Universidade Estadual de Campinas
Instituto de Computação

Jadisha Yarif Ramírez Cornejo

Pattern Recognition in Facial Expressions: Algorithms and Applications

Reconhecimento de Padrões em Expressões Faciais: Algoritmos e Aplicações

CAMPINAS
2020

Jadisha Yarif Ramírez Cornejo

Pattern Recognition in Facial Expressions: Algorithms and Applications

Reconhecimento de Padrões em Expressões Faciais: Algoritmos e Aplicações

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutora em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Tese defendida por Jadisha Yarif Ramírez Cornejo e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

R145p

Ramírez Cornejo, Jadisha Yarif, 1990-
Pattern recognition in facial expressions : algorithms and applications / Jadisha Yarif Ramírez Cornejo. – Campinas, SP : [s.n.], 2020.

Orientador: Hélio Pedrini.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Reconhecimento de emoções humanas. 2. Reconhecimento de expressões faciais. 3. Reconhecimento de padrões. 4. Síndromes genéticas. 5. Visão por computador. 6. Processamento de imagens – Técnicas digitais. 7. Redes neurais convolucionais. I. Pedrini, Hélio, 1963-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Reconhecimento de padrões em expressões faciais : algoritmos e aplicações
Palavras-chave em inglês:
Human emotion recognition
Facial expressions recognition
Pattern recognition
Genetic disorders
Computer vision
Image processing - Digital techniques
Convolutional neural networks
Área de concentração: Ciência da Computação
Titulação: Doutora em Ciência da Computação
Banca examinadora:
Hélio Pedrini [Orientador]
Sarajane Marques Peres
Edimilson Batista dos Santos
Fabricio Aparecido Breve
Marco Antonio Garcia de Carvalho
Data de defesa: 19-02-2020
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-7283-1482
- Currículo Lattes do autor: http://lattes.cnpq.br/1451212382988518

Universidade Estadual de Campinas
Instituto de Computação

Jadisha Yarif Ramírez Cornejo

Pattern Recognition in Facial Expressions: Algorithms and Applications

Reconhecimento de Padrões em Expressões Faciais: Algoritmos e Aplicações

Banca Examinadora:

- Prof. Dr. Hélio Pedrini
  IC / UNICAMP
- Profa. Dra. Sarajane Marques Peres
  EACH / USP
- Prof. Dr. Edimilson Batista dos Santos
  DCOMP / UFSJ
- Prof. Dr. Fabricio Aparecido Breve
  UNESP / Rio Claro
- Prof. Dr. Marco Antonio Garcia de Carvalho
  FT / UNICAMP

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 19 de fevereiro de 2020

Acknowledgements

- I would like to express my gratitude to my advisor, Prof. Hélio Pedrini, for his continuous support during my Ph.D. study and research.
- I would like to thank my parents, Alicia and Fredy, for their constant encouragement not to give up and reach my goals. I am grateful to my grandmother, Adriana, and to my brothers, Fredy and Shamir, for also being there throughout my life.
- I would also like to thank Jairo for his continued motivation and enthusiasm during the last stage of my doctorate. I am grateful to my friends and colleagues that I met during my PhD.
- Last but not least, I would like to express my gratitude to the Institute of Computing at the University of Campinas for being like my second home during these years. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível superior - Brasil (CAPES) - Code 001, and CNPq.

Resumo

O reconhecimento de emoções tem-se tornado um tópico relevante de pesquisa pela comunidade científica, uma vez que desempenha um papel essencial na melhoria contínua dos sistemas de interação humano-computador. Ele pode ser aplicado em diversas áreas, tais como medicina, entretenimento, vigilância, biometria, educação, redes sociais e computação afetiva. Há alguns desafios em aberto relacionados ao desenvolvimento de sistemas emocionais baseados em expressões faciais, como dados que refletem emoções mais espontâneas e cenários reais. Nesta tese de doutorado, apresentamos diferentes metodologias para o desenvolvimento de sistemas de reconhecimento de emoções baseado em expressões faciais, bem como sua aplicabilidade na resolução de outros problemas semelhantes. A primeira metodologia é apresentada para o reconhecimento de emoções em expressões faciais ocluídas baseada no Histograma da Transformada Census (CENTRIST). Expressões faciais ocluídas são reconstruídas usando a Análise Robusta de Componentes Principais (RPCA). A extração de características das expressões faciais é realizada pelo CENTRIST, bem como pelos Padrões Binários Locais (LBP), pela Codificação Local do Gradiente (LGC) e por uma extensão do LGC. O espaço de características gerado é reduzido aplicando-se a Análise de Componentes Principais (PCA) e a Análise Discriminante Linear (LDA). Os algoritmos K-Vizinhos mais Próximos (KNN) e Máquinas de Vetores de Suporte (SVM) são usados para classificação. O método alcançou taxas de acerto competitivas para expressões faciais ocluídas e não ocluídas. A segunda é proposta para o reconhecimento dinâmico de expressões faciais baseado em Ritmos Visuais (VR) e Imagens da História do Movimento (MHI), de modo que uma fusão de ambos descritores codifique informações de aparência, forma e movimento dos vídeos. Para extração das características, o Descritor Local de Weber (WLD), o CENTRIST, o Histograma de Gradientes Orientados (HOG) e a Matriz de Coocorrência em Nível de Cinza (GLCM) são empregados. A abordagem apresenta uma nova proposta para o reconhecimento dinâmico de expressões faciais e uma análise da relevância das partes faciais. A terceira é um método eficaz apresentado para o reconhecimento de emoções audiovisuais com base na fala e nas expressões faciais. A metodologia envolve uma rede neural híbrida para extrair características visuais e de áudio dos vídeos. Para extração de áudio, uma Rede Neural Convolucional (CNN) baseada no log-espectrograma de Mel é usada, enquanto uma CNN construída sobre a Transformada de Census é empregada para a extração das características visuais. Os atributos audiovisuais são reduzidos por PCA e LDA, então classificados por KNN, SVM, Regressão Logística (LR) e Gaussian Naïve Bayes (GNB). A abordagem obteve taxas de reconhecimento competitivas, especialmente em dados espontâneos. A penúltima investiga o problema de detectar a síndrome de Down a partir de fotografias. Um descritor geométrico é proposto para extrair características faciais. Experimentos realizados em uma base de dados pública mostram a eficácia da metodologia desenvolvida. A última metodologia trata do reconhecimento de síndromes genéticas em fotografias. O método visa extrair atributos faciais usando características de uma rede neural profunda e medidas antropométricas. Experimentos são realizados em uma base de dados pública, alcançando taxas de reconhecimento competitivas.

Abstract

Emotion recognition has become a relevant research topic by the scientific community, since it plays an essential role in the continuous improvement of human-computer interaction systems. It can be applied in various areas, for instance, medicine, entertainment, surveillance, biometrics, education, social networks, and affective computing. There are some open challenges related to the development of emotion systems based on facial expressions, such as data that reflect more spontaneous emotions and real scenarios. In this doctoral dissertation, we propose different methodologies to the development of emotion recognition systems based on facial expressions, as well as their applicability in the development of other similar problems. The first is an emotion recognition methodology for occluded facial expressions based on the Census Transform Histogram (CENTRIST). Occluded facial expressions are reconstructed using an algorithm based on Robust Principal Component Analysis (RPCA). Extraction of facial expression features is then performed by CENTRIST, as well as Local Binary Patterns (LBP), Local Gradient Coding (LGC), and an LGC extension. The generated feature space is reduced by applying Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms are used for classification. This method reached competitive accuracy rates for occluded and non-occluded facial expressions. The second proposes a dynamic facial expression recognition based on Visual Rhythms (VR) and Motion History Images (MHI), such that a fusion of both encodes appearance, shape, and motion information of the video sequences. For feature extraction, Weber Local Descriptor (WLD), CENTRIST, Histogram of Oriented Gradients (HOG), and Gray-Level Co-occurrence Matrix (GLCM) are employed. This approach shows a new direction for performing dynamic facial expression recognition, and an analysis of the relevance of facial parts. The third is an effective method for audio-visual emotion recognition based on speech and facial expressions. The methodology involves a hybrid neural network to extract audio and visual features from videos. For audio extraction, a Convolutional Neural Network (CNN) based on log Mel-spectrogram is used, whereas a CNN built on Census Transform is employed for visual extraction. The audio and visual features are reduced by PCA and LDA, and classified through KNN, SVM, Logistic Regression (LR), and Gaussian Naïve Bayes (GNB). This approach achieves competitive recognition rates, especially in a spontaneous data set. The second last investigates the problem of detecting Down syndrome from photographs. A geometric descriptor is proposed to extract facial features. Experiments performed on a public data set show the effectiveness of the developed methodology. The last methodology is about recognizing genetic disorders in photos. This method focuses on extracting facial features using deep features and anthropometric measurements. Experiments are conducted on a public data set, achieving competitive recognition rates.

List of Figures

2.1 A 3×3 neighborhood template of LGC operator. 24
2.2 Diagram with the main steps of the facial expression recognition methodology. 25
2.3 Cropped images with occluded facial regions from the JAFFE data set. 26
2.4 (a) Cropped images without occlusions from the JAFFE data set; (b) faces with occluded regions; (c) reconstructed faces; (d) filling the occluded facial regions from (c). 26
2.5 (a) Cropped image from the JAFFE data set; (b) LBP image from (a); (c) LBP image is divided into 63 regions. 27
2.6 (a) Cropped image from CK+ database; (b) census transformed image from (a). 27
2.7 (a) Cropped image from CK+ database; (b) LGC image from (a); (c) LGC-HD image from (a). 27

3.1 Diagram of the facial expression recognition methodology. 33
3.2 Cropped facial expression regions from BU-4DFE data set. From left to right, the emotions cataloged are anger, surprise, disgust, and sadness. (a) original video frame samples; (b) upper facial expression regions; (c) lower facial expression regions. 34
3.3 Visual rhythm representations from the top facial region sequences for different emotions. (a) anger; (b) disgust; (c) fear; (d) happiness; (e) sadness; (f) surprise. 35
3.4 (a)(c) Examples of video sequences from the upper and lower facial region of a sample from the BU-4DFE data set; (b)(d) MHI generated from the video sequence in (a)(c). 36
3.5 (a) MHI representation from upper facial region sequences extracted from BU-4DFE database; (b) filtered image by WLD from image (a); (c) WLD image divided into 63 regions. 37
3.6 (a) MHI representation from lower facial region sequences extracted from BU-4DFE database; (b) census transformed image from (a). 38

4.1 Diagram of the proposed methodology for audio-visual emotion recognition. . . 46
4.2 3×3 Census Transform for a single output pixel. 47
4.3 (a) Cropped image extracted from BAUM-1s database; (b) CLAHE processed image from (a); (c) census-transformed image from (a). 48
4.4 Cropped facial expression regions from (a) RML, (b) eNTERFACE05, and (c) BAUM-1s data sets. From left to right, the emotions cataloged are anger, disgust, fear, happiness, sadness, and surprise. 50

5.1 Diagram with the main steps of the proposed Down syndrome detection methodology. 57
5.2 Facial detection step. (a) detected face; (b) marked fiducial facial points. 58

5.3 Feature extraction. (a)-(b) sixteen facial landmark detected from the face; (c) fourteen distances extracted from the sixteen fiducial points. 59

6.1 Main steps of the genetic syndrome recognition methodology. 65
6.2 Feature extraction. (a)-(b) twenty seven facial fiducial points detected; (c) twenty one distances extracted from the twenty seven landmarks. 66

List of Tables

2.1 Average accuracy rates (%) for non-occluded facial images from CK+ and JAFFE data sets. . . . 28
2.2 Average accuracy rates (%) for occluded facial images from CK+ and JAFFE data sets. . . . 29
2.3 Comparison of average accuracy rates (%) for non-occluded (non-oc) and occluded facial expression images (oc) on CK+ and JAFFE data sets. . . . 29

3.1 Average accuracy rates (%) using Visual Rhythms following a diagonal orientation over the cropped facial images through GLCM, CENTRIST and WLD operators. . . . 40
3.2 Average accuracy rates (%) using Motion History Images over the cropped facial images through HOG, CENTRIST and WLD operators. . . . 40
3.3 Average accuracy rates (%) obtained with Motion History Images over different facial regions using the WLD operator. . . . 40
3.4 Average accuracy rates (%) for the Visual Rhythm with different orientation patterns over the upper and lower facial regions using the WLD operator. . . . 41
3.5 Average accuracy rates (%) employing Motion History Images and Visual Rhythms (horizontal orientation) over the upper and lower facial regions, using the WLD operator. . . . 41
3.6 Comparison to the state of the art on the BU-4DFE data set. . . . 42

