Section 3.6.1** the forehead region is not included in these tests. This is due the use of the mentioned facial region rather than improve, it reduces the recognition rate of FER systems.



L=M·N *L=N/3* *L=M/3* *L=M·N/2*

Figure 4.7. Examples of sub-block segmentation of non-square regions.

By adopting the analysis bias of **[66]** we test the LFC method with four square sizes (*L*=2, *L*=4, *L*=6 and *L*=12). In addition, three non-square windows are proposed: *L*=M·N/2, *L*=M/3 and *L*=N/3, which represent the segmentation of the facial region into four and three equal size parts (horizontal and vertical possibilities). Finally, the whole input facial region (*L*=M·N) without local segmentation is also tested. **Figure 4.7** illustrates an example of sub-block region segmentation of the described non-square windows applied to the mouth facial region. In summary, the following analysis presents the performance of LFC when eight different sizes of *L* in Equation (4-4) are used for feature vector calculation.

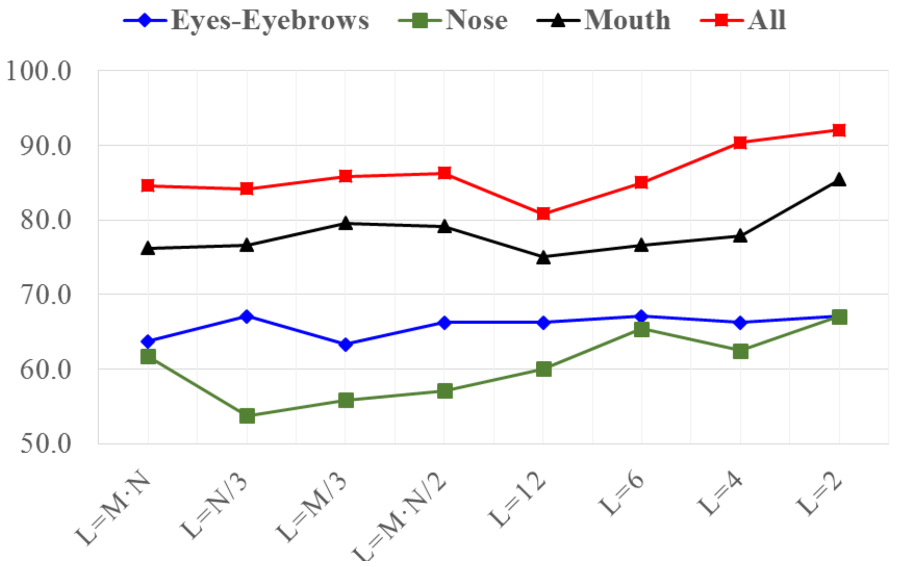


Figure 4.8. Results of LFC with different sub-block sizes using Eyes-Eyebrows, Nose, Mouth and All facial regions for feature extraction process.

The results of the eight different sub-block sizes are shown in **Figure 4.8**. From this graph we can easily see that the best recognition performance is obtained using the combination of all facial regions in the feature extraction process. In addition, the average recognition rate increases when the size of the sub-block decreases. Thus, the best results are obtained by using *L*=2 which represents the minimum square window of just 2x2 pixels. Finally, we can highlight that the best performance of LFC is reached when the sub-block size is equal to 2x2 pixels

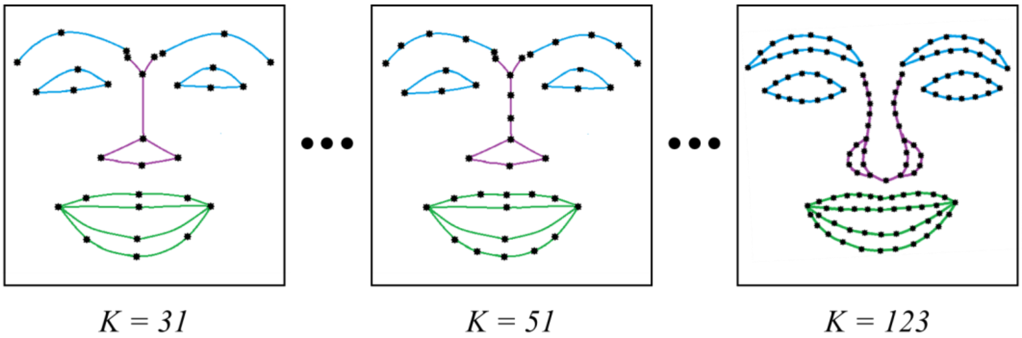


Figure 4.9. Examples of facial shapes represented by different number of fiducial points (*K*).

Choosing the number of landmarks that defines the facial shape is an important issue for every FER system based on geometric features. Therefore, a test for FFD using eight different number of fiducial points (*K*=31, 41, 51, 64, 81, 93, 115 and 123) is presented in this section. The test consists of analyzing FER performance based on different shape representations by changing the number of landmarks used in Equation (4-11). It is important to mention that for this particular test, the landmark estimation was manually annotated for all images of the CK+ subset. The main differences between the eight shape representations reside on the location and the number of facial landmarks of each facial region. For example, for *K*=31 the number of landmarks representing the nose region is 7 whereas for *K*=123 the same region is represented by 29 landmarks. **Figure 4.9** shows three examples of these different shape representations, i.e. *K*=31, *K*=51, and *K*=123.

Results of the eight *K* values for FFD are shown in **Figure 4.10**. This figure presents individual performance of eyes-eyebrows, nose lips and the combination of all of them. As expected, we can see that the results improve when the number of landmarks increases, thus *K*=123 presents the best performance for FFD. However, the improvement is not significant for some tests. For example, when all regions are used for feature extraction (All) the average recognition rates of *K*=51 and *K*=123 are 93.8% and 95.9% respectively, just 2% of improvement. Moreover, even when the nose region presents a remarkable improvement of accuracy, this is not reflected when all the regions are used for the feature extraction. Therefore, we decided to use the number of landmarks provided by **[77]** which conveniently is K=51.

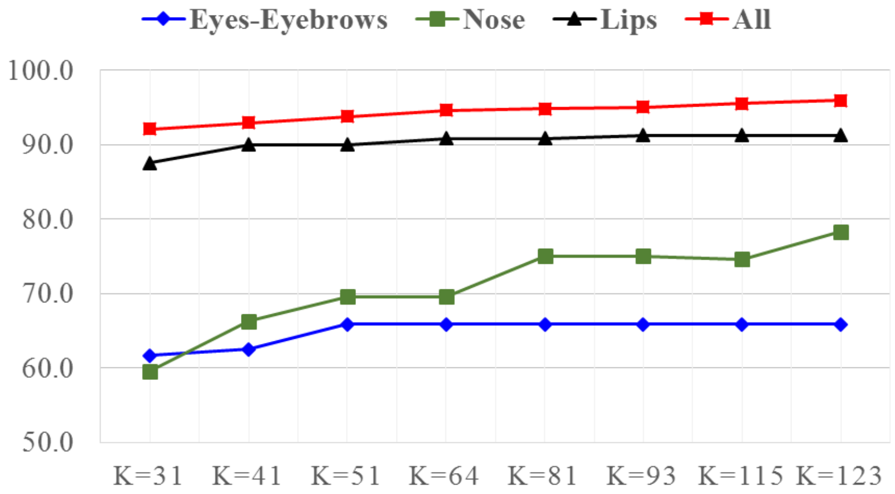


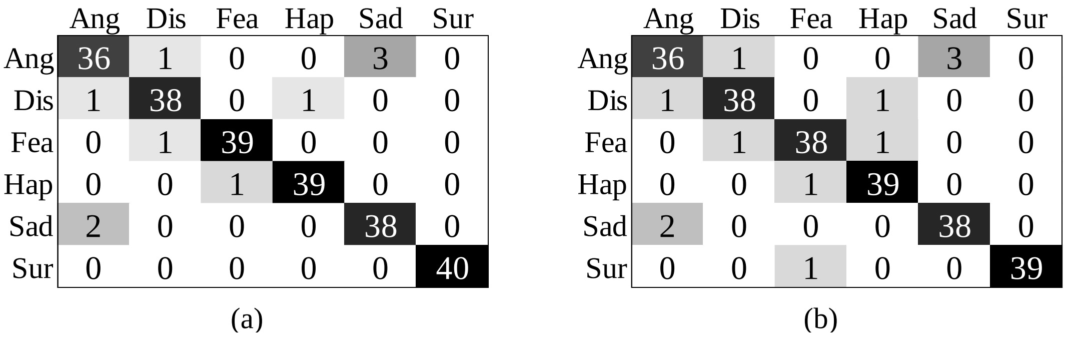
Figure 4.10. Results of FFD with different number of fiducial points using Eyes-Eyebrows, Nose, Mouth and All facial regions for feature extraction process.

Finally, the last test performed with the subset of CK+ is a comparison of the PCA baseline and the DFT-based proposals which results are shown in **Table 4.2**. The results are divided by the type of feature extraction process employed: appearance-, geometric- and hybrid-based. From this Table it can be seen that the proposal of hybrid features using DFT overcomes the average recognition rate of the rest of the tests, including that of PCA using the same hybrid method. However, the improvement is 0.83% which represents just 2 images from the database. Thus, the main improvement of the system is related to the proposed hybrid procedure. It is important to mention that the proposed hybrid method shows better performance than the other two feature-based methods from each combination of facial region.

Table 4.2. Average recognition rate of individual facial regions and its combinations obtained using the proposal and baseline methods.

	Appearance		*Geometric*		*Hybrid*	
Region	***DFT***	***PCA***	***DFT***	***PCA***	***DFT***	***PCA***
Eyes-eyebrows	67.08	66.67	68.33	65.83	71.67	69.17
Nose	67.08	66.25	72.92	72.08	75.83	75.00
Mouth	85.42	82.50	90.00	89.58	***90.83***	90.42
Eyes-Nose	76.25	75.42	73.75	73.33	78.75	77.50
Eyes-Mouth	90.42	90.42	92.08	90.83	***94.17***	93.75
Nose-Mouth	88.75	86.67	90.83	90.42	93.33	92.92
All regions	92.08	91.67	92.50	91.67	***95.83***	95.00

From **Table 4.2** it is possible to notice that the highest performance is reached by the combination of the three facial regions. Following by the combinations of Eyes-Mouth and Nose-Mouth. In addition, the best result provided by using only one facial region is obtained by using the mouth region, reaching more than 90% of accuracy. DFT and PCA approaches have similar recognition rates in almost all of the facial region combinations. However, in order to deeply analyze its performance, it is necessary to know the accuracy reached by each facial expression. To this end, **Figure 4.11** presents the confusion matrices of the proposed and baseline methods using hybrid features. Feature vectors used for obtaining these results are based on the combination of the three facial regions.



	Ang	Dis	Fea	Hap	Sad	Sur
Ang	36	1	0	0	3	0
Dis	1	38	0	1	0	0
Fea	0	1	39	0	0	0
Hap	0	0	1	39	0	0
Sad	2	0	0	0	38	0
Sur	0	0	0	0	0	40

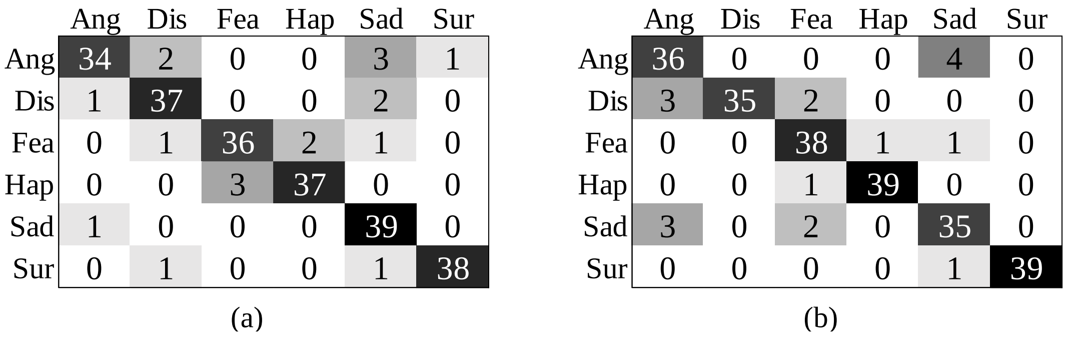
(a)

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	36	1	0	0	3	0
Dis	1	38	0	1	0	0
Fea	0	1	38	1	0	0
Hap	0	0	1	39	0	0
Sad	2	0	0	0	38	0
Sur	0	0	1	0	0	39

(b)

Figure 4.11. Confusion matrices of hybrid features using all regions. (a) DFT proposed method. (b) PCA baseline.

From this **Figure 4.11** it can be noticed that the misrecognition errors are almost identical, except for the extra mistakes of the PCA method, which are related to the expressions of surprise and fear misrecognized with fear and happiness respectively.



	Ang	Dis	Fea	Hap	Sad	Sur
Ang	34	2	0	0	3	1
Dis	1	37	0	0	2	0
Fea	0	1	36	2	1	0
Hap	0	0	3	37	0	0
Sad	1	0	0	0	39	0
Sur	0	1	0	0	1	38

(a)

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	36	0	0	0	4	0
Dis	3	35	2	0	0	0
Fea	0	0	38	1	1	0
Hap	0	0	1	39	0	0
Sad	3	0	2	0	35	0
Sur	0	0	0	0	1	39

(b)

Figure 4.12. Confusion matrices of the proposed method using all regions. (a) Appearance-based features. (b) Geometric-based features.

On the other hand, even the performance of the proposed method using appearance and geometric features are similar too, these results present different recognition accuracy by facial expression. **Figure 4.12** presents the confusion matrices of the appearance- and geometric-based methods which applied the DFT approach on the combination of all regions. Thus, geometric-based method presents better results for surprise, happiness and fear expressions, meanwhile appearance-based for sadness, surprise and disgust.

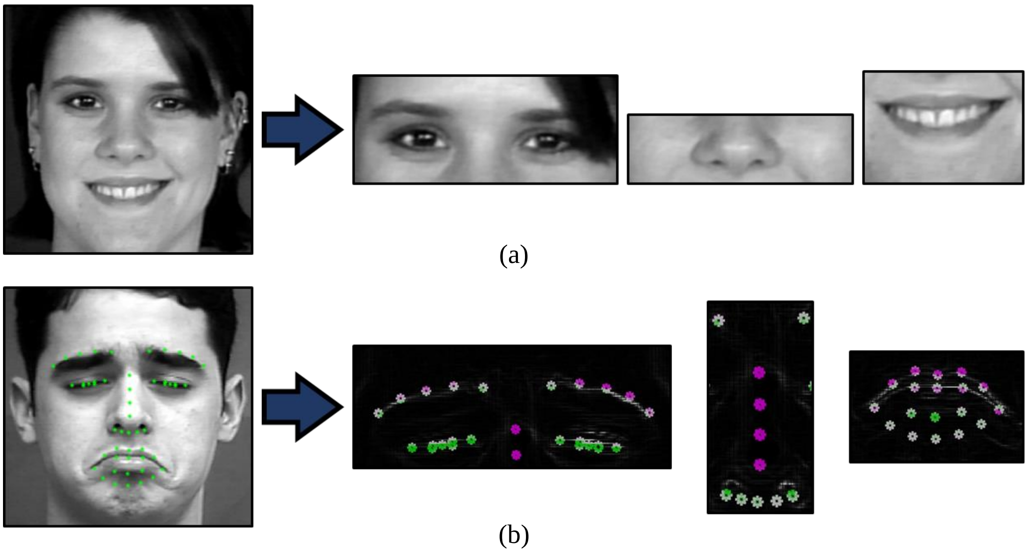


Figure 4.13. Example of mistakes of the proposed method using different kind of features. (a) Appearance-based: face labeled as happiness and misrecognized with fear. (b) Geometric-based: face labeled as sadness and misrecognized with fear.

The differences on the easiness for recognizing certain facial expressions are straight linked to the features used for the description of the facial image. That is why geometric-based method misrecognizes sadness with anger and fear. As well as, appearance-based misrecognizes fear with happiness. In order to appreciate these differences, **Figure 4.13** illustrates some examples of the misclassification errors generated by each feature-based methods.

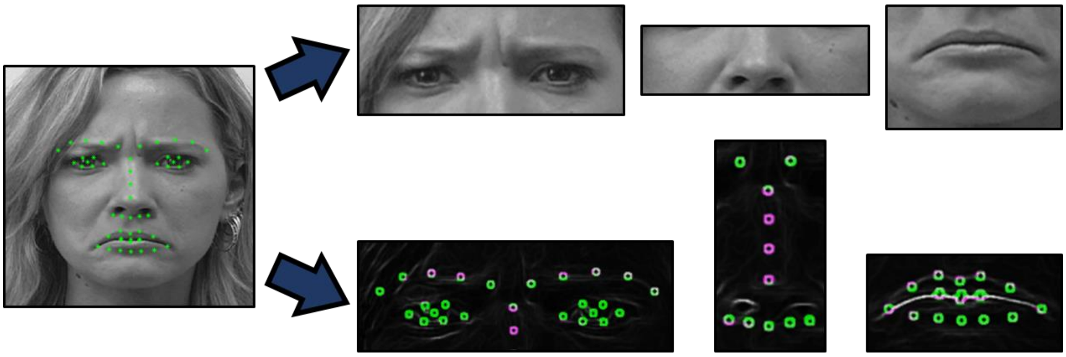


Figure 4.14. Example of a mistake of the proposed method using hybrid features. Face labeled as sadness and misrecognized with anger.

From **Figure 4.13** it is possible to see that the problem of the appearance-based may be associated with the mouth region. This is because the mouth presents certain similitudes with that of fear, i.e. both are open and showing the teeth. On the other hand, the problem of the geometric-based is related with the facial landmark localization of the mouth region. The shape detected by the algorithm shows an open mouth, meanwhile the truth is a clearly closed and sadly shaped mouth. Thus, the misrecognition problem of this particular example is due to the automatic detection instead of the feature extraction process. It is important to mention that the mistakes presented in **Figure 4.13** are exclusive of the mentioned feature-based methods, such that the face from (a) is well recognized by using geometric features and the opposite situation is presented in (b). In addition, **Figure 4.14** illustrates an example of a recurrent misrecognition problem. That sample is misrecognized with anger by the three types of feature-based methods. In this case, the landmarks are well detected but the problem seems to be in the subject's way of showing the sadness expression.

4.5.3 Results with Full Databases

As mentioned before, the fully automated system was evaluated using three standard databases: CK+, MUG and TFEID. **Table 4.3** shows the results of each databases divided by individual facial regions and its combinations.

Table 4.3. Average recognition rate of individual facial regions and its combinations by each database.

Region	*CK+*	*MUG*	*TFEID*
Eyes-eyebrows	78.7	81.1	77.8
Nose	86.2	80.2	74.7
Mouth	87.7	85.7	80.4
Eyes-Nose	89.8	88.5	86.7
Eyes-Mouth	96.4	94.0	93.0
Nose-Mouth	94.0	89.9	88.0
All regions	97.9	95.9	94.9

From **Table 4.3**, we can see that the best performance among all data sets is reached by using all regions for feature extraction (All region). Moreover, the best results using two and one facial regions are based on Eyes-Eyebrows-Mouth and Mouth respectively, for all data sets. Indeed, the performance of using only two facial regions is highly competitive, only approximately 1% of accuracy is decreased compared with "All regions". On the other hand, the results with CK+

presents a wider gap of the average recognition rate (10%) between Mouth and Eyes-Eyebrows regions. Furthermore, the TFEID test presents a significant decrease of performance when less than two regions are used for feature extraction. In other words, it is more difficult to recognize the six basic expressions using only one facial region with TFEID data set. In summary, the best performance reached by our proposal is based on all regions for feature extraction and the mouth seems to be the facial region which can better represent the six basic expressions.

A comparison with other approaches evaluated with same data sets is shown in this section. CK+ is one of the most used data sets for FER, therefore **Table 4.4** presents just some of many approaches which have employed it. From this table, we can see that our proposal overcomes all previous approaches. However, works **[43, 83, 84]** also present an average recognition rate higher than 97%. It is worth noting that two of these approaches used a combination of appearances and geometric features. In general, it can be noticed that the approaches based on both kinds of features reach higher performance. In addition, our proposal also overcomes results obtained by approaches based on Deep Neural Networks **[61, 83]**, which provide semantic features of expressive faces.

Table 4.4 Comparison with different approaches with CK+.

Ref. Year	*Method*	*Classifier*	*Data*	*Features*	*Protocol*	*Accuracy*
[80] ‘14	FPDRC + CARC + SDEP	NN	Image	Both	-	88.70
[85] ‘16	Weighted Feats.	SVM	Image	Geo.	2-fold	93.00
[60] ‘09	Boosted LBP	SVM	Image	App.	10-fold	95.10
[74] ‘11	PCA	LDCRF	Sequence	Geo.	4-fold	95.79
[86] ‘15	DVNP	RF	Sequence	Geo.	10-fold	96.38
[61] ‘17	CNN	LR	Image	App.	8-fold	96.76
[43] ‘14	PCA Dictionary	SRC	Image	App.	LOSO	97.19
[84] ‘16	LBP + NCM	SVM	Image	Both	5-fold	97.25
[83] ‘15	CNN + DNN	Joint F-N	Sequence	Both	10-fold	97.25
Proposed	*LFC + FFD*	*SVM*	*Image*	*Both*	*LOSO*	***97.90***

Tables 4.5 – 4.6 present the comparison of performance of different approaches with MUG and TFEID respectively. In both cases, our proposal obtains the highest recognition accuracy. This occurs, even when some approaches don’t use the complete data set of MUG, like **[87]**, and the process is based on sequence of frames, as in **[88]**. It is worth noting that the TFEID data set presents a bigger challenge for FER because instead of CK+ and MUG, the facial expressions are

shown only by Taiwanese people. Therefore, the face structure and some facial expressions may be affected by cultural differences.

Table 4.5. Comparison with different approaches with MUG.

Ref. Year	*Method*	*Classifier*	*Data*	*Features*	*Protocol*	*Accuracy*
[58] ‘15	Gabor + PCA	NN	Image	App.	2-fold	89.29
[89] ‘16	Landmark Dist.	SVM	Image	Geo.	2-fold	90.50
[87] ‘13	LFDA	kNN	Image	App.	LOSO	95.24
[88] ‘17	Triangle Land.	SVM	Sequence	Geo.	10-fold	95.50
Proposed	*LFC + FFD*	*SVM*	*Image*	*Both*	*LOSO*	***95.85***

Table 4.6. Comparison with different approaches with TFEID.

Ref. Year	*Method*	*Classifier*	*Data*	*Features*	*Protocol*	*Accuracy*
[90] ‘17	Haar Wavelet	LR	Image	App.	10-fold	89.58
[91] ‘14	LBP + MPC	SVM	Image	App.	10-fold	92.54
[92] ‘17	Pyramid Feat.	SVM	Image	App.	LOSO	93.38
[93] ‘15	DSNGE	kNN	Image	App.	LOSO	93.89
Proposed	*LFC + FFD*	*SVM*	*Image*	*Both*	*LOSO*	***94.94***

The last comparison with different approaches is focused on the capability to handle the partial occlusion problem. Methods **[42, 85, 94, 95]** proposed different approaches for solving this problem. Our potential solution consists of excluding the occluded facial region in the feature extraction process. For example, for eyes-eyebrows occlusion, our system only uses the regions of mouth and nose for feature vector estimation. **Tables 4.7** compares the results of methods under the occlusion of a specific facial region. In this situation, our proposal presents competitive results with other approaches. However, those are based on CK data set which is a previous version of CK+ known to be limited in size and lacked of spontaneous and non-exaggerated expression. On the other hand, **Table 4.8** presents an opposite situation, i.e. when only one part of the face is available because of occlusion problems. This extreme case is approached for only a few methods, such as **[94]** and **[85]**. From this table we can see that our proposal presents higher recognition rates for each extreme situation. In addition, it can be noticed that the recognition performance is higher when the mouth is available. Therefore, the most difficult scenario related to partial occlusion is when the mouth region is occluded. In this situation, our system can reach 89.8% of accuracy if eyes-eyebrows and nose regions are available.

Table 4.7. Comparison with different approaches under partial occlusion of specific facial regions.

				Occluded Part (%)		
Ref. Year	***Data Set***	***Method***	***Classifier***	***Eyes***	***Mouth***	***NO***
[94] '14	CK	Eigenphases	SVM	87.7	75.3	92.0
[95] '12	CK	Random Gabor Filters	SVM	90.5	82.9	91.5
[42] '14	CK	Radial Gabor Filters	LDA+kNN	**95.1**	**90.8**	95.3
Proposed	*CK+*	*LFC + FFD*	*SVM*	*94.0*	*89.8*	***97.9***

Table 4.8. Comparison with different approaches which present results with only one facial region.

				One Region Test (%)		
Ref. Year	***Data Set***	***Method***	***Features***	***Eyes***	***Nose***	***Mouth***
[85] '16	CK+	Weighted Feats.	Geo.	41.9	25.5	60.4
[94] '14	CK	Eigenphases	App.	53.3	61.0	79.3
Proposed	*CK+*	*LFC + FFD*	*Both*	***78.7***	***86.2***	***87.7***

4.6 Conclusion

This chapter presented a fully automated FER system based on the combination of two novel feature extraction methods: LFC and FFD, which are focused on appearance and geometric features obtained from individual facial regions of eyes-eyebrows, nose and mouth. Therefore, our proposal is robust to common FER problems such as illumination changes, image rotation and dimensionality reduction. In addition, different than the reviewed state-of-the-art approaches, our proposal could work well even when fiducial points are not accurately detected. This is possible because the appearance feature extraction does not depend on the extraction of geometric features. Thus, this proposal just depends on face and eyes detection, carried out by the robust algorithm of Viola-Jones, which achieved 100% of recognition with all data sets tested. Evaluation results also show that the proposed system can handle problems of partial occlusion without heavily decreasing its accuracy performance.

The best results obtained in this chapter were provided by the combination of all facial regions (eyes-eyebrows, nose and mouth), trend that appears when testing the three standard databases. In general, results obtained with the proposed algorithm overcome most of the previous works. In addition, compared with recently famous methods such as CNN and DNN, our system shows better performance with CK+, MUG and TFEID data sets, reaching 98%, 96% and 95% respectively.

In addition, the use of individual facial regions and its combinations enables to attack the problem of partial occlusion. Finally, analyzing the performance of the feature-based methods presented in this chapter, it can be noted that each kind of features are capable for recognizing specific facial expressions that are difficult for being done for its counterpart. Therefore, as a general conclusion, the fusion of both features is a valuable alternative for improving the recognition performance of FER systems.

CHAPTER V

5. ANALYSIS OF WSN AND ASN FACIAL EXPRESSIONS

This chapter presents a methodical analysis of Western-Caucasian and East-Asian prototypic expressions based on specific facial regions. This analysis is composed by facial expression recognition and visual analysis of expressive images. In addition, a cross-cultural human study applied to 40 subjects is presented as a baseline. Finally, two possible solutions for working with multicultural environments are also proposed in this chapter.