4.1 Subject-independent unimodality average accuracy rates (%) using audio features for each data set. . . . 51
4.2 Subject-independent unimodality average accuracy rates (%) using visual features for each data set. . . . 52
4.3 Subject-independent unimodality average accuracy rates (%) using audio features extracted by the original pre-trained AlexNet and visual features obtained through the CT based VGG for each data set. . . . 52
4.4 Comparison of subject-independent audio recognition average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets. . . . 52
4.5 Comparison of subject-independent visual recognition average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets. . . . 53
4.6 Subject-independent multimodality average accuracy rates (%) using the fusion of audio and visual features for each data set. . . . 53
4.7 Comparison of subject-independent multimodality average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets. . . . 54

5.1 Comparison of some methods for Down syndrome recognition according to accuracy rates (in percentage). . . . 57

5.2 Average accuracy, in percentage, using our geometric representation and the CENTRIST descriptor on the evaluated data set. 61
5.3 Comparison of average accuracy rates (%) for Down syndrome detection over our formed data set. 62
5.4 Confusion matrix for the proposed method. 62
5.5 Confusion matrix for the method developed by Zhao et al. [152]. 62

6.1 Average accuracy (%) using our geometric, ORB and deep features, and the fusion between them on the evaluated data set. 69
6.2 Comparison of average accuracy rates (%) for genetic syndrome recognition. . 70

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
BAUM-1s	Bahçeşehir University Multimodal Affective Database
BRIEF	Binary Robust Independent Elementary Features
BU-4DFE	Binghamton University 4D Facial Expression
CENTRIST	Census Transform Histogram
CK	Cohn-Kanade
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CT	Census Transform
DBN	Deep Belief Network
EBGM	Elastic Bunch Graph Matching
FAST	Features from Accelerated Segment Test
FBE	Filter Bank Energies
FISH	Fluorescence in Situ Hybridization
GLCM	Gray-Level Co-occurrence Matrix
GNB	Gaussian Naïve Bayes
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
ICA	Independent Component Analysis
JAFFE	Japanese Female Facial Expression
KNN	K-Nearest Neighbors
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LGC	Local Gradient Coding
LGC-HD	Local Gradient Coding of Horizontal and Diagonal gradient priority
LLD	Low-Level Descriptors
LOSGO	Leave-One-Speakers-Group-Out
LOSO	Leave-One-Subject-Out
LPQ	Local Phase Quantization
LR	Logistic Regression
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstrum Coefficients
MHI	Motion History Images
ORB	Oriented FAST and Rotated BRIEF

PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PNCC	Power Normalized Cepstral Coefficients
QIM	Quantized Image Matrix
RASTA	Relative Spectral
RML	Ryerson Multimedia Lab
RPCA	Robust Principal Component Analysis
SIFT	Scale Invariant Feature Transform
SURF	Speed up Robust Features
SVM	Support Vector Machine
VGG	Visual Geometry Group
VR	Visual Rhythm
WLD	Weber Local Descriptor

Contents

1 Introduction **16**
- 1.1 Problem Characterization . . . 16
- 1.2 Objectives and Contributions . . . 17
- 1.3 Research Questions . . . 18
- 1.4 List of Publications . . . 18
- 1.5 Text Organization . . . 19

2 Recognition of Occluded Facial Expressions based on CENTRIST Features **21**
- 2.1 Introduction . . . 21
- 2.2 Background . . . 22
- 2.3 Methodology . . . 25
- 2.4 Experimental Results . . . 28
- 2.5 Final Considerations . . . 30

3 Dynamic Facial Expression Recognition Based on Visual Rhythms and Motion History Images **31**
- 3.1 Introduction . . . 31
- 3.2 Dynamic Facial Expression Recognition Method . . . 32
  - 3.2.1 Preprocessing . . . 32
  - 3.2.2 Video Characterization . . . 34
  - 3.2.3 Facial Feature Extraction . . . 36
  - 3.2.4 Feature Reduction . . . 39
  - 3.2.5 Classification . . . 39
- 3.3 Experimental Results . . . 39
- 3.4 Final Considerations . . . 42

4 Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform **43**
- 4.1 Introduction . . . 43
- 4.2 Methodology . . . 46
  - 4.2.1 Audio Data . . . 46
  - 4.2.2 Visual Data . . . 47
  - 4.2.3 Audio Network . . . 48
  - 4.2.4 Visual Network . . . 48
  - 4.2.5 Audio Segment-based Fusion . . . 49
  - 4.2.6 Video Frame-based Fusion . . . 49
  - 4.2.7 Fusion and Classification . . . 49
- 4.3 Experimental Results . . . 49
- 4.4 Final Considerations . . . 54

5 Down Syndrome Detection based on Facial Features using a Geometric Descriptor 55
5.1 Introduction . . . 55
5.2 Background . . . 56
5.3 Proposed Methodology . . . 57
5.3.1 Facial Detection . . . 58
5.3.2 Feature Extraction . . . 58
5.3.3 Feature Reduction . . . 60
5.3.4 Classification . . . 60
5.4 Experimental Results . . . 60
5.5 Final Considerations . . . 63

6 Recognition of Genetic Disorders Based on Deep Features and Geometric Representation 64
6.1 Introduction . . . 64
6.2 Methodology . . . 65
6.2.1 Preprocessing . . . 66
6.2.2 Feature Extraction . . . 66
6.2.3 Feature Reduction and Classification . . . 68
6.3 Experimental Results . . . 68
6.4 Final Considerations . . . 70

7 Conclusions and Future Work 71

Bibliography 74

Chapter 1

Introduction

This chapter presents the problem under investigation in this doctoral dissertation, its motivations, challenges, objectives, contributions, list of publications, research questions, and text organization.

1.1 Problem Characterization

In recent years, emotion recognition has gained increasing attention by the scientific and industrial community in various knowledge domains, for instance, surveillance, education, healthcare, entertainment, biometrics, robotics, among others.

Emotions can be expressed through audio-visual forms or biological changes, such as facial expressions, voice intonation, cardiac rhythm, temperature, and brain signals. Facial expression analysis means a fundamental clue to perform emotion recognition. Facial expressions are the most basic, natural, and non-verbal form to convey the emotional state of a person in interpersonal communication. Facial expressions can also provide evidence about our cognitive state, such as physical state, pain, fatigue, degree of attention. As a result, many applications have emerged to improve human-computer interaction, such as driver fatigue detection, stress detection, physical pain detection, attention analysis in an online learning environment, recommendation systems, among others [24, 25, 26, 29].

Several decades ago, the facial expression analysis started using two-dimensional (2D) images, allowing the development of emotion recognition systems based on facial expression in constrained situations. A persistent problem for developing these systems is that they use image collections that do not reflect real-world scenarios. Most of them operate with image sets that do not contain occlusions, for instance, glasses, hats, scarves, facial hair, facial deformations, body gestures, and lighting [24, 25, 26, 29]. Moreover, most facial images are captured under controlled conditions, i.e., homogeneous background, frontal poses, controlled lightning, and acted expressions. The absence of these issues during the training phase may impact the development of accurate emotion recognition systems. Consequently, in this dissertation, we present a method for addressing facial expression recognition under simulate occlusions in 2D images.

However, 2D images do not allow an accurate analysis of spontaneous facial expressions due to their lack of temporal information. The accuracy of 2D facial expression recognition in real

scenes is considerably reduced due to several of the previously mentioned challenges. These difficulties motivated the exploration of dynamic 2D data to solve them, as the availability of spatial information could help to have different facial views. In addition, temporal information provides more information for recognizing facial expression on behalf of its dynamic nature [28]. Although the use of dynamic data can improve the facial expression recognition accuracy to a limit extent, it remains a major challenge. We also perform new procedures for the dynamic emotion recognition in this doctoral dissertation.

For many years, many efforts have been devoted to enhance automatic emotion recognition using unimodal approaches. Lately, these efforts have focused on exploring more complex forms of multimodal strategies, as does human-human interaction. In addition, many applications have emerged, such as social networks that use multimodal information, for instance, videos, photos, text and audio segments, and, in response, the need to analyze this information to continue improving human-computer interaction [22, 27]. Furthermore, it can also be used to diagnose emotional and behavioral disorders [104]. Facial expression and emotional speech are the most natural and efficient ways to convey emotions [22, 27, 89]. However, most systems based on audio-visual features show low accuracy rates. In this context, we propose a novel method for enhancing the bimodal emotion recognition rate.

Emotion recognition has been extensively used in several fields of knowledge. Emotion recognition is continuously investigated to assist in diagnosing and monitoring treatment of mental (anxiety and anorexia, for example) and genetic disorders (e.g., autism). Previous studies have shown that deficiencies in recognizing emotions are indicative of various mental disorders as well as genetic disorders [100, 101]. In addition, several studies prove that there are phenotypes that help medical professionals diagnose some disorders [1, 18, 20, 33, 39, 48, 50, 65]. In Down syndrome, for instance, there are facial phenotypes, such as flat and wide face, slanting eyes, epicanthic eye fold, short nose, which contribute to the detection of Down syndrome [33, 48]. Therefore, facial expression features may have the potential to assist in the recognition of these types of disorders.

1.2 Objectives and Contributions

The main purpose of this work is to analyze and develop an emotion recognition system based on facial expressions. Thus, we aim to explore new methods and algorithms based on hand-crafted and deep features that contribute to the automatic emotion recognition in images or videos. We also examine new channels (such as audio features) that reinforce the improvement of this task. In addition, we analyze and apply facial expression features to solve genetic disorders that present facial and cranial abnormalities.

In this research work, we introduce an effective methodology for the recognition of emotions in occluded facial expression. Furthermore, new hand-crafted features for dynamic facial expression recognition and an efficient strategy for audio-visual emotion recognition are introduced. We also propose an approach to the recognition of Down syndrome, as well as the recognition of other genetic disorders with competitive accuracy rates. The assessment of the developed approaches is conducted on public data sets.

1.3 Research Questions

In this dissertation, we explore various techniques to promote improved accuracy in facial expression recognition problems. The main research questions formulated in this work are listed as follows:

1. How can we use audio and visual features to enhance the multimodal emotion recognition approach?
2. How can we apply facial expression patterns for recognizing genetic syndromes?
3. Are deep features suitable to improve facial expression recognition and genetic disorder recognition rates?

We focus on the proposition of methodologies that assist in the implementation of accurate systems of automatic emotion recognition, as well as how facial expression patterns help in the development of approaches to genetic disorder recognition.

1.4 List of Publications

Some papers generated from the results directly related to the research topic of this doctoral dissertation are:

- J.Y.R. Cornejo, H. Pedrini. *Recognition of Occluded Facial Expressions based on CENTRIST Features.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Shanghai, China, pp. 1298-1302, March 20-25, 2016.
- J.Y.R. Cornejo, H. Pedrini, A. Machado-Lima, F.L.S. Nunes. *Down Syndrome Detection based on Facial Features using a Geometric Descriptor.* Journal of Medical Imaging, vol. 4, n. 4, pp. 044008_1-044008_6, October-December 2017.
- J.Y.R. Cornejo, H. Pedrini. *Recognition of Genetic Disorders Based on Deep Features and Geometric Representation.* 23rd Iberoamerican Congress on Pattern Recognition (CIARP). Madrid, Spain, pp. 665-672, November 19-22, 2018.
- J.Y.R. Cornejo, H. Pedrini. *Dynamic Facial Expression Recognition Based on Visual Rhythms and Motion History Images.* 15th International Conference on Machine Learning and Data Mining (MLDM). New York-NY, USA, pp. 599-613, July 20-24, 2019.
- J.Y.R. Cornejo, H. Pedrini. *Audio-Visual Emotion Recognition using a Hybrid Deep Convolutional Neural Network based on Census Transform.* IEEE International Conference on Systems, Man, and Cybernetics (SMC). Bari, Italy, pp. 3396-3402, October 06-09, 2019.