CONTENTS OF THE CHAPTER

5.1 Introduction 70
5.2 Datasets 71
5.3 FER Analysis..72
5.4 Visual Analysis...73
5.4.1 Appearance-Based 74
5.4.2 Geometric-Based 75
5.5 Human Study...77
5.6 Solutions for Multicultural FER...78
5.6.1 Early Ethnicity Detection 78
5.6.2 Consideration of Culture-specific Expressions 81
5.7 Experimental Results..82
5.8 Conclusion..89

5.1 Introduction

As mentioned in **Chapter 1**, Charles Darwin stated that facial expressions are innate and invariant for human beings and some mammals **[1]**. With this basis, many psychologists have agreed on the fact that facial expressions are straight linked with the six basic internal emotional states. This proposal defines the prototypic basic expressions of anger, disgust, fear, happiness, sadness and surprise which are recognized across all different races and cultures **[2]**. However, some cross-cultural studies have questioned and in some degree refuted this assumed cultural universality of facial expressions **[8-10]**. On the other hand, from the viewpoint of the human-computer interaction (HCI), the cultural universality of emotions is taken for granted **[4, 11]**. Therefore, most of the automatic facial expression recognition systems are based on the assumption that facial expressions are the same for all human beings. Besides some recent approaches reach a highly average recognition rate, none of them are considering the cultural specificity that some subjects could present on their facial expressions. Thus, in order to attain a complex HCI, FER systems have to take into account the differences which may appear between facial expressions from different races and cultures.

This chapter presents a methodical analysis of cultural specificity of facial expression recognition based on Western-Caucasian and East-Asian expressive faces using different feature extraction methods. This analysis is focused on in- and out-group performance as well as on specific differences presented for certain facial regions on the six basic expressions of each racial group. The proposed analysis is composed by facial expression recognition and visual analysis of facial expression images selected from four standard databases which are divided in three datasets of different cultural and ethnic regions: Western-Caucasian (WSN), East-Asian (ASN) and multicultural (MUL). As a baseline, we present a human study applied to 40 subjects composed by 20 Westerns and 20 East-Asians. The same datasets employed for the FER algorithms are used as stimulus in the human study.

The FER analysis is conducted by extracting appearance, geometric and hybrid features from expressive faces based on the regions of eyes-eyebrows, mouth, nose, forehead and face outline. To this end, the algorithms of feature extraction proposed in **Chapters 3** and **4** are used in this analysis. It is important to mention that, in order to precisely analyze facial expression differences between two racial groups, the total number of feature points per facial region is larger than that presented in **Section 4.3**, and those were manually obtained from each facial image. Finally, the visual analysis is based on two independent representations:

reconstructed images using PCA as in the Eigenfaces algorithm for appearance features; and caricature faces obtained from individual shapes of facial regions for geometric features. In this way, the facial differences of the six basic expressions that may appear among both cultural groups can be identified.

5.2 Datasets

FER analysis was evaluated using a total of 1,167 facial images. Specifically, 905 expressive faces from 262 subjects, which were selected from five standard datasets. This whole set, from now called multicultural dataset (MUL), was divided into two racial groups: Western-Caucasian dataset (WSN) and East-Asian (ASN) dataset.

WSN dataset is comprised of 552 expressive images from 165 different subjects selected from the extended Cohn-Kanade dataset (CK+) **[37]**, which in turn is composed of 327 facial image sequences from 123 subjects performing the 6 basic emotions plus the neutral state; and the Multimedia Understanding Group Facial Expression Database (MUG) **[81]**, which contains 1,462 sequence of frames from 52 subjects. It is worth noting that only Western-Caucasian subjects were selected from CK+, so that WSN includes 113 Euro-American subjects from CK+ and 52 Caucasians from MUG. **Figure 5.1** shows an example of the six basic expressions of WSN dataset, from left to right corresponds to the expressions of anger, disgust, fear, happiness, sadness and surprise.

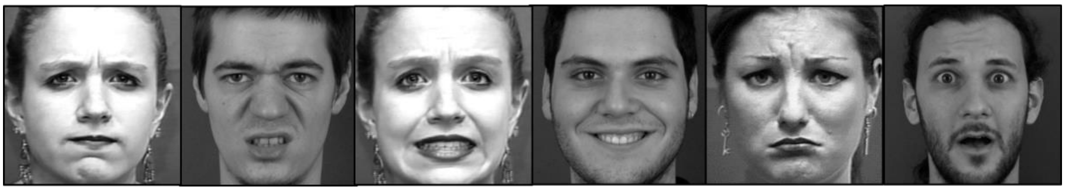


Figure 5.1. Example of images included in WSN dataset.

ASN dataset contains 353 expressive images displayed by 97 subjects selected from three different datasets: Japanese Female Facial Expression (JAFFE) dataset **[36]**, which comprises 213 images of 10 Japanese female models; Japanese and Caucasian Facial Expression of Emotion (JACFEE) dataset **[96]**, which contains 56 images from different individuals including 28 Japanese and 28 Caucasian subjects; and Taiwanese Facial Expression Image Database (TFEID) **[82]**, which includes 336 images from 40 Taiwanese models. In this way, ASN dataset is mainly formed with subjects from Japan and Taiwan. **Figure 5.2** illustrates some faces of the six basic expressions included in this dataset, from left to right corresponds to the expressions of anger, disgust, fear, happiness, sadness and surprise. Not shown are the images selected from JACFFE (Japanese people only) which cannot be reprinted due to copyright restrictions.

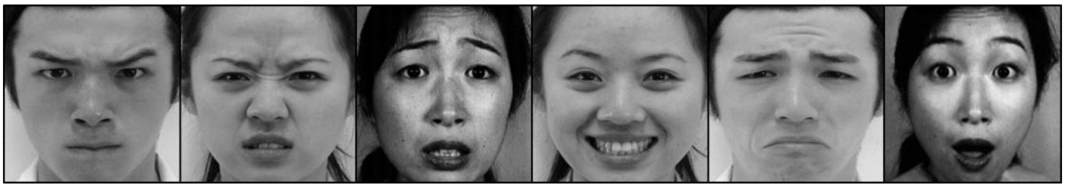


Figure 5.2. Example of images included in ASN dataset.

MUL dataset is a combination of WSN and ASN. Thus the number of images per facial expression is not equitable. **Table 5.1** presents detailed information about the number of images from each cultural dataset, where we can see that the minimum number of images per expression is 117, more than common multicultural works from the literature.

Table 5.1. Number of frames and subject of each cultural dataset.

Database	*Subjects*	*Ang*	*Disg*	*Fear*	*Happ*	*Sad*	*Surp*
WSN	165	84	93	68	116	64	127
ASN	97	57	65	57	69	53	52
MUL	262	141	158	125	185	117	179

Finally, the visual analysis and the human study were evaluated using sub-sets of the WSN and ASN datasets. Therefore, the number of images per facial expression is equitable among each cultural dataset, being 40 images per expression so that 240 expressive images correspond to WSN and ASN sub-sets respectively. It is worth noting that neutral images are not required for these analyses, thus a total of 480 expressive images were employed.

5.3 FER Analysis

The FER analysis is based on a conventional FER framework tested with different feature extraction methods and classified applying some particular cross-cultural modalities which enables a deep analysis of the cultural-specific differences on recognizing the six basic expressions. Basically, three different feature extraction methods were used in this analysis (next section describes them). Besides, classification stage was independently performed by SVM based on the cross-cultural recognition modalities of: in-group, out-group, multicultural and out-group multicultural.

In-group classification represents the FER performance when the same cultural-specific dataset is used for training and testing. The counterpart situation is presented by the out-group classification, because this modality performs the training phase whit a different dataset used for testing. Multicultural classification occur when the training and testing is conducted using a dataset comprised with a variety of cultures. Finally, out-group multicultural classification take place when the system is trained with a multicultural dataset but it is tested using a cultural-specific dataset.

Three feature extraction methods built on DFT are presented in this analysis: appearance-, geometric- and hybrid-based. These methods are based on facial region segmentation and facial landmark localization same as described in **Chapters 4**. It is worth noting that the hybrid-based method combines individual feature vectors from appearance and geometric features, and each facial part is fused with the same facial region of its counterpart kind of features. Finally, all of the feature extraction methods includes the consideration of the static structure of the face as reviewed in **Section 4.4.1**.

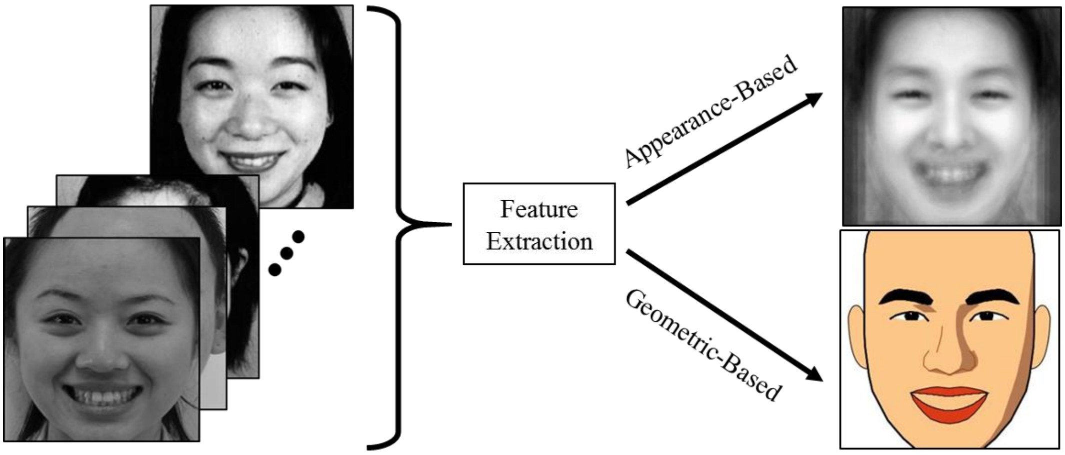


Figure 5.3. Average projected image of happiness from ASN dataset represented by using appearance and geometric features.

5.4 Visual Analysis

In order to make a robust analysis of facial expressions, visual changes presented in the face must be taken into account. Therefore, two methods for visually analyzing the differences which may appear among both racial groups are presented in this section. Those methods are based on the type of features employed for the analysis: appearance- and geometric-based. **Figure 5.3** shows an example of the visual representation of the average happy expression from both feature-based approaches using the ASN dataset.

5.4.1 Appearance-Based

The visual representation of appearance features is based on the well-known algorithm of Eigenfaces. The ability of this method for reconstructing images from projected vectors of a previously defined Eigenspace has been successfully applied for analyzing facial expressions **[97]**. Therefore, using reconstructed images from feature vectors gives the opportunity to analyze the differences and similarities which may appear among the basic expressions of different cultures in detail. To this end, average projected vectors by each expression have to be obtained, which can be calculated from a cultural-specific or multicultural dataset. These average vectors are given by:

$$\mathrm{Z}_{\mu}(r) = \frac{1}{P(r)} \sum_{i=0}^{P(r)} \mathrm{Z}(i) \quad , \tag{5-1}$$

for $r = 0,1,2,\ldots,5$ which represents the number of basic expressions and *P*(*r*) the number of images per expression, so that $\mathrm{Q} = P \times r$. As mentioned before, in order to have a better analysis of facial expressions the number of frames per expression should be equal, then $P(r) = \mathrm{Q}/6$. Finally, reconstructed images are the reshaped matrix of reconstructed average projected vectors given by:

$$\mathrm{R}(r) = \mathrm{W}\,\mathrm{Z}_{\mu}(r) + \mu \quad , \tag{5-2}$$

for $r = 0,1,2,\ldots,5$, where W is the subspace where the Z projections were made, and μ is the mean feature vector of all training images.

Figure 5.4 and **Figure 5.5** shows the visual representation of the average expressions of both datasets obtained from appearance. The expressions represented by each row from left to right are: anger, disgust, fear, happiness, sadness and surprise. It is easy to notice the same distinctive patterns to differentiate some specific expressions. For instance, disgust expression look different. Disgusted faces from WSN fulfill the necessary AUs to be classified as disgust (AU9, AU15, AU16). However, the same average face from ASN presents an extra AU22 and AU23 which are known to appear in anger expression.

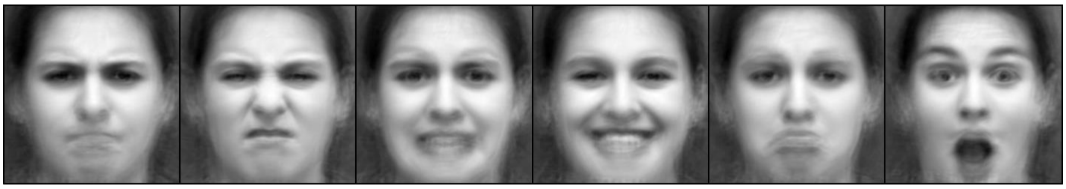


Figure 5.4. Visual representation of average expressions from WSN datasets using appearance features.

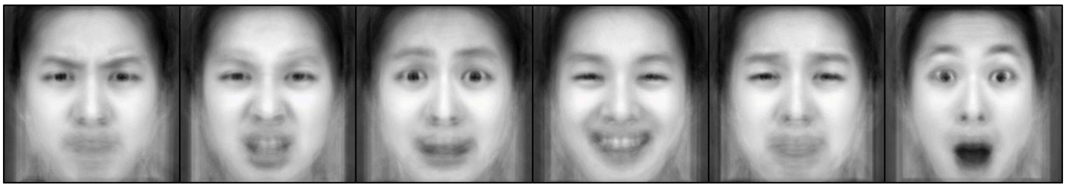


Figure 5.5. Visual representation of average expressions from ASN datasets using appearance features.

5.4.2 Geometric-Based

In order to precisely analyze the facial expression among both cultures, extremely attention has to be put on the landmark localization. Due to the subtle differences among specific facial regions, this process is crucial for the accuracy of feature extraction settling the analysis efficiency. Therefore, facial landmarks for this visual analysis were manually obtained. Thus, a total of 163 feature points for defining the whole facial shape was obtained by using the FaceFit software **[98]**, which is a GUI-based system based on a manual operation to pick up each feature point from the face. **Figure 5.6** shows an example of the wire frame obtained by FaceFit and the 163 feature points extracted to form the whole facial shape.

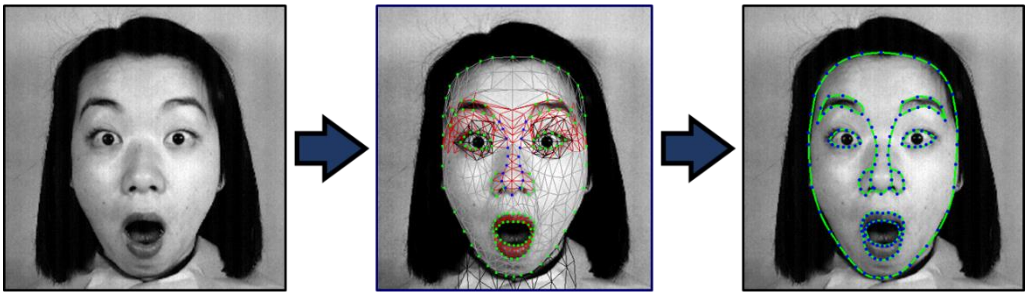


Figure 5.6. Example of the facial landmark localization using FaceFit software.

The raw data of geometric-based methods are easier to visualize than those of appearance-based. However, their projections made by PCA are more difficult to be visually analyzed. Therefore, in order to accurately analyze the geometric features, the DrawFace tool **[99]** was employed. This tool draws caricatures based on the Eigenfaces process, hence this tool requires to input individual eigenspaces of facial regions and the mean facial shape as initialization setup.

Figure 5.7 presents the general framework of DrawFace tool. Similar to the PCA process, the mean face of the complete dataset has to be subtracted from the input, but the eigenspaces are calculated individually for each facial region and the placement of them. Thus, the final caricature is drawn by integrating the projections of the input features into its respective eigenspace. In this way, these caricatures can be considered as a result of a projected feature vector into a subspace made by a set of specific facial shapes.

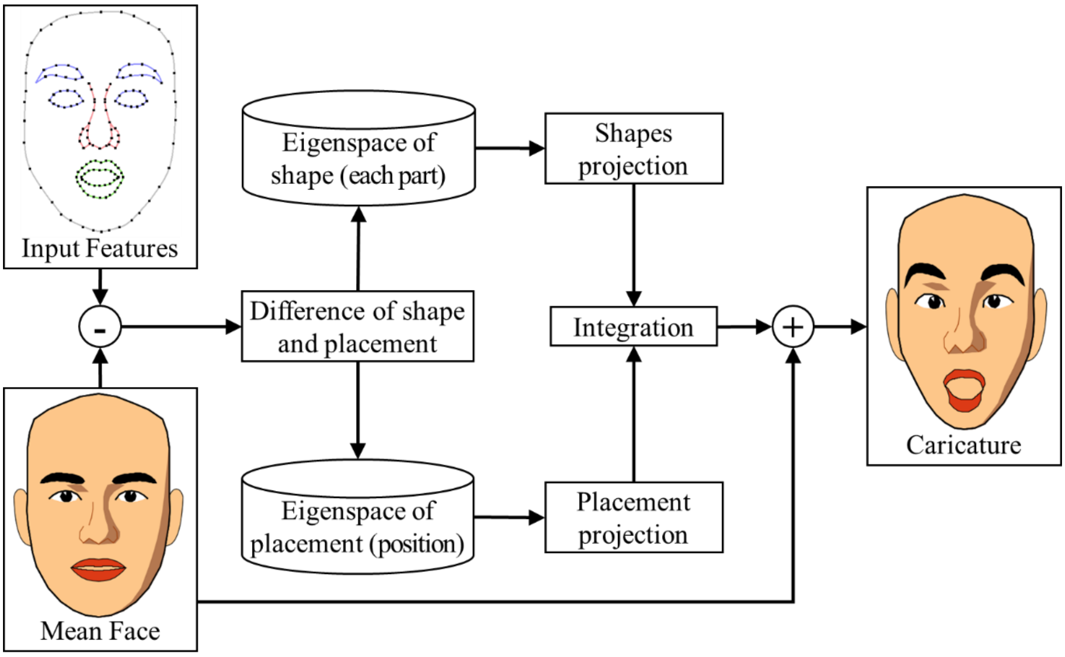

Input Features
-
Eigenspace of shape (each part)
Shapes projection
Difference of shape and placement
Integration
+
Caricature
Mean Face
Eigenspace of placement (position)
Placement projection

Figure 5.7. Framework of DrawFace tool [99] for developing facial caricatures.

Figure 5.8 and **Figure 5.9** shows the visual representation of the average expressions of both datasets obtained from geometric features. The expressions represented by each row from left to right are: anger, disgust, fear, happiness, sadness and surprise. It is easy to notice some distinctive patterns to differentiate specific expressions. For instance, disgust and fear expressions look different. The differences of disgusted faces present the same patterns found in appearance-based visualization described in the previous section. However, the average face of fear from WSN again covers the EMFACS for fear expression (AU1, AUN2, AU4, AU5, AU7, AU20, AU26). However, that of ASN lacks AU4 and AU20, given the impression of surprised in the eyes region and the mouth is not well defined.

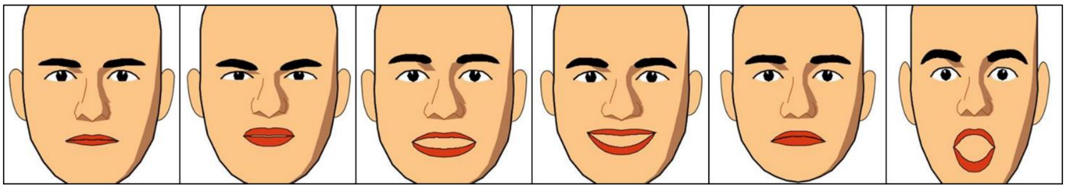


Figure 5.8. Visual representation of average expressions from WSN datasets using geometric features.

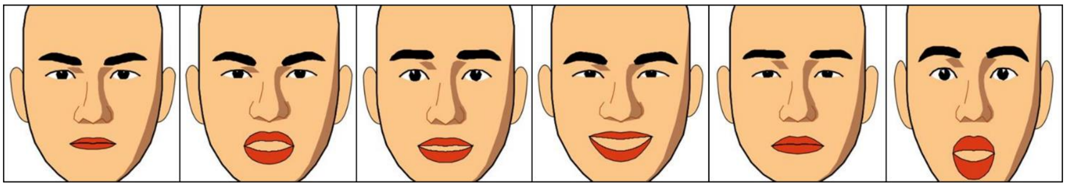


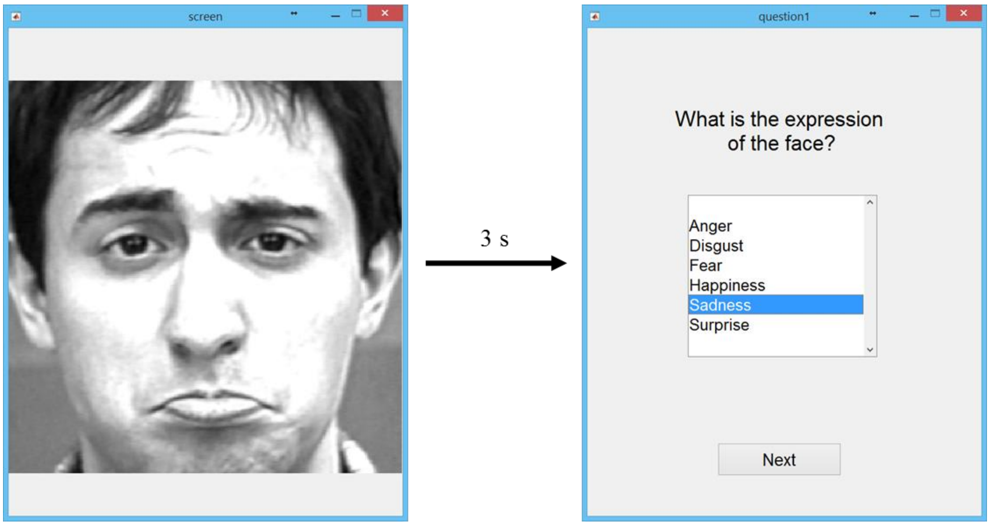
Figure 5.9. Visual representation of average expressions from ASN datasets using geometric features.

5.5 Human Study

As a baseline for the experimental results, a cross-cultural human study based on a survey applied to subjects from a different race and culture is presented. Forced-choice facial expression classification from each participant was collected by using the same datasets employed for the proposed experiments as stimuli. Relevant information about this study is presented as follows.

- ***Participants***. The experiment was applied to 40 students of the University of Electro-Communications in Tokyo Japan. Participants were divided into two groups: 20 East-Asians that include Japanese, Taiwanese and Chinese students (50% males); and 20 Western-Caucasians that include German, Swedish, American and Mexican students (50% males). Their ages ranged from 20 to 26 years old (mean 22). It is important to mention that the Western-Caucasians are currently exchange students and had newly arrived in an Asian country for the first time with a residence time no longer than 3 months on average
- ***Stimuli.*** The expressive faces of the WSN and ASN datasets presented in **Section 5.2** were used as stimuli for all of the participants. Thus, the complete stimuli set comprises 480 images displaying expressive faces of Western-Caucasian and East-Asian people.
- ***Procedure.*** A GUI-based script for collecting and presenting the survey was developed in MATLAB. After a brief explanation of the experiment by the software, the stimuli were automatically presented one by one. The instructions of the experiment were presented in Japanese and Chinese for East-Asian participants and in English for Westerns. Each stimulus appeared in the central visual field and

remained visible just for 3 seconds, followed by a 6-way forced choice decision question related to the 6 basic expressions. The question presented was “What is the expression of the face?” and the participant has to choose one answer before clicking the button “Next” for the following stimulus. We randomized trials within each participant and all of them have to recognize the 240 stimuli of each dataset, which have been presented by groups of 30 images with breaks of 30 seconds between them. **Figure 5.10** presents two screenshots of the application software used for this study.



What is the expression of the face?

Figure 5.10. Screenshots of the software application used for the human study.

5.6 Solutions for Multicultural FER

Even after finding cultural differences on recognizing and expressing some basic facial expressions, there are still some applications that have to work under multicultural environments. Therefore, taking into account the differences previously found, in this section, two possible solutions for working with this special situation are proposed. Specifically, trying to recognize the six basic facial expressions of emotion when Western-Caucasian and East-Asian subjects are involved in the process.

5.6.1 Early Ethnicity Detection

The first proposal builds its process on a logical solution, an early ethnicity detection. For this solution, the categorization of Western-Caucasian (WSN) and East-Asian (ASN) subjects is proposed as preprocessing stage. The general framework of FER employing this proposal is shown in **Figure 5.11**, highlighted you can see the steps which differ from traditional FER frameworks. In this process, culture-specific models obtained from WSN and ASN datasets are required for predicting the expression of a new face image, which has to be previously classified into one of the two possible racial groups, thus the decision is made based on the detected ethnicity of the input subject.

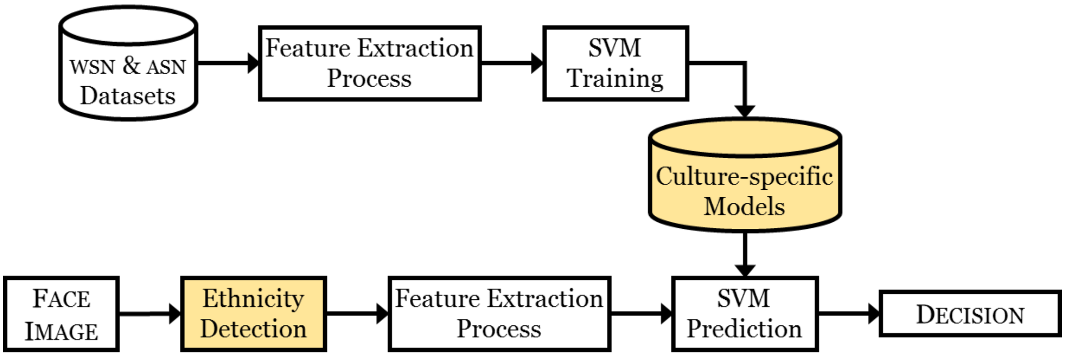


Figure 5.11. FER framework using an early ethnicity detection.

The ethnicity detection process is based on three different features obtained from the detected face of the input image, which are defined as color, texture and shape features. LFC and FFD processes are employed for extracting the texture and shape features, methods detailed in **Section 4.2.1 and 4.3.1** respectively. On the other hand, the color feature extraction is based on a proposed modification of the Dominant Color Correlogram Descriptor (DCCD), method which is used for content-based image retrieval **[100]**. After obtaining the different features, these are combined and classified by PCA and SVM respectively. The complete process of this proposal is shown in **Figure 5.12**.