While conducting doctoral research, other topics were explored that are not directly related to this dissertation. Most include the analysis and evaluation of discriminative features for facial expression. A list of papers published on topics complementary to those studied is presented as follows:

- J.Y.R. Cornejo, H. Pedrini. *Automatic Fruit and Vegetable Recognition based on CENTRIST and Color Representation.* Lecture Notes in Computer Science, vol. 10125, pp. 76-83, Springer-Verlag. Paper presented in the 21st Iberoamerican Congress on Pattern Recognition (CIARP). Lima, Peru, November 08-11, 2016.

- J.Y.R. Cornejo, H. Pedrini. *Emotion Recognition Based on Occluded Facial Expressions.* Lecture Notes in Computational Vision and Biomechanics, vol. 10484, pp. 309-319, Springer International Publishing. Paper presented in 19th International Conference on Image Analysis and Processing (ICIAP). Catania, Italy, September 11-15, 2017.

- J.Y.R. Cornejo, H. Pedrini. *Emotion Recognition From Occluded Facial Expressions using Weber Local Descriptor.* 25th International Conference on Systems, Signals and Image Processing (IWSSIP). Maribor, Slovenia, pp. 1-5, June 20-22, 2018.

- J.Y.R. Cornejo, H. Pedrini. *Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks.* 18th IEEE International Conference on Machine Learning and Applications (ICMLA). Boca Raton-FL, USA, pp. 111-117, December 16-19, 2019.

1.5 Text Organization

The text of this dissertation is organized as a compilation of papers published in international scientific venues in the fields of Image Processing and Machine Learning.

In Chapter 2, we propose a method for recognizing emotions in occluded facial expressions through static images. This approach is based on Census Transform Histogram features and follows a Robust Principal Component Analysis approach to occluded facial expression reconstruction. It is evaluated on two public data sets, achieving competitive accuracy rates for occluded and non-occluded facial expression images.

In Chapter 3, we present a method for recognizing dynamic facial expressions through Visual Rhythms and Motion History Images that encode video appearance, shape, and motion information. Texture analysis is performed through several techniques, such as Weber Local Descriptor, Census Transform Histogram, Histogram of Oriented Gradients, and Gray-Level Co-Occurrence Matrix. In addition, experiments are conducted to show the relevance of upper and lower facial expression regions.

Due to a demand to explore and analyze other relevant channels that may contribute to improve the accuracy of automatic emotion recognition, we propose a bimodal approach to recognizing emotions using audio and visual data. In Chapter 4, we describe a method for recognizing emotion in audio-visual videos using a hybrid deep convolutional neural network. An audio neural network based on Log Mel-spectrogram is presented and evaluated, whereas a visual neural network based on two facial expression representations, Census Transform and Contrast-limited Adaptive Histogram Equalization, is described and discussed. Experiments are conducted on three public data sets.

In Chapters 5 and 6, we apply facial expression features to address two genetic disorder recognition problems. In Chapter 5, we describe a method for Down syndrome recognition in facial photographs based on facial geometry features. In Chapter 6, we present a method for

genetic disorder recognition based primarily on deep features, as well as ORB and geometric features.

Chapter 2

Recognition of Occluded Facial Expressions based on CENTRIST Features

Abstract

Emotion recognition based on facial expressions plays an important role in numerous applications, such as affective computing, behavior prediction, human-computer interactions, psychological health services, interpersonal relations, and social monitoring. In this work, we describe and analyze an emotion recognition system based on facial expressions robust to occlusions through Census Transform Histogram (CENTRIST) features. Initially, occluded facial regions are reconstructed by applying Robust Principal Component Analysis (RPCA). CENTRIST features are extracted from the facial expression representation, as well as Local Binary Patterns (LBP), Local Gradient Coding (LGC) and an extended Local Gradient Coding (LGC-HD). Then, the feature vector is reduced through Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). For facial expression recognition, K-nearest neighbor (KNN) and Support Vector Machine (SVM) classifiers are applied and tested. Experimental results on two public data sets demonstrated that the CENTRIST representation achieved competitive accuracy rates for occluded and non-occluded facial expressions compared to other state-of-the-art approaches available in the literature.

2.1 Introduction

Although human emotion has been recently investigated in several knowledge areas, interest in emotions has its roots in Charles Darwin's pioneering studies [36]. He conjectured that emotional expression is universal, that is, independent on culture, race or gender.

Emotion can be defined as a subjective experience or physiological reaction of human beings, which can be expressed through several forms, such as body movement, voice intonation, facial expressions, and cardiac rhythm [133]. A facial expression is comprised of one or more facial musculature movements, which is functionally the same for newborns and adults. Facial expressions [105] are considered a universal and non-verbal communication mode that exhibit emotions in all human beings, allowing emotional information to be conveyed in a simple and

natural way. There is strong evidence of universal facial expressions for seven emotions [43]: anger, contempt, disgust, fear, happiness, sadness and surprise.

A challenging and interesting task is to automatically classify emotions through computer vision analysis [62, 70, 90, 114, 116, 122]. Over the past few years, there has been increasing attention to the development of robust devices that can help understand emotions and moods of human beings [118, 133, 134, 137]. Furthermore, research has been conducted to apply these devices in the development of tools for video surveillance systems, airport security, behavioral research, aggression detector for closed-circuit television, on-board emotion detector for drivers, among other applications. Therefore, facial expression recognition is an important issue for affective computing research [81].

Three main components are usually distinguished in automatic facial expression recognition systems: facial detection, feature extraction and representation of facial expression, and expression recognition. It is important to mention that most of the available facial expression recognition systems are based on data sets that do not reflect real and natural scenes. Furthermore, the majority of them do not address occlusions caused by hats, beard, sunglasses or scarves. Consequently, the omission of these factors during the training stage might affect the accuracy of the facial expression recognition process.

A novel and effective facial expression recognition methodology robust to occlusions is proposed and analyzed in this work. The method consists in five main stages. Initially, occluded facial regions are reconstructed through the Dual Approach [58], which is based on RPCA principles. Then, facial fiducial points are automatically detected. A set of features is extracted from the facial expressions, including the Census Transform Histogram. To reduce the dimensionality of the extracted features, the resulting descriptor is transformed to a lower dimensional space. Finally, the occluded facial expressions are recognized.

Experiments are conducted on two public data sets. The results obtained with the proposed method were compared to other approaches available in the literature. Our approach achieved high recognition accuracy rates for occluded and non-occluded images without demanding high computational resources.

The remainder of the paper is organized as follows. Section 2.2 briefly describes some important concepts related to the topic under investigation. Section 2.3 presents the methodology proposed in this work, describing the preprocessing, the facial expression reconstruction, the facial feature extraction, the feature reduction, as well as the classification process. Section 2.4 describes and analyzes the experimental results. Section 2.5 concludes the paper with final remarks and directions for future work.

2.2 Background

This section briefly describes some concepts related to the facial expression recognition problem addressed in our work.

Robust Principal Component Analysis (RPCA) [13] is a variant of the Principal Component Analysis (PCA) [73], which allows to recover a low-rank matrix A from a corrupted data matrix $D = A + E$, with gross but sparse errors E, through the solution of the following convex

optimization problem

$$\min_{A,E} \|A\|_* + \lambda |E|_1 \qquad \text{such that } D = A + E \tag{2.1}$$

where $\|A\|_*$ is the nuclear norm of a matrix A, $|E|_1$ represents the sum of the absolute values of E, and λ denotes a positive weighting parameter. For the facial expression reconstruction task, our work uses the Dual Approach [58] over the samples of the training set, which solves the RPCA problem (Equation 2.1) via its dual

$$\max_{Y} \langle D, Y \rangle \qquad \text{such that } J(Y) \leq 1 \tag{2.2}$$

where

$$\langle A, B \rangle = tr(A^T B) \qquad J(Y) = \max(\|Y\|_2, \lambda^{-1}|Y|_\infty) \tag{2.3}$$

such that $\|Y\|_2$ implies the spectral norm of a matrix Y, and $|Y|_\infty$ is the maximum absolute value of the matrix entries [58, 82].

Local Binary Patterns (LBP) [103] operator is a robust texture descriptor known by its discriminative power and computational simplicity, allowing real-time image processing. It is invariant to monotonic gray-scales transformations, e.g., caused by illumination variations. LBP encodes the difference between the central pixel value and its $n \times n$ neighborhood pixels sequentially, considering the result as a binary number.

CENsus TRansform hISTogram (CENTRIST) [136] is a visual descriptor, initially proposed for topological place and scene category recognition. This operator is based on Census Transform (CT), a non-parametric local transform. It is characterized for being a holistic representation, which captures the structural properties and has high generalization for categorization, suppressing detailed textural information. CENTRIST compares the central pixel intensity value with its 8-neighborhood. If the central pixel intensity value is higher than or equal to one of its neighbors, bit 1 is set in the corresponding location, otherwise bit 0 is set. The obtained 8 bits can be put together in any order, converting the resulting stream to a base-10 number, as the CT value of the current central pixel. Thus, CENTRIST is a histogram vector with 256 bins that represents the appearance frequency of CT.

Local Gradient Coding (LGC) [125], unlike LBP, considers the graylevel relationship between the central pixel and its neighbors. This operator describes graylevel distribution. The LGC algorithm, using a 3×3 neighborhood template as shown in Figure 2.1, is defined as

$$\begin{aligned} \text{LGC} &= s(g_1 - g_3)2^7 + s(g_4 - g_5)2^6 + s(g_6 - g_8)2^5 \\ &+ s(g_1 - g_6)2^4 + s(g_2 - g_7)2^3 + s(g_3 - g_8)2^2 \\ &+ s(g_1 - g_8)2^1 + s(g_3 - g_6)2^0 \end{aligned} \tag{2.4}$$

where

$$s(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The LGC algorithm compares the vertical, horizontal and diagonal gradients of its 8-neighbors and converts the binary stream into a base-10 number.

g_1	g_2	g_3
g_4	g_c	g_5
g_6	g_7	g_8

Figure 2.1: A 3×3 neighborhood template of LGC operator.

Local Gradient Coding of Horizontal and Diagonal gradient priority (LGC-HD) [125] is an extension of the LGC operator. The LGC-HD operator is capable of decreasing the feature space, reducing the processing time and improving the recognition accuracy. LGC-HD is defined as

$$\begin{aligned} \text{LGC-HD} &= s(g_1 - g_3)2^4 + s(g_4 - g_5)2^3 + s(g_6 - g_8)2^2 \\ &+ s(g_1 - g_8)2^1 + s(g_3 - g_6)2^0. \end{aligned} \tag{2.5}$$

Principal Component Analysis (PCA) [73] is a common technique for dimensionality reduction, which searches for a smaller set of new composite dimensions that represents a multidimensional feature space with minimum loss of information. PCA models the variance-covariance structure in a set of linearly uncorrelated variables, called principal components, through linear combinations from the original variables. A t-dimensional feature vector of N samples of the training set is projected onto a f-dimensional feature space, through PCA, resulting the new feature vector defined as

$$y_i = W_{PCA}^T x_i \qquad i = 1, \ldots, N \tag{2.6}$$

where W_{PCA}^T represents the linear transformations matrix, whose columns represents the eigenvectors associated with the largest eigenvalues of the scatter matrix S_T, which is expressed as

$$S_T = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T \tag{2.7}$$

where μ is the mean of all the samples of the training set [40].