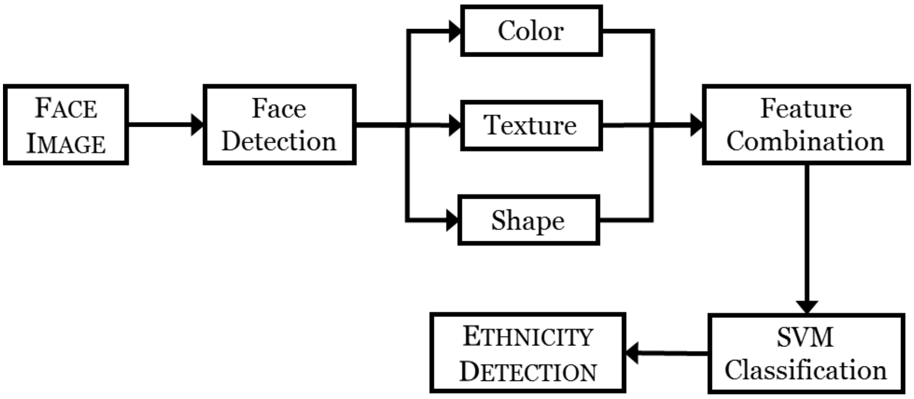


Figure 5.12. Ethnicity detection process.

For the color feature extraction process, the facial region should be first converted from RGB to HSV color space, subsequently the HSV is quantified so that only 72 different skin colors are considered. Finally, the histogram calculated from the skin colors of the face image is taken as color feature vector.

The HSV quantization is based on the fact that all possible skin tones of the different human races are part of a sub-set color from elements of HSV **[101]**. In addition, many studies have found that the skin tone related to the hue component of HSV falls into to sub-group of H that 300° ≤ H ≤ 60° **[102-104]**. Therefore, non-interval quantization of the HSV color space is employed, so that H components are divided into eight shares, taking into account only the sub-group which includes the skin colors. Thus,

$$H = \begin{cases} 0, & \text{if} \quad 280 < h \le 300 \\ 1, & \text{if} \quad 300 < h \le 320 \\ 2, & \text{if} \quad 320 < h \le 340 \\ 3, & \text{if} \quad 340 < h \le 360 \\ 4, & \text{if} \quad 0 < h \le 20 \\ 5, & \text{if} \quad 20 < h \le 40 \\ 6, & \text{if} \quad 40 < h \le 60 \\ 7, & \text{if} \quad 60 < h \le 80 \end{cases}, \tag{5-3}$$

where *h* is the value of hue component of certain pixel of the face image, and *H* is the new quantized value. Subsequently, the S and V components are divided into three shares respectively, as given by

$$S = \begin{cases} 0, & \text{if} \quad 0 \le s \le 1/3 \\ 1, & \text{if} \quad 1/3 < s \le 2/3 \\ 2, & \text{if} \quad 2/3 < s \le 1 \end{cases}, \quad V = \begin{cases} 0, & \text{if} \quad 0 \le v \le 1/3 \\ 1, & \text{if} \quad 1/3 < v \le 2/3 \\ 2, & \text{if} \quad 2/3 < v \le 1 \end{cases}, \tag{5-4}$$

where *s* and *v* represent values of saturation and “value” (from HSV hexcone model) respectively, whereas *S* and *V* their quantized values. The final step of the quantization is to obtain the combination of the three individual values, which is defined by:

$$C = 9H + 3S + V, \tag{5-5}$$

where *C* represents one of the 72 possible colors of human skin.

Finally, from the new matrix obtained by all pixels evaluated with **Equation (5-5)**, the histogram which calculates the most recurrent colors from the face image is used as color feature vector.

5.6.2 Consideration of Culture-specific Expressions

Based on the two specific facial expressions that have found to be different between Western-Caucasians and East-Asians (see **Section 5.4**), the second proposal is focused on the different ways to training the system. Thus, the training step takes special attention on the expressions of disgust and fear. It is worth noting that these expressions not only show visual differences but also present difficulties for being accurately recognized by observers from out-cultural-groups (in-group advantage, **Section 2.1**). **Figure 5.13** shows the proposed FER framework, where WSN and ASN datasets are needed in order to apply special training methods for considering the cultural-specific differences of disgust and fear. In this way, the multicultural models obtained after the training dismiss the need of information about the ethnicity of the input face.

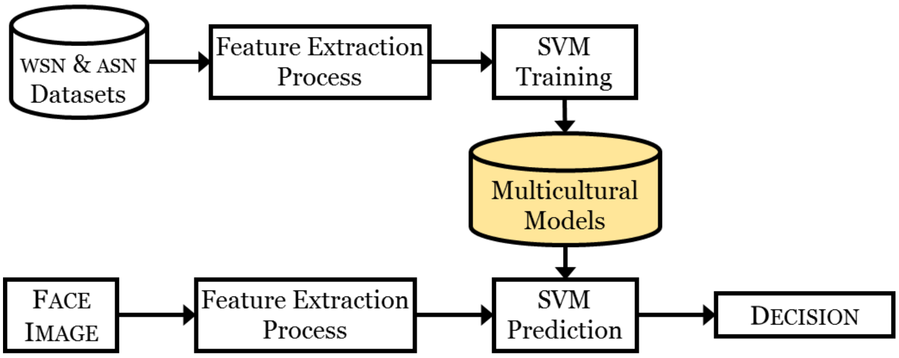


Figure 5.13. FER framework based on multicultural training for taking into account the differences on specific facial expressions.

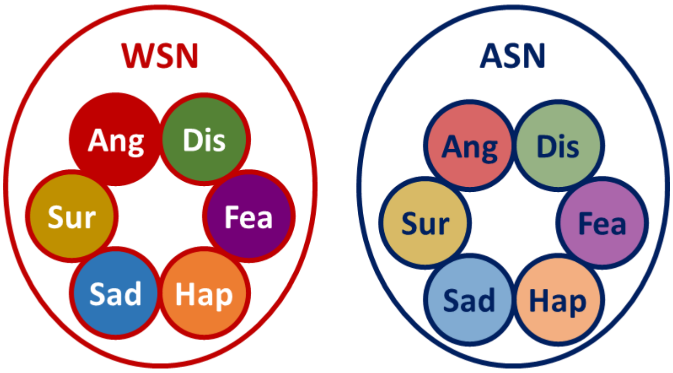


Figure 5.14. Two training methods based on cultural-specific groups (WSN and ASN).

For instance, consider the two possible training methods for working with individual cultural groups such as WSN and ASN (**Figure 5.14**). In each case, the models related to the six basic facial expressions are based only on expressive images from their respective cultural groups. Thus, two different training rounds have to be done for classifying culture-specific expressions of WSN and ASN. In other words, with this training method, only culture-specific models are obtained. Therefore, information about the ethnicity of the input subject is required in order to have an accurately multicultural recognition.

An alternative to have just one training round is proposed in this section. Consider the three different diagrams illustrated in **Figure 5.15**, each of them presents a possible solution that covers the main cultural differences of disgust and fear expressions. For instance, test A obtains multicultural models for anger, happiness, sadness and surprise (merging expressive faces from WSN and ASN), whereas specific-cultural sub-models are trained for disgust and fear. In other words, the multicultural models of each of these two facial expressions are calculated by dividing expressive faces form WSN and ASN. It is worth noting that this division is made only in the training. Thus, the classification stage just considers six multicultural models related to each basic facial expression of emotion. On the other hand, test B and C follow the same protocol but employing only one specific-cultural sub-model, either related to disgust or fear. Results from each of these three tests are presented in the next section.

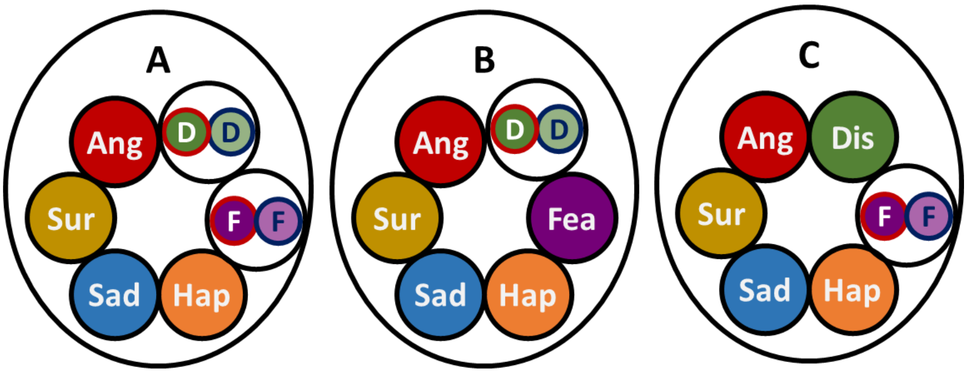


Figure 5.15. Three possible training options for working with multicultural databases.

5.7 Experimental Results

The six basic expressions of all feature extraction methods and cross-cultural classification modalities were classified by multi-class SVMs with RBF kernels **[72]**, and evaluated by leave one subject out cross-validation. Average recognition rates and confusion matrices are presented to show the accuracy of each trial. It is important to mention that all facial images were pre-processed in order to have the same inter-ocular distance and eye position as well as cropped with 280x280 pixels. For the appearance-based approach, facial regions of forehead, eyes-eyebrows, mouth and nose their sizes were normalized at 100x70, 200x80, 140x80 and 175x50 pixels respectively. In addition, the results obtained by the human study are also presented in this section.

Table 5.2. Average recognition rate of individual facial regions and its combinations obtained by each feature extraction method using sub-sets of WSN and ASN.

Feature-based:	*Appearance*		*Geometric*		*Hybrid*	
Dataset:	*WSN*	*ASN*	*WSN*	*ASN*	*WSN*	*ASN*
Eyes-eyebrows	67.1	66.7	65.8	71.7	75.0	71.3
Nose	67.1	50.8	78.3	58.8	81.7	63.8
Mouth	85.4	72.5	91.3	81.3	95.4	84.2
Eyes-Nose	76.3	70.0	76.3	79.2	81.7	80.4
Eyes-Mouth	90.4	89.6	95.0	92.9	98.3	94.6
Nose-Mouth	88.8	78.8	94.6	87.1	94.6	86.3
All regions	92.1	90.8	97.9	94.2	***98.8***	95.0

In general, in-group analysis represents the performance of FER when the people from the same race attempt to recognize facial expressions from their own racial group. In terms of FER systems, this happens when the training and testing sets correspond to the same dataset. **Table 5.2** shows the results of different facial regions using the three feature extraction methods divided by the dataset employed for the evaluation. *FrOt* refers to the forehead region, the outline shape or the fusion of both as it may apply. From this Table it can be noticed that the best results are provided by the proposed hybrid method. In addition, the WSN test reaches higher accuracy than the ASN for all feature extraction methods and facial regions.

Since the best performance was achieved by employing hybrid features, the rest of the analysis is made using only this feature extraction method. Thus, **Table 5.3** shows all results obtained by this proposed method and the human study. The results of in-group, out-group, multicultural and out-group multicultural are presented from left to right in this order.

From **Table 5.3** a trend of the results can be observed. This trend follows the order from top to bottom the accuracy obtained by the tests of in-group, out-multicultural, multicultural and out-group. Thus, the best result is reached by the in-group test of WSN (WSNvsWSN), followed by the out-multicultural test of the same dataset (MULvsWSN) and the third place is for the multicultural test (MULvsMUL). This trend is also followed by the human study, showing that the WSN dataset is easier for classifying the prototypic expressions.

Table 5.3. Average recognition rate of individual facial regions and its combinations obtained using the proposal and baseline methods.

Training Dataset: *Testing Dataset:*	*WSN* *WSN*	*ASN* *ASN*	*ASN* *WSN*	*WSN* *ASN*	*MUL* *MUL*	*MUL* *WSN*	*MUL* *ASN*
Eyes-eyebrows	75.0	71.3	54.2	50.4	68.5	68.3	68.8
Nose	81.7	63.8	46.7	60.8	72.1	80.0	64.2
Mouth	95.4	84.2	68.3	78.8	89.4	94.6	84.2
Eyes-Nose	81.7	80.4	54.6	55.8	80.2	82.9	77.5
Eyes-Mouth	98.3	94.6	**72.5**	82.5	94.4	**95.8**	92.9
Nose-Mouth	94.6	86.3	68.3	81.7	92.1	94.6	89.6
All regions	**98.8**	**95.0**	72.1	**85.0**	**95.0**	95.4	**94.6**
Human Study	76.8	71.7	66.7	67.2	-	-	-

The analysis per facial region combinations shows that the combination of E-N-M performs better in most of the training combinations. On the other hand, the best single region for FER is the mouth, followed by the nose. However, the eyes region present interesting results, especially for the out-group and out-multicultural analysis, where this region seems to define the expressions of ASN better. In order to analyze the performance per expression of those results,

Table 5.4 and **Table 5.5** show the accuracy of the mouth region and E-N-M combination. In these tables, we can notice that the mentioned trend is not followed by all the expressions. For example, the out-multicultural test of the mouth shows better accuracy when anger is tested using ASN dataset rather than WSN. In summary, it is confirmed that the clearer difference between cultures is presented in the out-group test.