Linear Discriminant Analysis (LDA) [55] is a classification method for producing models with high accuracy. LDA searches for a linear combination of features that best separates two or more classes, preserving as much of the discriminatory information of the class as possible. Two measures are defined for all the samples of the training set, one is the within-class scatter matrix S_W and the other is the between-class scatter matrix S_b

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (y_i^j - \mu_j)(y_i^j - \mu_j)^T \tag{2.8}$$

$$S_b = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T \tag{2.9}$$

where y_i^j is the ith sample of class j, μ_j denotes the mean of class j, μ represents the mean of all samples of all classes, C is the number of classes, and N_j is the number of samples per

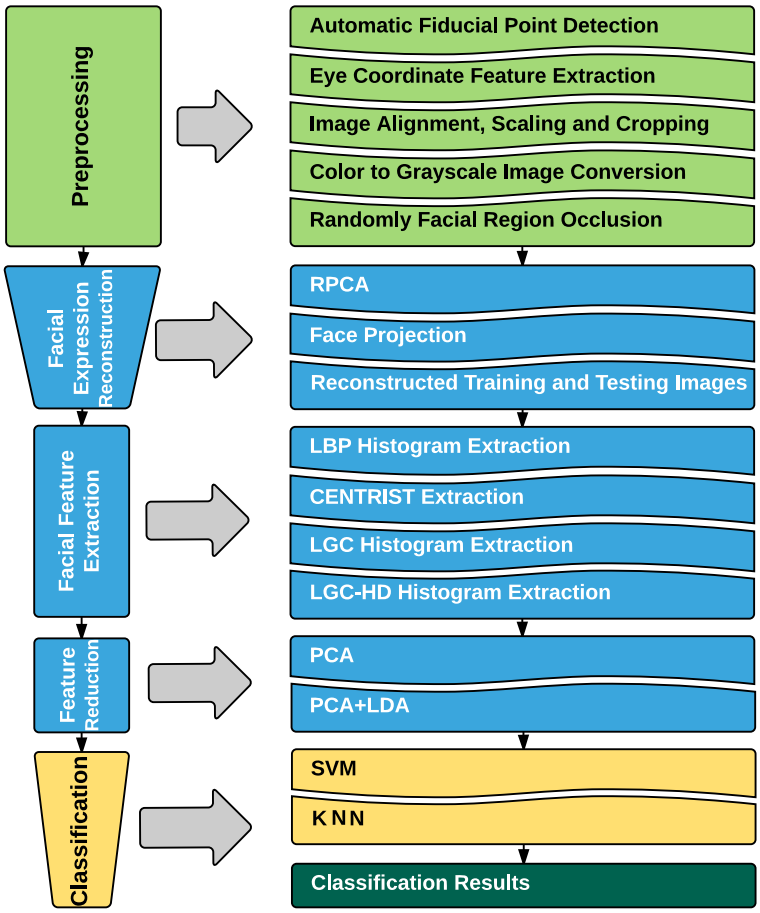


Figure 2.2: Diagram with the main steps of the facial expression recognition methodology.

class j [40].

It is known that the application of PCA followed by LDA can achieve higher recognition accuracy [40] than using only the individual approaches. The original t-dimensional space for the training set samples is projected onto an intermediate f-dimensional space using PCA. Then, the latter is projected onto a final g-dimensional space using LDA.

2.3 Methodology

The proposed methodology for facial expression recognition under occlusion presence is composed of five main stages: preprocessing, facial expression reconstruction, facial feature extraction, feature reduction and classification. These steps are illustrated in Figure 2.2 and described as follows.

Initially, a preprocessing step is applied to the images in order to provide aligned faces, uniform size and shape, and randomized occluded facial regions. This preprocessing task consists in the following seven steps: (i) automatic fiducial point detection through Chehra Face and Eyes Tracking Software [6]; (ii) extraction of eye coordinate features; (iii) image rotation to align the eye coordinates; (iv) image scaling proportionally to the minimum distance between the eyes; (v) face region cropping using a proper bounding rectangle; (vi) conversion of the color images to grayscale; (vii) addition of randomized black rectangles to simulate facial occlusion,

including left eye, right eye, two eyes, bottom left side of the face, bottom right side of the face or bottom side of the face, as illustrated in Figure 2.3.

Figure 2.3: Cropped images with occluded facial regions from the JAFFE data set.

Although PCA is widely used as a technique for reducing high-dimensional feature subspaces, it does not work well with grossly corrupted observations, e.g., occluded faces, variations of facial expressions, illumination conditions, image noise. On the other hand, RPCA [58] performs in a more effective way with missing data and outliers. RPCA is an extension of the classical PCA procedure and it has been demonstrated to be more robust, among other applications, for the reconstruction of occluded facial expressions [72] and to contribute in achieving better facial expression recognition accuracy [94]. As suggested in [29], RPCA approach is performed for facial expression reconstruction using 150 iterations and a parameter regularization [63]

$$\lambda = \frac{1}{\sqrt{\max(m, n)}} \tag{2.10}$$

where m and n are the size of matrix D.

After the facial expression reconstruction step, we projected all samples of the testing set onto the space generated by RPCA, such that all occluded facial regions set from the reconstructed faces, for both training and testing sets, were filled. Then, we applied the contrast-limited adaptive histogram equalization (CLAHE) over the reconstructed facial regions to enhance the image contrast. This process is illustrated in Figure 2.4.

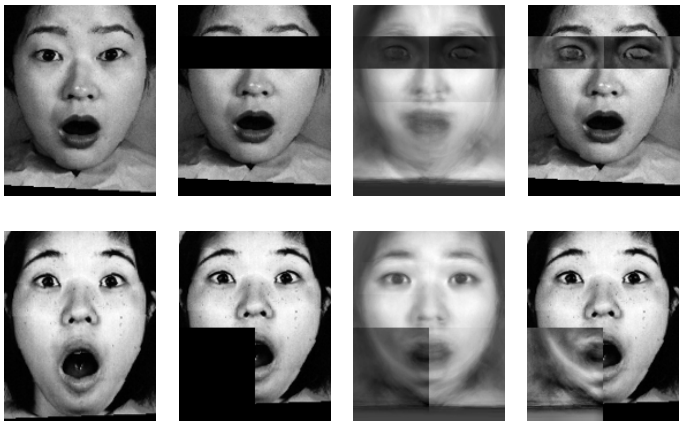


Figure 2.4: (a) Cropped images without occlusions from the JAFFE data set; (b) faces with occluded regions; (c) reconstructed faces; (d) filling the occluded facial regions from (c).

Four visual descriptor types were used in the facial expression recognition: Local Binary Patterns (LBP), Census Transform Histogram (CENTRIST), Local Gradient Coding (LGC) and

an extension of the Local Gradient Coding based on the principle of the horizontal and diagonal gradients (LGC-HD).

The LBP descriptor was applied over the entire image to extract the LBP code from each pixel. After obtaining an LBP labeled image and conducting several experiments, we divided the image into 63 (= 7×9) regions, as shown in Figure 2.5. LBP histograms were extracted from each generated region and concatenated all of them into one feature vector of length 16128 (= 256×63), which describes local texture and global shape information of the image.

(a) (b) (c)

Figure 2.5: (a) Cropped image from the JAFFE data set; (b) LBP image from (a); (c) LBP image is divided into 63 regions.

The images were also divided into 63 regions to extract the CENTRIST features from each region. The resulting vectors are concatenated, forming a vector of 16128 (= 256×63). CENTRIST is able to capture local structures of the image. Figure 2.6 shows a sample of a census transformed facial expression image.

(a) (b)

Figure 2.6: (a) Cropped image from CK+ database; (b) census transformed image from (a).

In a similar way to LBP feature extraction, LGC and LGC-HD were applied separately over the entire image. Figure 2.7 shows the obtained images after applying these operators. The images were then divided into 63 regions to extract their histograms. Hence, the resulting feature vectors have also 16128 (= 256×63) dimensions.

(a) (b) (c)

Figure 2.7: (a) Cropped image from CK+ database; (b) LGC image from (a); (c) LGC-HD image from (a).

Two techniques for feature reduction, PCA and LDA, were used sequentially, that is, each approach was applied individually for each feature vector. SVM and KNN classifiers were used to compare the recognition accuracy rates.

2.4 Experimental Results

Our method has been tested on the Cohn-Kanade (CK+) [87] data set and the Japanese Female Facial Expression (JAFFE) [88] data set. The CK data set is available in two versions. We used the second one (CK+) that contains 593 sequential images of posed and non-posed expressions from 123 subjects, categorized into one of seven facial expressions: anger, contempt, disgust, fear, happy, sadness and surprise. The CK+ data set also includes some metadata, such as 68 facial landmarks [87]. The JAFFE data set is a collection of 213 images from 10 Japanese female models who perform seven facial expressions: anger, disgust, fear, happiness, neutral, sadness and surprise [88].

For each data set, we randomly select 80% of samples of each class for the training set and the remaining 20% for the testing set. Then, 50% of the training set samples of each class were occluded and a similar procedure was applied to the testing set. We set 20 different randomized collections of occluded and non-occluded data to perform experiments for both data sets.

From these image collections, we conducted experiments using LBP, CENTRIST, LGC and LGC-HD operators through four methods: PCA+KNN, PCA+LDA+KNN, PCA+SVM and PCA+LDA+SVM. The results are shown in Tables 2.1 and 2.2, whose values represent the average facial expression recognition accuracy rate from the performed experiments. It is relevant to clarify that RPCA is always applied independently of the feature reduction and classification methods applied.

Table 2.1: Average accuracy rates (%) for non-occluded facial images from CK+ and JAFFE data sets.

	Method	LBP	CENTRIST	LGC	LGC-HD
CK+	PCA + KNN	43.74	**55.82**	37.70	36.81
	PCA + LDA + KNN	92.62	**93.66**	83.44	87.17
	PCA + SVM	77.17	**81.27**	71.80	69.78
	PCA + LDA + SVM	92.84	**94.10**	83.89	85.82
JAFFE	PCA + KNN	64.41	**87.50**	66.91	73.10
	PCA + LDA + KNN	**93.00**	91.60	83.10	88.34
	PCA + SVM	84.18	83.40	85.84	**86.31**
	PCA + LDA + SVM	**92.50**	92.00	82.03	87.51

From our experiments, we can observe that the CENTRIST operator allows to achieve more than 90% of accuracy rate for non-occluded image collections, even for occluded ones. Results from CK+ data set showed that the CENTRIST operator is always superior to the other feature extraction methods. From the JAFFE data set results, we can see that the LBP operator is slightly

Table 2.2: Average accuracy rates (%) for occluded facial images from CK+ and JAFFE data sets.

	Method	LBP	CENTRIST	LGC	LGC-HD
CK+	PCA + KNN	42.06	**54.04**	36.35	33.81
	PCA + LDA + KNN	88.06	**90.30**	78.06	80.18
	PCA + SVM	75.01	**78.51**	67.97	66.65
	PCA + LDA + SVM	88.44	**90.01**	78.51	79.33
JAFFE	PCA + KNN	45.84	**87.50**	66.91	74.06
	PCA + LDA + KNN	83.10	**91.60**	83.10	88.57
	PCA + SVM	70.60	83.40	85.84	**88.67**
	PCA + LDA + SVM	81.44	**92.00**	82.03	87.47

superior to CENTRIST operator for non-occluded images, whereas far better than the others. However, the CENTRIST operator is much better among the other texture operators for occluded images.

Furthermore, despite lower accuracy rates achieved by LGC and LGC-HD operators, we can notice that LGC-HD provides better results than using LGC operator. From the experiments conducted over occluded and non-occluded collections, we can see that following PCA+LDA method always provides higher accuracy rates than only applying PCA.

There are only few similar works available in the literature that consider random partial facial occlusions, especially on both training and testing stages. Then, we compared our approach against other state-of-the-art methods. Table 2.3 summarizes the best results reached by our approach, including the other methods available in the literature, on both data sets. It is possible to see that the proposed approach (CENTRIST+PCA+LDA+SVM), unlike other visual descriptors, achieves the best results for CK+ and JAFFE data sets under occlusion presence. Table 2.3 is sorted in descending order by occluded recognition accuracy rate.

Table 2.3: Comparison of average accuracy rates (%) for non-occluded (non-oc) and occluded facial expression images (oc) on CK+ and JAFFE data sets.

	Approach	Strategy	non-oc	oc
CK+	Proposed method	CENTRIST+PCA+LDA+SVM	94.10	90.01
	Proposed method	LBP+PCA+LDA+SVM	92.62	88.44
	Ramírez et al. [29]	Gabor+PCA+LDA+SVM	94.03	85.68
	Liu et al. [84]	Maximum Likelihood Estimation Sparse Representation	94.29	85.24
JAFFE	Proposed method	CENTRIST+PCA+LDA+SVM	92.00	92.00
	Liu et al. [84]	Maximum Likelihood Estimation Sparse Representation	93.42	86.73
	Proposed method	LBP+PCA+LDA+KNN	93.00	83.10
	Ramírez et al. [29]	Gabor+PCA+LDA+SVM	95.12	82.86
	Zhang et al. [147]	Gabor template and SVM	81.20	48.80

2.5 Final Considerations

In this work, we introduced the CENTRIST operator as a potential visual descriptor for emotion recognition. CENTRIST has proven to be robust to occluded and non-occluded facial expressions. Some advantage of this operator include its simple implementation, good performance and high computational speed. Furthermore, experimental results have shown that the use of the PCA+LDA approach increases the recognition rates significantly.