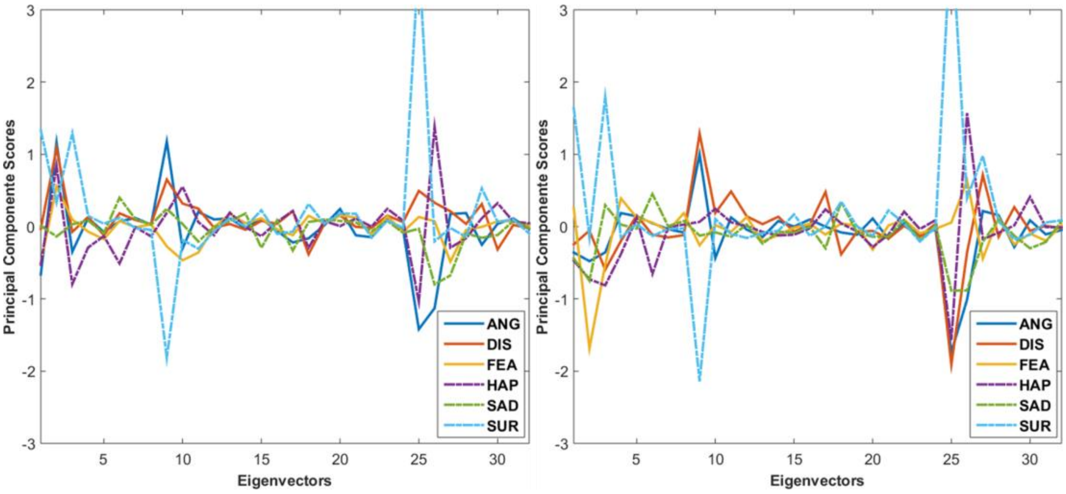
Table 5.4. Classification accuracy per expression using the region of Mouth.

Training Dataset: *Testing Dataset:*	*WSN* *WSN*	*ASN* *ASN*	*ASN* *WSN*	*WSN* *ASN*	*MUL* *MUL*	*MUL* *WSN*	*MUL* *ASN*
Anger	95.0	90.0	85.0	72.5	92.5	95.0	90.0
Disgust	90.0	72.5	32.5	60.0	81.3	90.0	72.5
Fear	95.0	72.5	80.0	72.5	87.5	97.5	77.5
Happiness	95.0	85.0	70.0	92.5	88.8	95.0	82.5
Sadness	100.0	87.5	60.0	85.0	88.8	92.5	85.0
Surprise	97.5	97.5	82.5	90.0	97.5	97.5	97.5
Average	95.4	84.2	68.3	78.8	89.4	94.6	84.2

Table 5.5. Classification accuracy per expression using the "all" combination.

Training Dataset:	*WSN*	*ASN*	*ASN*	*WSN*	*MUL*	*MUL*	*MUL*
Testing Dataset:	*WSN*	*ASN*	*WSN*	*ASN*	*MUL*	*WSN*	*ASN*
Anger	100.0	92.5	82.5	97.5	93.8	95.0	92.5
Disgust	97.5	92.5	40.0	87.5	92.5	92.5	92.5
Fear	95.0	92.5	72.5	50.0	93.8	90.0	97.5
Happiness	100.0	97.5	60.0	100.0	92.5	100.0	85.0
Sadness	100.0	97.5	90.0	80.0	97.5	95.0	100.0
Surprise	100.0	97.5	87.5	95.0	100.0	100.0	100.0
Average	98.8	95.0	72.1	85.0	95.0	95.4	94.6

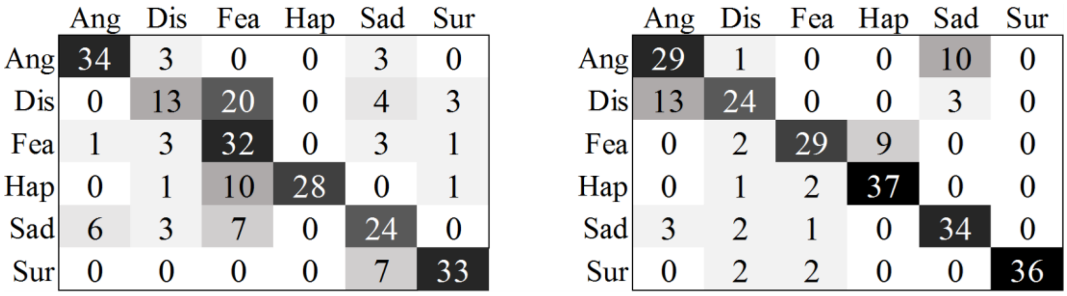
Another way to analyze the distinctive properties of average projected vectors is simply by plotting them. In this way, it can be visualized the behavior of each feature vector and the capability of discrimination of the six basic expressions based on them. **Figure 5.16** shows the six average expression vectors of each dataset. It is easy to see that the six vectors of WSN present better distinctiveness among themselves rather than those of ASN, which have problems especially for the expressions of fear, disgust and anger.



(a) (b)

Figure 5.16. Average projected vectors of six expressions, (a) from ASN and (b) from WSN.

Thanks to the visual analysis, the misrecognition problems that appear among the six basic expressions can be deeply studied. **Figure 5.17** and **Figure 5.18** show the confusion matrices for WSN and ASN of Mouth and Eyes region, respectively. Note that these matrices present a gray scale color map for simplifying the visualization, showing higher results in black and lower in white. The confusion matrices represent the true expression on the vertical axis and the decision made by the classifier is presented on the horizontal axis so that each row of the matrix indicates the level of confusion of each expression with its counterparts.



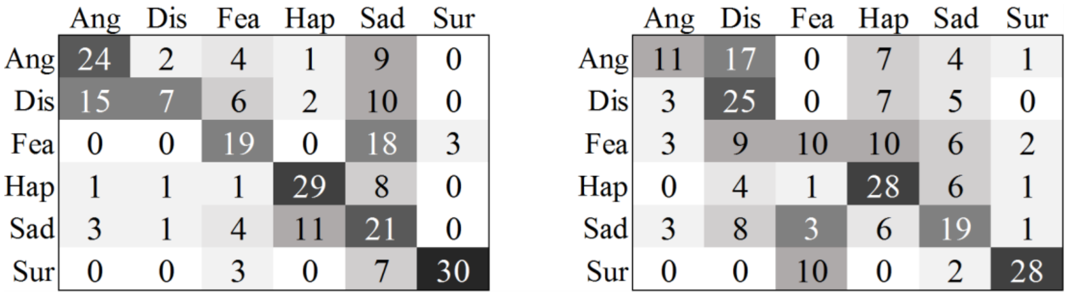
	Ang	Dis	Fea	Hap	Sad	Sur
Ang	34	3	0	0	3	0
Dis	0	13	20	0	4	3
Fea	1	3	32	0	3	1
Hap	0	1	10	28	0	1
Sad	6	3	7	0	24	0
Sur	0	0	0	0	7	33

(a)

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	29	1	0	0	10	0
Dis	13	24	0	0	3	0
Fea	0	2	29	9	0	0
Hap	0	1	2	37	0	0
Sad	3	2	1	0	34	0
Sur	0	2	2	0	0	36

(b)

Figure 5.17. Confusion matrices of mouth region for out-group test. (a) ASNvsWSN (b) WSNvsASN.



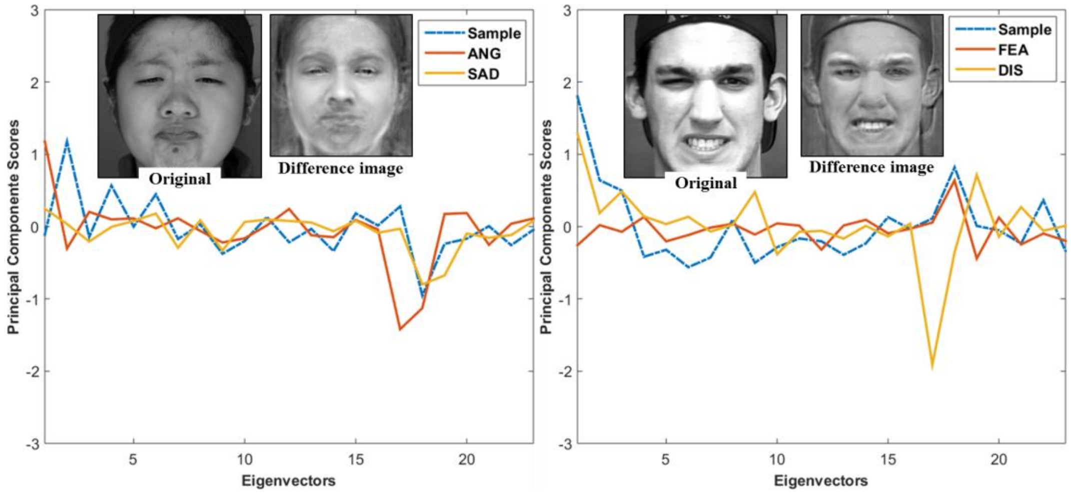
	Ang	Dis	Fea	Hap	Sad	Sur
Ang	24	2	4	1	9	0
Dis	15	7	6	2	10	0
Fea	0	0	19	0	18	3
Hap	1	1	1	29	8	0
Sad	3	1	4	11	21	0
Sur	0	0	3	0	7	30

(a)

	Ang	Dis	Fea	Hap	Sad	Sur
Ang	11	17	0	7	4	1
Dis	3	25	0	7	5	0
Fea	3	9	10	10	6	2
Hap	0	4	1	28	6	1
Sad	3	8	3	6	19	1
Sur	0	0	10	0	2	28

(b)

Figure 5.18. Confusion matrices of eyes region for out-group test. (a) ASNvsWSN (b) WSNvsASN.



(a) (b)

Figure 5.19. Examples of misclassification. (a) sample from ASN dataset showing anger misrecognized with sadness. (b) sample from WSN dataset showing fear misrecognized with disgust.

From these confusion matrices, it can be noticed that the misclassification problems mainly reside in the expressions of fear and disgust for both facial regions. Some examples of the faces misrecognized even for the human study are presented in **Figure 5.19**, which shows the original expressive facial image, its feature vector extracted using E-M-N combination and the visual representation using appearance features (difference image).

As mentioned in **Section 5.6**, there are some applications that require the recognition of the six basic expressions of emotion based on multicultural datasets. This problem is usually solved by training the system with multicultural datasets, however as seen in **Table 5.5**, this solution decreases the recognition accuracy, especially for WSN subjects. Therefore, two possible solutions to this problem are presented in this work, the first one includes a preprocessing stage based on ethnicity detection (**Section 5.6.1**), and the second is focused on the two expressions that present more visual differences among cultures (**Section 5.6.2**). The comparison between these proposals and the traditional solution that just employs multicultural datasets are presented in **Table 5.6**. In this table "***Multi.***" refers to the test based on the traditional solution; "***Ethnic***" to the proposal based on an extra preprocessing stage; "***Test A***" to the test that calculates specific-cultural sub-models of both expressions, disgust and fear; whereas "***Test B***" and "***Test C***" to the tests that only calculates it for disgust and fear respectively.

From **Table 5.6** we can see that the best solution for the multicultural scenario is the one that includes the early ethnic detection (97%). Indeed, just tests ***Ethnic*** and ***B*** overcome the average recognition rate of the traditional solution. In this way, it is possible to confirm that multicultural problems are related to the cultural differences of a few facial expressions, specifically disgust. However, the best solution still is based on the individual training for each cultural group. It is worth noting that none of this results overcome the recognition accuracy obtained by the in-group test of WSN (98.8 %).

Table 5.6. Average recognition rate of individual facial regions and its combinations obtained using multicultural datasets only.

Test:	***Multi.***	***Ethnic***	***Test A***	***Test B***	***Test C***
Eyes-eyebrows	68.5	73.2	68.4	69.6	68.2
Nose	72.1	72.8	72.0	73.3	71.8
Mouth	89.4	89.9	89.2	90.8	89.0
Eyes-Nose	80.2	81.1	80.0	81.5	79.9
Eyes-Mouth	94.4	96.5	94.2	95.9	94.0
Nose-Mouth	92.1	90.5	91.9	93.6	91.7
All regions	*95.0*	***<u>97.0</u>***	94.8	**96.5**	94.6

Finally, **Table 5.7** presents a comparison of the results obtained with those found in the literature. A few studies have analyzed the cross-cultural recognition capabilities of their proposals. Nevertheless, none of them have covered the out-group multicultural analysis. Even though the recognition accuracy is vastly ranged, the same trend is found, the results of in-group tests are better than those of out-group and WSN datasets achieve higher accuracy than ASN. Trend that still happens when using the two proposals for achieving the multicultural scenario problem.

Table 5.7. Classification accuracy of the proposed method and previous works with cross-cultural tests.

Training Dataset: *Testing Dataset:*	*WSN* *WSN*	*ASN* *ASN*	*ASN* *WSN*	*WSN* *ASN*	*MUL* *MUL*	*MUL* *WSN*	*MUL* *ASN*
Gabor **[42]**	91.5	89.7	54.1	55.9	-	-	-
LBP+SVM **[60]**	91.4	81.0	-	41.3	-	-	-
HOG+NNE **[59]**	-	-	55.9	63.6	93.8	-	-
CNN **[61]**	98.9	86.7	-	79.6	-	-	-
HOG+SVM **[58]**	93.5	88.6	39.9	42.3	84.7	-	-
Human Study	76.8	71.7	66.7	67.2	-	-	-
Proposed	98.8	95.0	72.5	85.0	97.0	97.8	96.4

It is worth noting that the results of related works presented in **Table 5.7** widely range from those obtained by the proposed method, mainly because of some specific reasons: the inconsistency on the number of samples per expression for training (some expressions used more than 200% of samples than others); the algorithms used for feature extraction (all of them only used appearance features for this task); and the null consideration of the static structure of individual faces for the analysis (some of the studies consider the neutral face as an extra class for the recognition task but not for the analysis itself). As an extra reference from the setup of mentioned works, **Table 5.8** shows the datasets used for defining the datasets of WSN and ASN.

Table 5.8. Datasets included in each cultural group of previous cross-cultural studies.