Despite the fact that the LGC and LGC-HD did not provide high accuracy for facial expression recognition, it is important to remark that LGC-HD is much superior than the LGC operator due to its improvement in the classifier recognition accuracy and its computational speed.

As directions for future work, we intend to explore new feature extraction methods, as well as improve the CENTRIST approach. We also plan to conduct experiments by considering real facial occlusions, e.g., sunglasses, scarves, facial hair, caps and beard, to be sure that the proposed approach is robust in real scenes. In addition, we pretend performing experiments for facial expression recognition in video scenes. Finally, we consider crucial the research on the development of an automatic occlusion detector for emotion recognition systems robust to occlusions.

Chapter 3

Dynamic Facial Expression Recognition Based on Visual Rhythms and Motion History Images

Abstract

Facial expressions provide important indications about human emotions. The development of an automatic emotion recognition system based on facial expressions has received significant attention due to its applications in several knowledge domains, such as affective computing, behavior prediction, security, biometrics, and human-computer interactions. In this work, we propose a novel and automatic methodology for emotion recognition from dynamic facial expressions. Our approach encodes video appearance, shape, and motion information as the fusion of Visual Rhythms (VR) and Motion History Images (MHI) for further describing them through texture analysis. These texture analysis methods include Weber Local Descriptor (WLD), Census Transform Histogram (CENTRIST), Histogram of Oriented Gradients (HOG), and Gray-Level Co-Occurrence Matrix (GLCM). For assessing our methodology, we conducted experiments on the Binghamton University 4D Facial Expression (BU-4DFE) data set, achieving accuracy improvements on partial facial regions. In summary, we demonstrate that the combination of Visual Rhythms and Motion History Images aids on the automatic recognition of dynamic facial expressions.

3.1 Introduction

Emotion is a physiological reaction or subjective experience of human beings, which can be expressed in audiovisual forms, such as speech, prosody, respiration sound, pupil size, facial expression, gesture, posture, among others. Emotions can also be expressed by biological changes, for example, heart rate, respiration rate, sweating, temperature, muscle tension, brain waves, and facial expressions. However, facial expressions are a universal and non-verbal communication channel that shows emotions in all human beings, allowing to communicate emotional information in an easier, simple and natural way. A facial expression consists of one or more facial musculature movements, which are functionally the same in adults and newborns [4].

There is a strong evidence of universal facial expressions for seven emotions [4, 43]: anger, contempt, disgust, fear, happiness, sadness, and surprise.

Recently, research has been conducted to develop robust devices that can help to understand emotions and moods of human beings. Computers should ideally generate an artificial emotional response to human expressions in a believable way. Investigations have been carried out to apply these devices in the development of automated tools for social mobile applications, social augmented reality, behavioral prediction, airport security, aggression detector, video surveillance, mapping of facial expression for human biometric systems, among other applications.

Furthermore, facial expression recognition plays an important role in affective computing research [109]. Facial expressions can also provide strong evidence about the cognitive state of human beings, such as physical state, pain, fatigue, attention degree. More recently, many applications have been proposed to improve human-computer interaction, such as on-board fatigue detector for drivers, stress detector, physical pain detector, attention assessment application in an online learning environment, or recommendation systems.

As a main contribution of this work, a novel dynamic facial expression recognition approach is proposed. The methodology consists of five main stages: (i) automatic detection of facial fiducial points and cropping of the upper and lower facial regions; (ii) characterization of facial regions from the video sequences through Visual Rhythms (VR) and Motion History Images (MHI); (iii) extraction of a set of features from the facial expression regions, including the Weber Local Descriptor (WLD) features; (iv) reduction of the extracted feature dimensionality, such that the resulting vector is transformed into a lower dimensional space; (v) recognition of the dynamic facial expressions.

The remainder of the paper is structured as follows. Section 3.2 presents the approach proposed in this work, including details on preprocessing, video characterization, feature extraction, feature reduction and dynamic facial expression classification. Experiments performed on the BU-4DFE data set are described and discussed in Section 3.3. Finally, some concluding remarks are presented in Section 3.4.

3.2 Dynamic Facial Expression Recognition Method

The proposed facial expression recognition method is explained in this section. The main stages of the method are illustrated in Figure 3.1 and described as follows.

3.2.1 Preprocessing

The video preprocessing stage is crucial for performing facial expression recognition. Hence, this stage is required to reduce the complexity of the subsequent steps, such as extraction and recognition. This phase is also fundamental towards the success of the emotion recognition task.

Initially, an automatic detection of facial fiducial points is performed over the first frame of the video sequences using the Dlib Facial Landmark Detector [75], which tracks 68 facial landmarks points. Then, the upper and lower facial regions are cropped by proper bounding rectangles, which were defined based on the distance between the facial fiducial points. The top bounding rectangle width is calculated as the distance between the outer end points of the eyebrows, whereas the height is determined by the maximum distance between the trisection

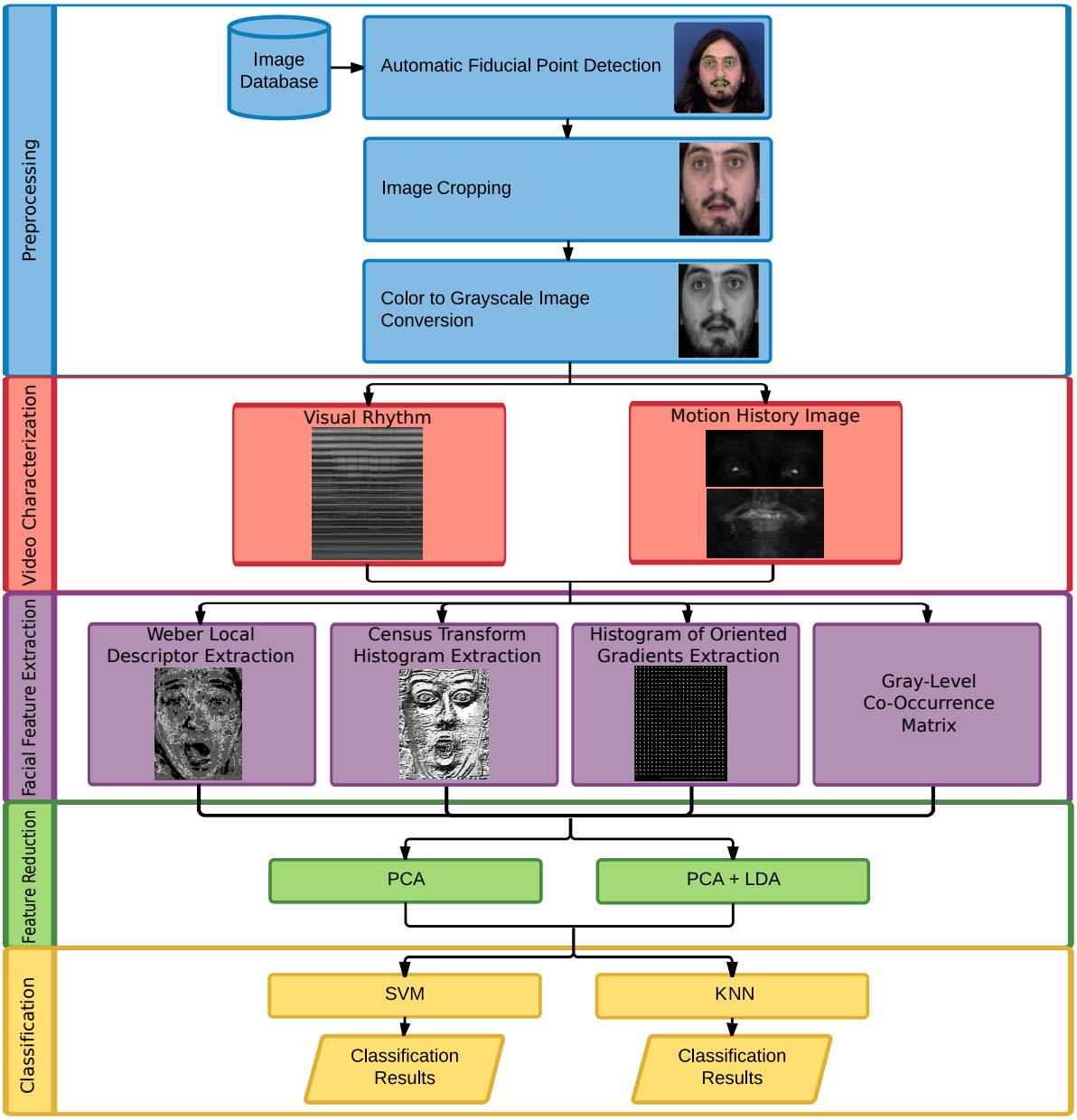


Figure 3.1: Diagram of the facial expression recognition methodology.

points of the eyebrow upper arcs and the bottom points of the lower eyelids. Figure 3.2.(b) depicts some upper facial region samples from the BU-4DFE data set. For the bottom bounding rectangle, the width is defined as the distance between the mouth corners and its height is set as the length between the top and bottom points of mouth on the midline. Some samples of the upper regions from the BU-4DFE data set are shown in Figure 3.2.(c).

After preprocessing the first frame, a facial landmark tracking based on a dense optical flow tracking algorithm [86] is executed on the next frames, whose upper and lower facial regions are also cropped using the bounding rectangles defined previously. Finally, all color video frames are converted into grayscale images.

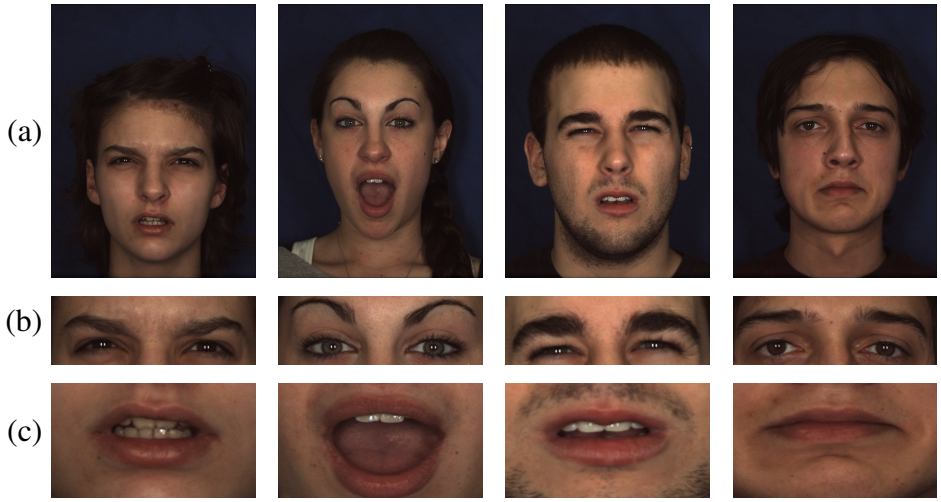


Figure 3.2: Cropped facial expression regions from BU-4DFE data set. From left to right, the emotions cataloged are anger, surprise, disgust, and sadness. (a) original video frame samples; (b) upper facial expression regions; (c) lower facial expression regions.

3.2.2 Video Characterization

The Visual Rhythms and Motion History Images (MHI), applied for characterizing the sequences of facial expression videos, are briefly described as follows.

Visual Rhythms (VR)

This strategy [17] is employed over the upper and lower facial regions from the video sequences. This technique allows the extraction of spatio-temporal information and summarizes the video sequences into one single image. Visual Rhythm (VR) has been used for diverse computer vision tasks [110, 126], such as action recognition, video caption detection and facial spoofing attack detection. VR consists of joining pixel slices from all the frames of a video. The choice about how to form the pixel slices is arbitrary, which depends on the type of information desired. The pixel set can be extracted from the entire image or just from some image regions. For instance, a video V in the domain $2D + t$ has t frames of dimension $H \times W = N$ pixels, where pixels of each frame t can be concatenated following any orientation pattern - that is, horizontal, vertical, diagonal - to form an image of dimension $N \times t$ pixels, called visual rhythm.

After executing several experiments following different orientation patterns for concatenating the pixels, we decided to work with a horizontal orientation pattern, since it allowed to obtain higher accuracy rates for the emotion recognition task. Therefore, the visual rhythm is generated for both upper and lower facial regions. The top of Figure 3.3 shows a sample frame of the upper facial region sequences, where the bottom displays the visual rhythm image generated from the entire upper region sequences.

Motion History Images (MHI)

This technique [9, 38] is a temporal template that describes where the motion occurs and how the object is moving. MHI has been largely adopted in image processing tasks, such as facial

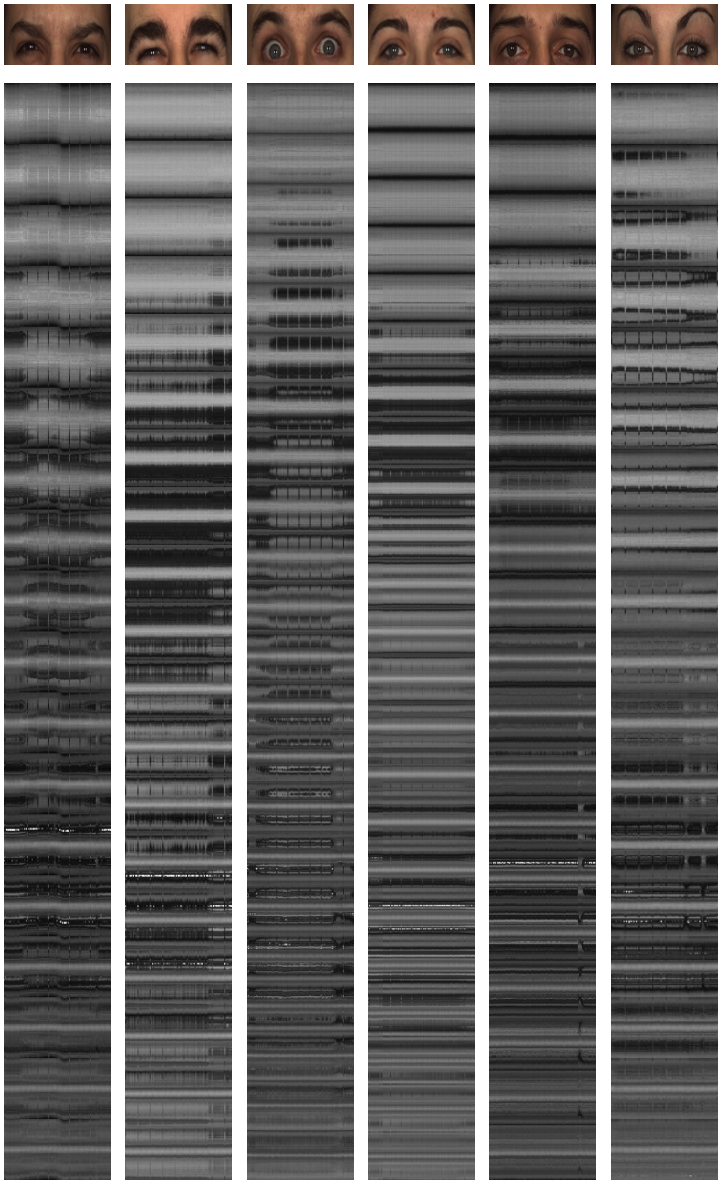
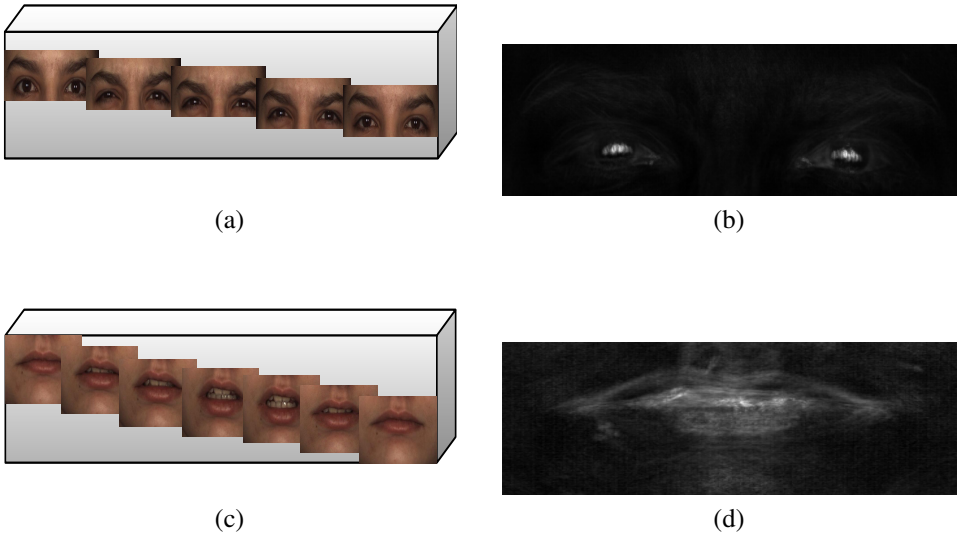


Figure 3.3: Visual rhythm representations from the top facial region sequences for different emotions. (a) anger; (b) disgust; (c) fear; (d) happiness; (e) sadness; (f) surprise.

action unit detection [51, 128], action recognition [67], posture recognition [139], among others. MHI extracts silhouettes through background subtraction and adds the differences between video subsequences. MHI represents the motion history from stacking the video sequence into a single image.

Therefore, a motion history image is generated from the sequences of cropped facial regions. Figures 3.4(a) and 3.4(c) show a sequence of facial regions (eyes and mouth), whereas Figures 3.4(b) and 3.4(d) exhibit their respective motion history images.



(a) (b)

(c) (d)

Figure 3.4: (a)(c) Examples of video sequences from the upper and lower facial region of a sample from the BU-4DFE data set; (b)(d) MHI generated from the video sequence in (a)(c).

3.2.3 Facial Feature Extraction

Four feature extraction techniques - Weber Local Descriptor (WLD), Census Transform Histogram (CENTRIST), Histogram of Oriented Gradients (HOG) and Grey Level Co-occurrence Matrices (GLCM) - are used for facial expression recognition.

Weber Local Descriptor (WLD)

This method [16] is a powerful image descriptor inspired by the Weber's Law, a historically important psychological law which states that the change in a stimulus that will be just noticeable is a constant ratio of the original stimulus. The WLD descriptor has been used for several image processing and computer vision tasks, such as face recognition, emotion recognition [2, 83, 99], gender classification, and image forgery detection.

The WLD descriptor is defined by two components, a differential excitation and an orientation, which are both calculated for each pixel of a given image. Differential excitation denotes the intensity differences between a central pixel and its neighbors. This ratio is set by the sum of the intensity differences, divided by the central pixel, defined as

$$\xi(x_c) = \arctan\left[\sum_i \frac{(x_i - x_c)}{x_c}\right] \tag{3.1}$$

where $\xi(x_c)$ is the differential excitation, x_c represents each pixel with its pixel neighbors x_i, and $\arctan$ function is used since it can restrict the result value in order to prevent it from increasing or decreasing too much when the input data becomes greater or smaller. Orientation describes the ratio of the change in the horizontal and vertical direction of a pixel. This is calculated over x and y gradients generated by $[1, 0, 1]$ and $[-1, 0, 1]^T$ filters that operate at the same differential

excitation neighborhood. It can be computed as follows:

$$\theta(x_c) = \arctan\left[\frac{G_x}{G_y}\right] \tag{3.2}$$

where G_i is the gradient calculated by the pixel difference in direction $i \in \{x, y\}$. Therefore, the differential excitation ξ and orientation θ are linearly quantized into T dominant differential magnitudes and orientations, respectively. Then, in order to build the WLD histogram, the differential excitation ξ and orientation θ are computed for each pixel, generating T subhistograms. Each subhistogram is divided into M segments, which is evenly divided into S bins.

The WLD descriptor [16] is applied over the VR and MHI representations obtained from the upper and lower facial region sequences, as shown in Figures 3.3 and 3.4, to extract textural features. After executing several experiments with different WLD parameters, we set: $(T) = 8$ dominant orientations, $(M) = 6$ subhistograms, and $(S) = 5$ bins. We split each image into 7 columns and 9 rows, which results in 63 subregions, as shown in Figure 3.5(c). For each facial subregion, we extracted WLD histograms and concatenated all of them into one with a length of 15600 (= $M \times T \times S \times 63$). The intensity values of the WLD image describe the differential excitation information of the facial expression image. The entire process is illustrated in Figure 3.5.

(a) (b) (c)

Figure 3.5: (a) MHI representation from upper facial region sequences extracted from BU-4DFE database; (b) filtered image by WLD from image (a); (c) WLD image divided into 63 regions.

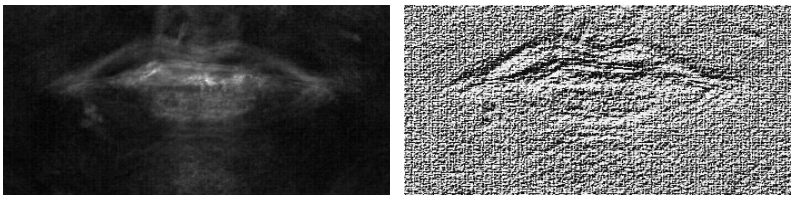
Census Transform Histogram (CENTRIST)

This visual descriptor was originally purposed for topological place recognition and scene categorization [135, 136]. It has also been applied to pedestrian detection [68, 69] and emotion recognition [24] tasks. CENTRIST is based on Census Transform (CT), which is a non-parametric local transform. This descriptor is a holistic representation, which captures general structural properties besides retains the local structures, while suppressing detailed textural information.

CT compares the center pixel intensity value with its eight neighboring pixels. If the value of the center pixel intensity is higher than or equal to one of its 8-neighborhood, bit 1 is set in the corresponding location, otherwise, bit 0 is set. The generated 8 bits can be joined in any order, which is converted to a base-10 number. This is the CT value of the current center pixel. Thus, a histogram of CT values is computed with 256 bins, which is called CENTRIST.

In a similar way to WLD feature extraction, CENTRIST was applied over the representations of VR and MHI from the upper and lower facial region sequences. The images were also divided

into 63 regions in order to extract the CENTRIST features from each subregion. For each image, the resulting CENTRIST vectors are concatenated, forming a vector of length $256 \times 63 = 16128$. Figure 3.6 shows a sample of a census transformed MHI representation.



(a) (b)

Figure 3.6: (a) MHI representation from lower facial region sequences extracted from BU-4DFE database; (b) census transformed image from (a).

Histogram of Oriented Gradients (HOG)

This feature descriptor [34] has been widely applied to object detection and recognition tasks [41, 124]. HOG descriptor represents local patterns at every pixel using gradient orientations. The HOG algorithm divides an image into small connected regions, called cells. For each cell, a histogram of oriented gradients is computed. Groups of adjacent cells correspond to spatial regions, called blocks. Cell grouping forms the basis for combination and normalization of histograms. The final HOG descriptor is achieved by grouping all normalized groups of histograms into a single feature vector.

We employ HOG for extracting features from the MHI representations, using the following parameter set: block size (bs) = 2×2, cell size (cs) = 8×8, block overlap (bo) = 1×1, bin number (bn) = 9, and block normalization (bn) = L2. The HOG feature vector encodes local shape information from regions within an image. The length N of the feature vector for an image I is expressed as

$$N = bpi * bs * bn \tag{3.3}$$

where $bpi = ((size(I)/cs) - bs)/((bs - bo) + 1)$ and $size(.)$ is the matrix dimension.

Gray-Level Co-Occurrence Matrix (GLCM)

This method [64] characterizes the texture of the image by describing how often pixel pairs with specific values and in a specific orientation θ and distance d occur in an image. Then a GLCM is created, from which statistical measures, such as contrast, correlation, among others, are extracted.

We use GLCM to extract features from the Visual Rhythm representations. Then, we calculated the co-occurrence matrix for four orientations $\theta = \{0, \pi/4, \pi/2, 3\pi/4\}$ and one distance $d = 1$. In order to summarize the textural information, for each generated matrix, we computed the following five measures: contrast, dissimilarity, homogeneity, energy, and correlation. The length of the resulting feature vector is $\theta \times d \times 5 = 20$.