Dataset:	*WSN*	*ASN*
Gabor **[42]**	CK	JAFFE
LBP+SVM **[60]**	CK	JAFFE
HOG+NNE **[59]**	RAFD	JAFFE, TFEID
CNN **[61]**	CK+	JAFFE
HOG+SVM **[58]**	CK+, MUG	JAFFE
Proposed	CK+	JAFFE, JACFEE, TFEID

5.8 Conclusion

This chapter presented a methodical analysis of Western-Caucasian and East-Asian basic expressions focused on four facial regions. Based on the literature and from a psychological viewpoint, it is known that there exist in-group advantages for recognizing facial expressions when using cross-cultural datasets and Western-Caucasian facial expressions are closer related to the six basic emotions than those of East-Asian. Moreover, from an HCI viewpoint it is known that multicultural training is necessary for increasing the accuracy of FER systems. Following these findings, the cross-cultural tests analyzed in this paper present the same trend about the design of the training phase which suggests that WSN basic expressions are easier to recognize than those of ASN.

In addition, as indicated by previous works, strong cross-cultural similarities on the expressions of happiness and surprise were found in the analyzed datasets. On the other hand, two possible solutions were proposed for the multicultural problem and even though one of them reached higher accuracy, the proposed analysis of out-group multicultural test follows the same trend by exposing difficulties for recognizing ASN expressions. This issue points out the culture-specific differences that can be found on showing the six basic expressions. To this end, and thanks to the facial region segmentation presented in this paper, it was possible to fully analyze the cross-cultural recognition performance of individual facial regions and its combinations. In this way, the presented analysis also contributes to identifying the differences between WSN and ASN expressive faces which are primarily found in the regions of mouth and eyes-eyebrows, specifically for expressions of disgust and fear. In summary and as a general conclusion, it is better to set up specific training for each cultural group when working with multicultural datasets for FER. Based on the presented results, this issue could be pointed out by including an extra pre-processing stage for ethnicity recognition.

CHAPTER VI

6. CONCLUSIONS

This chapter presents the general conclusions reached after the culmination of this research work. In addition, some of the possible continuation lines related to the main topics of this thesis are also presented.

CONTENTS OF THE CHAPTER

6.1 Conclusion 92

6.2 Future Works 94

REFERENCES 95

LIST OF PUBLICATIONS 103

6.1 Conclusion

This thesis presented a methodical analysis of the prototypic facial expressions of Western-Caucasians and East-Asians for Facial Expression Recognition from a composite perspective of HCI and psychology. The analysis was built on the fact that rather than the debate related to the specificity of facial expressions, HCI approaches take the cultural universality of six basic expressions for granted. The proposed analysis was based on FER and visual analysis, and in general the conclusions of this thesis can be divided as follows.

1) About the techniques proposed for FER systems

In order to achieve the main goal of this thesis, some solutions for the problems attached to FER systems have been developed. Firstly, working with different facial regions represents an advantage for overcoming the problem of partial occlusion. The proposal based on the facial region segmentation of forehead, eyes-eyebrows, nose and mouth achieves high accuracy recognition when working with sub-block eigenphases algorithm, and even better when the static structure of the faces altogether with the 2D-DFT procedure were applied, reaching 92% and 95.8% of average recognition rate respectively. The best combination of individual feature vectors was obtained by using the regions of eyes-eyebrows, nose and mouth. In this way, it is possible to conclude that the forehead region is dispensable when trying to recognize facial expressions. Finally, as a result of employing the facial region segmentation and because of the proposal of DFT+PCA for the fusion of appearance and geometric features, the proposed hybrid-based method presented the highest accuracy performance. Reaching almost 99% of accuracy when enough facial landmarks are employed in the feature extraction. Once more, this performance was obtained by the combination of Eyes-Nose-Mouth. Hence, an extra conclusion based on the analysis of individual facial regions and its combinations, it was demonstrated that the mouth is the most important part of the face for developing facial expression recognition.

2) About the culture-based analysis of facial expressions

Taking advantage of the feature extraction methods proposed in the way to fulfill the aim of this thesis, the analysis of Western-Caucasian and East-Asian expressive faces were performed with many variations. However, all of the FER systems tested using the methodical analysis produced the same trend: WSN reached higher accuracy than ASN. Thus, the best results are obtained by in-group modality followed by out-multicultural, multicultural and at bottom the out-group. In this way, it was proved that there exist differences in the cross-cultural recognition of the six basic expressions even for three different FER systems (based on appearance- geometric- and hybrid-features). In addition, it can be seen that WSN facial expressions fit better for the prototypic expressions than those of ASN. An interesting finding observed when using out-group performance is that WSN recognizes the expressions of ASN better than ASN handles those of WSN, just exactly as human beings. Finally, thanks to the visual analysis and the confusion matrices, it was confirmed that the way of showing certain facial expressions differs between both racial groups, especially for the expressions of disgust and fear in the regions of mouth and eyes-eyebrows respectively. Hence, by identifying the differences between WSN and ASN expressive faces, two possible solutions for the multicultural problem were proposed. Even though these proposals overcome the traditional solution, the out-multicultural results cannot reach the accuracy of the in-group test. Therefore, we can conclude that the differences of showing facial expressions among cultures straightly affect the performance of automatic FER systems.

In summary and based on the experimental results presented in this thesis, it can be concluded that it is better to set up specific training for each cultural group when working with multicultural datasets for FER. This issue could be pointed out by including an extra pre-processing stage for ethnicity recognition when the situation warrants it. In addition, the culture-specific training strongly depends on the application of the FER system due to the cross-cultural similarities found on the called positive expressions. Therefore, looking back to the question formulated in **Chapter 1** and based on the findings obtained in this thesis, the correct answer for that question is: “companion robots and digital avatars should be adapted to express cultural-specific emotions”.

On the other hand, it has to be admitted that the proposed analysis deals with the problem of reliability of the datasets. Hence, most of the standard datasets are taken under controlled environments and expressions are shown by professional actors which sometimes exaggerate and break the spontaneity of a true facial expression. However, despite the limitations of the datasets, the proposed analysis helps to find cultural differences of specific facial expressions and introduces a methodical process for analyzing the cross-cultural capabilities of any FER systems.

6.2 Future Works

As a future work, it is possible to expand the limited size of the datasets by including databases from different countries of specific cultures (e.g. China and Korea for East-Asian) and by adding extra cultures (e.g. African, Latino, Indian, etc.). In addition, it is also necessary to analyze the facial expressions shown under non-controlled environments, situation known as "in-the-wild". Hence, the conclusions reached in this thesis could be covered.

Another option for contributing to the cultural specificity theory of facial expressions, is to analyze the categorization capability of the six prototypic expressions among different racial groups by applying unsupervised classification methods to each cultural dataset, thus it will enable the possibility to measure the cultural-specificity of the assumed basic expressions and the hypothesis of less than six basic expressions could be supported.

Finally, another alternative for the analysis is to employ semantic features in order to describe the facial expression of specific cultural groups. Hence, developing methods based on CNN or other deep learning algorithms, could provide results different than those obtained by the conventional methods.

References

[1] C. Darwin, *The expression of the emotions in man and animals* vol. 526: University of Chicago press, 1965.

[2] P. Ekman, "Cross-cultural studies of facial expression," *Darwin and facial expression: A century of research in review,* pp. 169-222, 1973.

[3] Y. Tian, T. Kanade, and J. F. Cohn, "Facial Expression Recognition," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds., ed London: Springer London, 2011, pp. 487-519.

[4] S. Deshmukh, M. Patwardhan, and A. Mahajan. (2016, Survey on real-time facial expression recognition techniques. *IET Biometrics 5(3)*, 155-163. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2014.0104

[5] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition,* vol. 36, pp. 259-275, 2003.

[6] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, pp. 1424-1445, 2000.

[7] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 921-926.

[8] R. E. Jack, "Culture and facial expressions of emotion," *Visual Cognition,* vol. 21, pp. 1248-1286, 2013/09/01 2013.

[9] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 7241-7244, May 8, 2012 2012.

[10] M. N. Dailey, C. Joyce, M. J. Lyons, M. Kamachi, H. Ishi, J. Gyoba, and G. W. Cottrell, "Evidence and a computational explanation of cultural differences in facial expression recognition," *Emotion,* vol. 10, p. 874, 2010.

[11] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 37, pp. 1113-1133, 2015.

[12] B. E. Holt, "Animal Drive and the Learning Process. An Essay Toward Radical Empiricism," *The Journal of Nervous and Mental Disease,* vol. 78, p. 554, 1933.

[13] L. Hearn, "In a Japanese Garden, Glimpses of Unfamiliar Japan," ed: New York: Harper, 1894.

[14] M. Mead, "DARWIN AND FACIAL EXPRESSION-CENTURY OF RESEARCH IN REVIEW-EKMAN, P," ed: OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513, 1975.

[15] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[16] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis," *Psychological bulletin,* vol. 128, p. 203, 2002.

[17] D. Matsumoto and M. Assar, "The effects of language on judgments of universal facial expressions of emotion," *Journal of Nonverbal Behavior,* vol. 16, pp. 85-99, June 01 1992.

[18] H. A. Elfenbein, "In-Group Advantage and Other-Group Bias in Facial Emotion Recognition," in *Understanding Facial Expressions in Communication: Cross-cultural and Multidisciplinary Perspectives*, M. K. Mandal and A. Awasthi, Eds., ed New Delhi: Springer India, 2015, pp. 57-71.

[19] D. Matsumoto, "Methodological requirements to test a possible in-group advantage in judging emotions across cultures: comment on Elfenbein and Ambady (2002) and evidence," 2002.

[20] R. E. Jack, R. Caldara, and P. G. Schyns, "Internal representations reveal cultural diversity in expectations of facial expressions of emotion," *Journal of Experimental Psychology: General,* vol. 141, p. 19, 2012.

[21] S.-M. Kang and A. S. Lau, "Revisiting the out-group advantage in emotion recognition in a multicultural society: Further evidence for the in-group advantage," *Emotion,* vol. 13, p. 203, 2013.

[22] A. J. O'Toole and V. Natu, "Computational perspectives on the other-race effect," *Visual Cognition,* vol. 21, pp. 1121-1137, 2013/09/01 2013.

[23] M. L. Smith, G. W. Cottrell, F. Gosselin, and P. G. Schyns, "Transmitting and Decoding Facial Expressions," *Psychological Science,* vol. 16, pp. 184-189, 2005.

[24] M. Yuki, W. W. Maddux, and T. Masuda, "Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States," *Journal of Experimental Social Psychology,* vol. 43, pp. 303-311, 2007/03/01/ 2007.

[25] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural Confusions Show that Facial Expressions Are Not Universal," *Current Biology,* vol. 19, pp. 1543-1548, 2009/09/29/ 2009.

[26] A. F. Shariff and J. L. Tracy, "What Are Emotion Expressions For?," *Current Directions in Psychological Science,* vol. 20, pp. 395-399, 2011.

[27] R. Jack, O. Garrod, and P. Schyns, "Dynamic signaling of facial expressions transmit social information in a hierarchical manner over time," *Journal of Vision,* vol. 13, p. 1277, 2013.

[28] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences,* vol. 107, pp. 2408-2412, February 9, 2010 2010.

[29] C.-L. Huang and Y.-M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification,"

Journal of Visual Communication and Image Representation, vol. 8, pp. 278-290, 1997/09/01 1997.

[30] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing,* vol. 18, pp. 881-905, 2000.

[31] J. Yang and A. Waibel, "A real-time face tracker," in *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on,* 1996, pp. 142-147.

[32] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, pp. 137-154, 2004.

[33] C. Huang, H. Ai, Y. Li, and S. Lao, "High-Performance Rotation Invariant Multiview Face Detection," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 29, pp. 671-686, 2007.

[34] L. Deligiannidis and H. R. Arabnia, *Emerging trends in image processing, computer vision and pattern recognition*: Morgan Kaufmann, 2014.

[35] M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and S. Sarkar, "Orthogonal projection images for 3D face detection," *Pattern Recognition Letters,* vol. 50, pp. 72-81, 2014.

[36] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition,* 1998, pp. 200-205.

[37] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580),* 2000, pp. 46-53.

[38] 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.

[39] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," in *18th International Conference on Pattern Recognition (ICPR'06),* 2006, pp. 1148-1153.

[40] Y. Lijun, W. Xiaozhou, S. Yi, W. Jun, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06),* 2006, pp. 211-216.

[41] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of multimedia,* vol. 1, pp. 22-35, 2006.

[42] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognition,* vol. 45, pp. 80-91, 2012.

[43] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "PCA-based dictionary building for accurate facial expression recognition via sparse representation," *Journal of Visual Communication and Image Representation,* vol. 25, pp. 1082-1092, 2014.

[44] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference," *IEEE Transactions on Multimedia,* vol. 12, pp. 682-691, 2010.

[45] X. Zhao and S. Zhang, "Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding," *EURASIP Journal on Advances in Signal Processing,* vol. 2012, p. 20, 2012.

[46] R. Xiao, Q. Zhao, D. Zhang, and P. Shi, "Facial expression recognition on multiple manifolds," *Pattern Recogn,* vol. 44, 2011.

[47] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. l. Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1-7.

[48] M. F. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 42, pp. 28-43, 2012.

[49] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition,* vol. 47, pp. 1282-1293, 2014.

[50] Z. Li, J.-i. Imai, and M. Kaneko, "Facial Expression Recognition Using Facial-component-based Bag of Words and PHOG Descriptors," *The Journal of The Institute of Image Information and Television Engineers,* vol. 64, pp. 230-236, 2010.

[51] L. Zhang, S. Chen, T. Wang, and Z. Liu, "Automatic Facial Expression Recognition Based on Hybrid Features," *Energy Procedia,* vol. 17, pp. 1817-1823, 2012/01/01 2012.