3.2.4 Feature Reduction

Feature dimensionality reduction techniques are applied over the generated feature vector. This step alters the original data representation, such that the new feature vector has a smaller number of dimensions compared to the initial representation, preserving the most representative features.

We follow two approaches PCA and PCA+LDA for performing feature reduction. Firstly, PCA is employed over the resulted feature vector set, obtaining the principal feature vectors. Finally, Linear Discriminant Analysis (LDA) is applied over the feature vector, after its reduction with Principal Component Analysis (PCA), which generates a new feature space.

3.2.5 Classification

During this stage, we used Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifiers to compare the dynamic facial expression recognition accuracy rates. After performing dimensionality reduction, the classifiers are trained with reduced feature vectors. We employ k-fold cross-validation to validate our methodology. Furthermore, we apply a grid search strategy to look for a proper combination of parameters that provides the most accurate classification model.

3.3 Experimental Results

Experiments were conducted on the Binghamton University 4D Facial Expression (BU-4DFE) [143] data set to evaluate the effectiveness of the proposed methodology. This data set is composed of 606 3D videos performed by 101 subjects, of which 58 are female and 43 are male models from different ethnics including Asian, Black, Hispanic and White. For each subject, there are six model sequences categorized into one of six facial expressions: anger, disgust, happiness, fear, sadness, and surprise. Each video contains about 100 frames, considering posed and non-posed (spontaneous) expressions. Each 3D model of a 3D video presents a resolution of approximately 35,000 vertices. While texture videos have a resolution of 1040×1392 pixels/frame. The 3D videos were captured at a rate of 25 frames per second (fps).

We perform experiments using different strategies, as shown in Figure 3.1, that is, applying Visual Rhythms and Motion History Images over cropped facial regions, extracting features using Weber Local Descriptor, Census Transform Histogram, Gray-Level Co-Occurrence Matrix and Histogram of Oriented Gradients, and classifying through KNN and SVM algorithms. Tables 3.1 to 3.5 present the average facial expression recognition accuracy rates. It is worth mentioning that all the results were obtained using the PCA + LDA approach, since it increases the recognition accuracy and preserves the most representative and discriminative information, as reported in the literature [29, 40]. Our method was validated through a 10-fold cross-validation protocol.

As it can be observed from Tables 3.1 and 3.2, the recognition rate using WLD features is superior to the accuracy obtained by the other feature extraction methods. The WLD operator allows to achieve higher accuracy rates for both video characterization techniques, that is, Visual Rhythms and Motion History Images, which were applied over the cropped face, which refers to the entire face without background.

Table 3.1: Average accuracy rates (%) using Visual Rhythms following a diagonal orientation over the cropped facial images through GLCM, CENTRIST and WLD operators.

		Feature Descriptor		
		GLCM	CENTRIST	WLD
Recognition	KNN	35.80	47.20	50.20
Method	SVM	42.00	50.60	54.44

Table 3.2: Average accuracy rates (%) using Motion History Images over the cropped facial images through HOG, CENTRIST and WLD operators.

		Feature Descriptor		
		HOG	CENTRIST	WLD
Recognition	KNN	37.20	41.47	42.30
Method	SVM	40.20	42.29	44.50

Table 3.3: Average accuracy rates (%) obtained with Motion History Images over different facial regions using the WLD operator.

		Facial Regions			
		Cropped Face	Upper	Lower	Upper & Bottom
Recognition	KNN	42.30	41.80	59.00	67.40
Method	SVM	44.50	45.00	63.50	69.40

Since WLD performed better than other descriptors, we conducted experiments over different facial regions. The results are shown in Table 3.3. Employing the cropped face provides the lowest accuracy. However, using both upper and lower facial expression regions allows increasing recognition accuracy rates notably. We can also observe that the emotion recognition rate considering the lower region is 20% more accurate than using only the upper region. This result could contribute to a significant improvement in terms of recognition accuracy rate on facial expression images with occlusions, such as glasses or facial hair.

After concluding that using the upper and lower facial regions allows to reach a higher accuracy rate, we decide to generate visual rhythms from the upper and lower regions of the video sequences. Then, Table 3.4 presents the results of the experiments performed over the visual rhythms following different orientations, i.e., diagonal, vertical, and horizontal. The highest accuracy rates were achieved using a horizontal orientation pattern, whereas lower accuracy rates were reached using a diagonal orientation. This is an interesting fact since diagonal orientation has been shown to be appropriate and effective for action detection [97]. However, for the dynamic facial expression recognition task, using a horizontal pattern increased its effectiveness by 5% compared to the use of a diagonal orientation.

Table 3.4: Average accuracy rates (%) for the Visual Rhythm with different orientation patterns over the upper and lower facial regions using the WLD operator.

		Visual Rhythm Orientation		
		Diagonal	Vertical	Horizontal
Recognition	KNN	67.70	65.40	68.40
Method	SVM	66.60	69.00	71.30

Table 3.5: Average accuracy rates (%) employing Motion History Images and Visual Rhythms (horizontal orientation) over the upper and lower facial regions, using the WLD operator.

		Strategy		
		MHI	Visual Rhythm	MHI & Visual Rhythm
Recognition	KNN	67.40	68.30	74.00
Method	SVM	69.40	71.30	76.40

Table 3.5 summarizes the highest accuracy rates obtained using Visual Rhythms with a horizontal orientation and Motion History Images. The features were extracted using WLD. It is possible to see that there is no significant difference between dynamic facial expression recognition rates using Visual Rhythm and MHI representations independently. However, when combining both video characterization methods, there is a positive impact on the recognition accuracy rate.

We compared our method to others available in the literature considering different experimental settings. A comparison of the results is shown in Table 3.6, which is sorted in descending order by accuracy rate. Table 3.6 also provides information about the number of classes, number of subjects, whether or not the full sequence was employed, and the validation protocol.

Sandbach et al. [117] achieved a recognition accuracy rate of 64.60% on six facial expressions, while a higher accuracy (81.93%) on only three classes. Reale et al. [112] obtained a recognition rate of 71.50%, but their method requires a manual intervention for selecting the onset frame. Canavan et al. [12] reached an accuracy of 75.10% for recognizing low-intensity facial expressions. Berretti et al. [8] achieved an accuracy of 76.30%, nevertheless, their proposal depends on the landmark detection precision.

Yao et al. [141] reached a recognition rate of 86.87% based on deep neural networks. Zhen et al. [154] obtained a rate of 87.06% using the facial musculature. Amor et al. [5] reported the highest accuracy rate of 93.21%, using Hidden Markov Model (HMM) and a sliding window to sample frames. However, this approach requires a manual detection of key frames, such as the onset and apex frames. Approaches proposed by Reale et al. [112], Sandbach et al. [117], Jeni et al. [71] and Yao et al. [142] also face this issue. In order to avoid an unprecise and manual selection, we use all video sequences to recognize facial expression, as the following works also do [8, 138, 141, 154].

Table 3.6: Comparison to the state of the art on the BU-4DFE data set.

Method	Classes	Subjects	Full Sequence	Validation Protocol	Accuracy (%)
Amor et al. [5]	6	60	No	10-fold CV	93.21
Yao et al. [142]	6	60	No	06-fold CV	90.00
Zhen et al. [154]	6	60	Yes	10-fold CV	87.06
Yao et al. [141]	6	60	Yes	10-fold CV	86.87
Xue et al. [138]	6	60	Yes	10-fold CV	78.80
Jeni et al. [71]	6	-	No	LOSOCV	78.18
Proposed Method	6	101	Yes	10-fold CV	76.40
Berretti et al. [8]	6	60	Yes	10-fold CV	76.30
Canavan et al. [12]	6	60	-	10-fold CV	75.10
Fang et al. [52]	6	100	-	10-fold CV	74.63
Reale et al. [112]	6	100	No	-	71.50
Sandbach et al. [117]	6	60	No	06-fold CV	64.60

3.4 Final Considerations

In our work, we introduced the Visual Rhythm representation as a potential video characterization technique for dynamic emotion recognition. Some advantages of Visual Rhythms include their simple and fast implementation, as well as high computational performance. Furthermore, experimental results have demonstrated that the combination of Visual Rhythms and Motion History Images positively influences recognition rates.

It is also important to remark the relevance of using the lower facial expression region for emotion recognition since it contains sufficient discriminative information. Moreover, the combination of the upper and lower regions presents even more representative information.

Chapter 4

Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform

Abstract

Over the last years, recognition of emotions based on multimodal channels has received increasing attention from the scientific community. Many application fields can benefit from multimodal emotion recognition, such as human-computer interactions, educational software, behavior prediction, interpersonal relations. Speech and facial expressions are two natural and effective ways to express emotions in human-human interaction. In this work, we introduce a hybrid deep convolutional neural network to extract audio and visual features from videos. Initially, for extracting audio data, we transform the audio signal into an image representation as input to a 2D-Convolutional Neural Network (CNN). For extracting visual data, we introduce a Census-Transform (CT) based on CNN. Then, we fuse both audio and visual features, reducing them through Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Finally, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR) and Gaussian Naïve Bayes (GNB) classifiers are employed for emotion recognition. Experimental results on RML, eNTERFACE05 and BAUM-1s data sets demonstrated that our model reached competitive recognition rates compared to other state-of-the-art approaches.

4.1 Introduction

In recent years, emotion recognition [32, 45, 98] has gained increasing interest since it plays a crucial role in developing machines capable of recognizing, understanding, expressing, and reacting to emotions. Furthermore, it means a key for enhancing the interactive experience of human-computer interaction (HCI) applications in different knowledgeable fields, which can be achieved by following a multimodal emotion recognition approach, as natural human-human interaction (HHI) does. It emerges the need for exploring multimodal approaches in order to analyze which channels afford valuable information for automatic emotion recognition. Human emotion can be expressed through several audio-visual channels and internal biological changes,

such as voice intonation, facial expressions, temperature, body gestures, pupil size, brain signals, and heart rate.

Facial and vocal expressions have shown to be a natural, efficient and simple form to express emotions in human interactions. Facial expressions are a universal and non-verbal form to communicate emotions, existing proof of universal facial expressions for seven emotions: contempt, anger, fear, disgust, happiness, sadness and surprise [43, 140]. Facial expressions are independent on gender, ethnicity, and age. A facial expression consists of motion of one or more facial muscles (functionally the same for adults and newborns), whereas the vocal expression depends on the tone and energy.

Speech carries linguistic and paralinguistic information that expresses emotions. However, some studies [3, 57] show that linguistic messages are fairly unreliable due to the difficulty to associate and predict the person words with the emotion that intents to express since it depends on the semantic context and language [145]. In the case of paralinguistic messages, it is stated that prosodic, acoustic and voice quality features provide relevant emotion information. Some of these features are pitch, speaking rate, formant frequencies, spectral tilt, energy, audible speech duration [89].

Lately, several multimodal applications based on automatic emotion recognition have been investigated due to their high potential in intelligent systems for diverse fields, such as behavioral prediction, online learning, social applications, biometric systems, video surveillance, and healthcare. For example, in an online learning system, the behavior of the software could be dynamically modified according to the user feelings. It means that if the user is bored, the application could increase the pace and the difficulty level; otherwise, it could decrease them to keep the user motivated for more extended periods, improving learning progress [61]. In driver behavioral prediction, if a driver is fatigued or annoyed, the system could play relaxing or loud music to calm or put her or him on alert, depending on the case, for avoiding to put driver life in risk and car accidents. In autism healthcare, it can aid to children with autism spectrum disorder to receive real-time feedback, while they are interacting with their caretakers since autism patients typically have difficulty to understand other people emotions [123].

More recently, there have been unimodal investigations focused on facial expression recognition [15, 24, 26, 29], as well as on audio emotion recognition [80, 127, 131]. For facial expression recognition, the feature extraction can be categorized into geometric and appearance features. The geometric features [29, 102] include, for instance, the facial landmark location, such as the corners of mouth, eyes, and eyebrows; facial region shapes, i.e., nose, mouth, eyes; and distances and angles computed from the facial fiducial points. For facial expression recognition in videos, Noroozi et al. [102] computed 60 geometrical features, including distances and angles. For static images, Ramirez et al. [29] used 22 geometric features, which comprises a set of selected points and distances. The appearance features consist of extracting features from the whole face or a specific facial region, using texture filters, such as Gabor wavelets and Local Binary Patterns (LBP) [24, 74]. Kaya et al. [74] extracted and compared Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Local Phase Quantization (LPQ), LBP and its extension Local Gabor Binary Patterns, and deep features for videos. In the case of static images, Cornejo et al. [24] performed experiments by means of LBP, Census Transform Histogram (CENTRIST), Local Gradient Coding (LGC) and its extension Local Gradient Coding based on the principle of the horizontal and diagonal gradients (LGC-HD).