[52] S. Wan and J. K. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognition,* vol. 47, pp. 1859-1868, 2014.

[53] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition,* vol. 41, pp. 833-851, 2008.

[54] W. Zhen and T. S. Huang, "Capturing subtle facial motions in 3D face tracking," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1343-1350 vol.2.

[55] P. Yang, Q. Liu, and D. N. Metaxas, "RankBoost with l1 regularization for facial expression recognition and intensity estimation," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1018-1025.

[56] T. Moriyama, T. Kanade, J. F. Cohn, X. Jing, Z. Ambadar, G. Jiang, and H. Imamura, "Automatic recognition of eye blinking in spontaneously occurring behavior," in *Object recognition supported by user interaction for service robots*, 2002, pp. 78-81 vol.4.

[57] H. Fang, N. Mac Parthaláin, A. J. Aubrey, G. K. L. Tam, R. Borgo, P. L. Rosin, P. W. Grant, D. Marshall, and M. Chen, "Facial expression recognition in dynamic sequences: An integrated approach," *Pattern Recognition,* vol. 47, pp. 1271-1281, 2014.

[58] F. A. M. da Silva and H. Pedrini, "Effects of cultural characteristics on building an emotion classifier through facial expression analysis," *Journal of Electronic Imaging,* vol. 24, pp. 023015-023015, 2015.

[59] G. Ali, M. A. Iqbal, and T.-S. Choi, "Boosted NNE collections for multicultural facial expression recognition," *Pattern Recognition,* vol. 55, pp. 14-27, 2016.

[60] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image Vis Comput,* vol. 27, 2009.

[61] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognition,* vol. 61, pp. 610-628, 2017.

[62] H. Towner and M. Slater, "Reconstruction and Recognition of Occluded Facial Expressions Using PCA," in *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings,* A. C. R. Paiva, R. Prada, and R. W. Picard, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 36-47.

[63] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing,* vol. 26, pp. 1052-1067, 2008.

[64] Z. Ligang, D. Tjondronegoro, and V. Chandran, "Toward a more robust facial expression recognition in occluded images using randomly sampled Gabor based templates," in *2011 IEEE International Conference on Multimedia and Expo*, 2011, pp. 1-6.

[65] Y. Miyakoshi and S. Kato, "Facial emotion detection considering partial occlusion of face using Bayesian network," in *2011 IEEE Symposium on Computers & Informatics,* 2011, pp. 96-101.

[66] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, M. Nakano-Miyatake, and H. Perez-Meana, "A sub-block-based eigenphases algorithm with optimum sub-block size," *Knowledge-Based Systems,* vol. 37, pp. 415-426, 2013.

[67] Y. Luo, C.-m. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of PCA and LBP with SVM," *Optik - International Journal for Light and Electron Optics,* vol. 124, pp. 2767-2770, 2013.

[68] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, 2005, pp. 1692-1698 Vol. 2.

[69] M. Savvides, B. V. K. V. Kumar, and P. K. Khosla, "Eigenphases vs eigenfaces," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004, pp. oo810-oo813 Vol.3.

[70] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE,* vol. 69, pp. 529-541, 1981.

[71] V. N. Vapnik and V. Vapnik, *Statistical learning theory* vol. 1: Wiley New York, 1998.

[72] C. Chih-Chung, "LIBSVM : A library for support vector machines," *ACM Trans. Intelligent Systems and Technology,* vol. 2, pp. 27:1-27:27, 2011 2011.

[73] Z. Zhengyou, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 454-459.

[74] S. Jain, H. Changbo, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1642-1649.

[75] X. Xie and K.-M. Lam, "Facial expression recognition based on shape and texture," *Pattern Recognition,* vol. 42, pp. 1003-1011, 2009.

[76] W. Hwang, H. Wang, H. Kim, S. C. Kee, and J. Kim, "Face Recognition System Using Multiple Face Model of Hybrid Fourier Feature Under Uncontrolled Illumination Variation," *IEEE Transactions on Image Processing,* vol. 20, pp. 1152-1165, 2011.

[77] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859-1866.

[78] G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A Comprehensive Performance Evaluation of Deformable Face Tracking “In-the-Wild”," *International Journal of Computer Vision,* pp. 1-35, 2017.

[79] H. M. El-Bakry, "Automatic human face recognition using modular neural networks," *Machine Graphics and Vision,* vol. 10, pp. 47-73, 2001.

[80] J. Yi, X. Mao, L. Chen, Y. Xue, and A. Compare, "Facial expression recognition considering individual differences in facial structure and texture," *IET Computer Vision,* vol. 8, pp. 429-440, 2014.

[81] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, 2010, pp. 1-4.

[82] L.-F. Chen and Y.-S. Yen, "Taiwanese facial expression image database," *Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan,* 2007.

[83] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2983-2991.

[84] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications,* vol. 76, pp. 7803-7821, March 01 2017.

[85] W. Wei and Q. Jia, "Weighted Feature Gaussian Kernel SVM for Emotion Recognition," *Intell. Neuroscience,* vol. 2016, p. 11, 2016.

[86] X. Pu, K. Fan, X. Chen, L. Ji, and Z. Zhou, "Facial expression recognition from image sequences using twofold random forest classifier," *Neurocomputing,* vol. 168, pp. 1173-1180, 2015/11/30/ 2015.

[87] 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, pp. 83-92, 2013.

[88] D. Ghimire, J. Lee, Z.-N. Li, and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications,* vol. 76, pp. 7921-7946, March 01 2017.

[89] F. A. Maximiano da Silva and H. Pedrini, "Geometrical Features and Active Appearance Model Applied to Facial Expression Recognition," *International Journal of Image and Graphics,* vol. 16, p. 1650019, 2016.

[90] M. Goyani and N. Patel, "Multi-Level Haar Wavelet based Facial Expression Recognition using Logistic Regression," *Indian Journal of Science and Technology,* vol. 10, 2017.

[91] N. Farajzadeh, G. Pan, and Z. Wu, "Facial expression recognition based on meta probability codes," *Pattern Analysis and Applications,* vol. 17, pp. 763-781, November 01 2014.

[92] A. M. Ashir and A. Eleyan, "Facial expression recognition based on image pyramid and single-branch decision tree," *Signal, Image and Video Processing,* vol. 11, pp. 1017-1024, September 01 2017.

[93] H. W. Kung, Y. H. Tu, and C. T. Hsu, "Dual Subspace Nonnegative Graph Embedding for Identity-Independent Expression Recognition," *IEEE Transactions on Information Forensics and Security,* vol. 10, pp. 626-639, 2015.

[94] G. Benitez-Garcia, G. Sanchez-Perez, H. Perez-Meana, K. Takahashi, and M. Kaneko, "Facial expression recognition based on facial region segmentation and modal value approach," *IEICE TRANSACTIONS on Information and Systems,* vol. 97, pp. 928-935, 2014.

[95] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random Gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing,* vol. 145, pp. 451-464, 2014.

[96] M. Biehl, D. Matsumoto, P. Ekman, V. Hearn, K. Heider, T. Kudoh, and V. Ton, "Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences," *Journal of Nonverbal Behavior,* vol. 21, pp. 3-21, 1997.

[97] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Research,* vol. 41, pp. 1179-1208, 2001.

[98] S. Morishima, Y. Yagi, M. Kaneko, H. Harashima, M. Yachida, and F. Hara, "Construction of standard software for face recognition and synthesis," *Tec Rep IEICE PRMU97–282,* pp. 129-136, 1998.

[99] 金子正秀, "6. コンピュータ似顔絵," *映像情報メディア学会誌,* vol. 62, pp. 1938-1943, 2008.

[100] A. Fierro-Radilla, K. Perez-Daniel, M. Nakano-Miyatakea, H. Perez-Meana, and J. Benois-Pineau, "An Effective Visual Descriptor Based on Color and Shape Features for Image Retrieval," in *Human-Inspired Computing and Its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I,* A. Gelbukh, F. C. Espinoza, and S. N.

Galicia-Haro, Eds., ed Cham: Springer International Publishing, 2014, pp. 336-348.

[101] S. Fu, H. He, and Z. G. Hou, "Learning Race from Face: A Survey," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 36, pp. 2483-2509, 2014.

[102] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transactions on Multimedia,* vol. 1, pp. 264-277, 1999.

[103] H.-T. Quan, M. Meguro, and M. Kaneko, "Skin-color extraction in images with complex background and varying illumination," in *Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings.*, 2002, pp. 280-285.

[104] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition,* vol. 40, pp. 1106-1122, 2007/03/01/ 2007.

LIST OF PUBLICATIONS

Journal Papers:

1. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Methodical Analysis of Western-Caucasian and East-Asian Basic Facial Expressions of Emotions based on Specific Facial Regions," *Journal of Signal and Information Processing,* vol. 8, no. 2, pp. 78-98, 2017.
   (Related to the contents of Chapter 5)

2. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Facial Expression Recognition Based on Local Fourier Coefficients and Facial Fourier Descriptors," *Journal of Signal and Information Processing,* vol. 8, no. 3, pp. 132-151, 2017. (Related to the contents of Chapter 4)

International Conference Papers:

3. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Analysis of In- and Out-group Differences between Western and East-Asian Facial Expression Recognition," 2017 15*th IAPR International Conference on Machine Vision Applications* (*MVA*2017), pp. 402-405, Nagoya, Japan, 2017.
   (Related to the contents of Chapter 5)

Other Publications and Presentations:

4. **G. Benitez-Garcia,** G. Sanchez-Perez, H. Perez-Meana, K. Takahashi, and M. Kaneko, "Facial expression recognition based on facial region segmentation and modal value approach," *IEICE Transactions on Information and Systems,* vol. E97-D, pp. 928-935, 2014.

5. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Analysis of differences between Western and East-Asian faces based on facial region segmentation and PCA for facial expression recognition," *The Irago Conference* 2016, pp. 020025.1-020025.10, Tokyo, Japan, 2017.

6. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Appearance- and Geometric-based Methodical Feature Analysis of Facial Parts using PCA for Facial Expression Recognition," *Vision Engineering Workshop* 2016 (*ViEW*2016), pp. 122-127, 2016.

7. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Geometric-based Feature Analysis of Differences between Western and East-Asian Expressive Faces Based on Principal Component Scores," *Human Interface Symposium* 2016 (*HIS*2016), pp. 943-946, 2016.

8. **G. Benitez-Garcia,** T. Nakamura, and M. Kaneko, "Analysis of the Effect of Static Structure of Faces for Facial Expression Recognition," 21*st JFACE Annual Conference* (*Forum Kaogaku* 2016), p. 81, 2016.

9. **G. Benitez-Garcia,** G. Sanchez-Perez, H. Perez-Meana, K. Takahashi, and M. Kaneko, "Facial Expression Recognition Under Partial Occlusion Based on Facial Region Segmentation," 映像情報メディア学会技術報告, vol. 37, pp. 95-98, 2013.

10. **G. Benitez-Garcia,** G. Sanchez-Perez, H. Perez-Meana, K. Takahashi, and M. Kaneko, "Comparison of Facial Expression Recognition Rates Depending on Combination of Different Facial Regions," 映像情報メディア学会冬季大会講演予稿集, pp. 12-7-1, 2012.

ACKNOWLEDGMENTS

The present thesis is a summary of my doctoral studies at the University of Electro-Communications and represents the culmination of my formal student life. This work would not have been accomplished without the help of many people. All of you who have been with me through this journey deserves more than my unconditional gratitude. THANK YOU.

First of all, I would like to express my deepest thanks to my supervisor, Prof. Masahide Kaneko. He generously supported and guided me for getting my Ph.D. degree. I had the opportunity to work under his supervision as an exchange student when I first came to Japan in April 2012, and after this first approach, Prof. Kaneko helped me to come back for the doctoral course by promoting me to the MEXT scholarship that provided me financial support from October 2014 to the conclusion of this work in September 2017. I can't properly express my gratitude and admiration to Prof. Kaneko. It is simply amazing how diligent and hard worker he is. Indisputably, Prof. Kaneko is a role model. I hope to humbly follow his steps. Also, I am especially thankful to Dr. Nakamura, for his suggestions and revisions on my research work.

I would like to extend my thanks to the members of my supervisory committee, which includes Prof. Kazuo Tanaka, Prof. Hiroshi Yokoi, Prof. Masafumi Uchida and Prof. Takayuki Nagai. Thanks to their meaningful suggestions I was able to improve the present work.

Also, I would like to especially thank the members of my lab, the English professors, the office staff members, my classmates, the Mexican student community and all the people who made my student life happier in all senses.

Last but not least, I would like to express my gratitude and love to my beloved wife and my newborn baby, who were always with me no matter how difficult and painful was this journey. Frankly speaking, their love pushed me to overcome all adversities. On the other hand, I owe special thanks to my father who always cares and supports me from the distance and all my family in Mexico. I can't finish writing these acknowledgements without mentioning my mother, I am definitely convinced that she is still taking care of me. *True love may live forever*.

AUTHOR'S BIOGRAPHY

Benitez Garcia Gibran de Jesus was born in Mexico City, Mexico on October 6th, 1988. He received the B.S. degree on Computer Science Engineer and the M.S. degree from the Mechanical Engineering School of the National Polytechnic Institute, Mexico City in 2011 and 2014, respectively. He was graduated with honors for his excellent M.S. research work. He stayed in the University of Electro-Communications in Tokyo, Japan, from April 2012 to March 2013 as a JUSST student. He entered to the doctoral course at the same university on October 2014. His research interests include face and facial expression recognition, automatic human-behavior detection, pattern recognition and computer vision.