Regarding audio emotion recognition [80, 127, 131], audio features can be divided into prosody, voice quality and spectral features. Prosody features are composed of pitch, intensity, energy, duration, among others. Formants, jitter (pitch irregularity), shimmer (amplitude irregularity), harmonics to noise ratio, and spectral energy distribution are some samples of voice quality features. The most common spectral feature is the Mel-Frequency Cepstrum Coefficients (MFCC). Zhalehpour et al. [146] extracted MFCC and relative spectral features (RASTA) based on perceptual linear prediction (PLP). Noroozi et al. [102] employed MFCC, Filter Bank Energies (FBE), and statistics and acoustics features. Ma [89] and Zhang et al. [149] used audio features based on Mel-spectrogram. Both audio and visual unimodal approaches performed emotion recognition through machine learning techniques, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Long Short-Term Memory (LSTM), CNN, and so on.

There has also been some research related to multimodal emotion recognition [7, 46, 47, 59, 89, 132, 148, 149], which usually involves the combination of video and speech due to these can be captured in a non-invasive mode and be more expressive. For example, Zhang et al. [149] proposed a hybrid deep model composed of a CNN for extracting audio features and 3D-CNN for visual features, then used a Deep Belief Network (DBN) for fusing features and SVM for classification. A persistent problem in developing audio-visual emotion recognition systems is that most of them present lower accuracy rates. One of the main issues is due to the procedure in how audio and visual features are extracted and fused. Both unimodal and multimodal emotion recognition are still challenging task, facing an open problem that is the emotion feature extraction from raw data.

In order to address the mentioned problems, deep learning techniques seem to be appropriate due to their powerful feature discriminator and feature learning capacity into many research fields, such as object detection, recognition, and segmentation. Deep learning became very popular in 2006 thanks to an advance in speech recognition [153]. Among them, Convolutional Neural Networks (CNN) are one of the most representative deep learning models. CNN presents some advantages against the traditional methods, such as hierarchical feature representation, weight sharing, sparse interaction, and performed well at feature extraction of data with a specific structure, i.e., image data. CNN also employs raw data as input instead of using hand-crafted features. Therefore, CNN has been extensively applied in a wide range of domains, for instance, emotion recognition, pedestrian detection, image reconstruction, video analysis, image super-resolution, among others.

In this work, we introduce a novel and effective methodology for audio-visual emotion recognition based on a hybrid deep convolutional neural network for audio and visual feature extraction. The approach is comprised of three stages. The first one converts the raw audio signal into an image representation as the input to the audio-network, which is a pre-trained Convolutional Neural Network (CNN). The second phase consists of performing facial detection and cropping the facial images from videos and transformed them into a Census-Transform (CT) representation as the input to the visual-network. This network is a pre-trained CNN, which is fine-tuned to learn facial expression features. The third stage aims to extract the audio and visual features from the hybrid deep convolutional neural network and merged them. Consequently, the feature vector dimensionality is reduced using PCA and LDA. Finally, audio-visual emotion recognition is conducted through KNN, SVM, Logistic Regression (LR) and Gaussian Naïve

Bayes (GNB) for comparing the accuracy rates.

Experiments are conducted on three challenging public data sets. As main contribution, the results achieved with the proposed recognition method are compared to other approaches available in the literature, demonstrating that our solution reached higher recognition accuracies for audio-visual emotion recognition.

The remainder of the paper is organized as follows. Section 4.2 describes the methodology proposed in this work. Section 4.3 presents and discusses our experimental results. Section 4.4 concludes the paper with final remarks and directions for future work.

4.2 Methodology

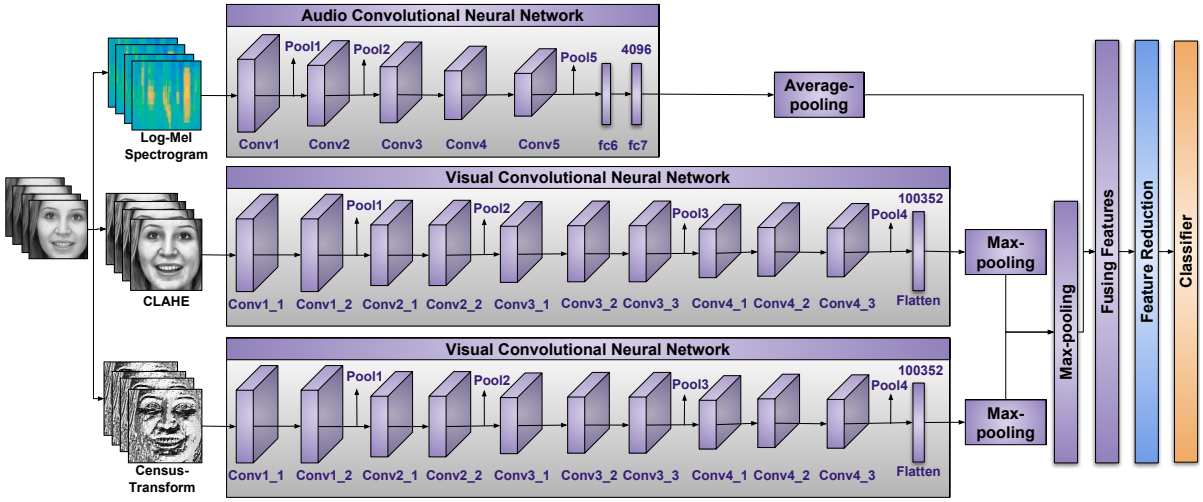


Figure 4.1: Diagram of the proposed methodology for audio-visual emotion recognition.

A diagram of the proposed methodology is illustrated in Figure 4.1. Our hybrid convolutional neural network receives three input streams, i.e., the audio signal representation and two visual representations processed by CNN models. The audio and visual features are extracted from fully-connected layers and then integrated through pooling strategies. The outputs are reduced dimensionally and the classification stage is then performed.

In the following subsections, we explain how to generate the inputs to the audio and visual neural networks, how our hybrid deep model is trained, as well as how the fusion and classification steps work.

4.2.1 Audio Data

Log Mel-spectrogram has been demonstrated to be efficient for discriminating features for emotion recognition in audio clips [89, 149]. It is a representation of how the frequency content of a signal changes along time.

Initially, we compute the log Mel-spectrogram for the entire audio signal, using $M = 64$ Mel-filter banks from 20 to 8000Hz with an *offset* $= 0.01$, defined as

$$MelSpectrogram_{frame_t} \approx \log(melbank(M) * \text{abs}(rfft(frame_t))^2 + offset) \tag{4.1}$$

where *frame*$_t$ denotes the t frame, *melbank* means to convert hertz frequency into Mel-scale applying M Mel-filters banks, *rfft* implies the fast Fourier transform, *abs* represents the complex number amplitude, and *offset* is set to avoid taking the logarithm 0. Moreover, we employ a 0.025s Hamming window size, and a 0.010 s hop size. Then, we divide the log Mel-spectrogram into audio segments of 64×64 size by using a window of 64 frames and a hop of 34 frames. Then, the segment duration time is 655ms (= 10ms × (64-1) + 25ms). After obtaining log Mel-spectrogram segments of 64×64 size, we calculate their first (delta) and second (delta-delta) order differentials. The delta coefficients are computed as

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{4.2}$$

where d_t is a delta coefficient of a frame t calculated through the static coefficients c_{t+n} to c_{t-n}. A common value for N is 2. Therefore, we calculate delta-delta coefficients using the generated delta coefficients. As a result, we generate three-channels for each log Mel-spectrogram segment of 64×64×3, which can be seen as an image representation. Hence, as each formed audio segment will be the input to a pre-trained audio network, we resize each segment into 227×227×3 size by applying inter-cubic interpolation.

4.2.2 Visual Data

For each video frame, we perform automatic facial detection by Dlib library and crop the face region employing a proper bounding rectangle. Then, we apply the Census-Transform (CT) operator over the cropped regions, due to its robustness to monotonic changes in illumination and its holistic representation, suppressing detailed textural information. It has also been shown to perform well as a visual representation for facial expression recognition [24].

CT operator implies the intensity value comparison between a center pixel and its eight-neighborhood. If the intensity value of the center pixel is greater than or equal to one of its neighboring pixels, bit 1 is set in the corresponding position; otherwise, bit 0 is set. Thus, the eight bits generated can be joined together arbitrarily and converted to a base-10 number, which represents the Census-Transform value of the current center pixel. Figure 4.2 illustrates an example of CT encoding for one pixel of the input image. Therefore, CT is applied to all cropped region pixels, obtaining a grayscale image, as shown in Figure 4.3(c).

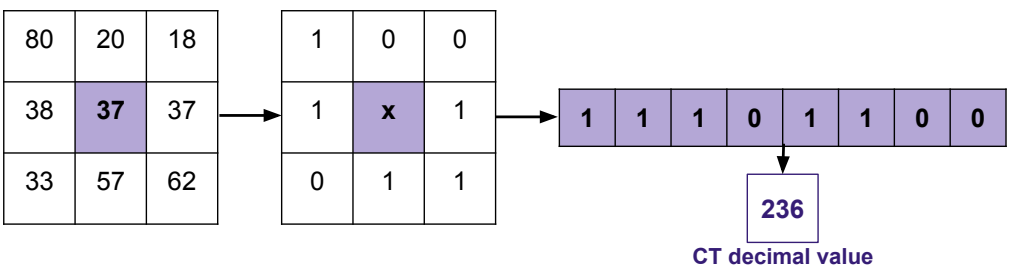


Figure 4.2: 3×3 Census Transform for a single output pixel.

(a) (b) (c)

Figure 4.3: (a) Cropped image extracted from BAUM-1s database; (b) CLAHE processed image from (a); (c) census-transformed image from (a).

We also apply the contrast-limited adaptive histogram equalization (CLAHE) [155] to the original cropped regions to enhance the image contrast. Consequently, we obtain two image representations for each cropped facial region. Therefore, we convert each generated image into color images and scale them into $224 \times 224 \times 3$ pixels using an inter-area interpolation for shrinking the image and an inter-cubic interpolation for enlarging it. Figure 4.3 shows the obtained images after applying CLAHE and CT operators.

4.2.3 Audio Network

We employ a 2D convolutional neural network (CNN) model for extracting audio features. We selected the AlexNet [76] architecture with pre-trained weights on ImageNet data set. This network is composed of five convolutional layers (*Conv1-Conv2-Conv3-Conv4-Conv5*), three max-pooling layers (*pool1-pool2-pool5*) and three fully connected layers (*fc6-fc7-fc8*). Then, we normalize the training set through min-max-normalization and scale the testing and validation set using the training set parameters. We extract 4096-D audio features from the *fc7* layer.

4.2.4 Visual Network

For facial expression feature extraction, we use a 2D convolutional neural network based on the VGG-Face network [107]. The VGG-Face network architecture consists of thirteen convolutional layers (*Conv1_1-Conv1_2-Conv2_1-Conv2_2-Conv3_1-Conv3_2-Conv3_3-Conv4_1-Conv4_2-Conv4_3-Conv5_1-Conv5_2*), five max-pooling layers (*pool1-pool2-pool3-pool4-pool5*) and three fully connected layers (*fc6-fc7-fc8*). This network was pre-trained with about 2.6 million face images from 2622 subjects. Our visual network comprises the first fourteen VGG-Face layers, involving ten convolutional layers (from *Conv1_1* to *Conv4_3*) and four max-pooling layers (from *pool1* to *pool4*), one flatten layer, and one fully connected layer with six neurons since there are six emotions to be classified.

Our visual network receives as input facial images based on Census Transform and CLAHE. Then, for training our network, we perform data augmentation over the training set images, which includes rotating the images by a random angle in a range of -15° to 15°, and shifting them vertically and horizontally by no more than 20% of image size. Then, we employ a 128 batch size and a stochastic gradient descent (SGD) with a learning rate of 0.001, a momentum of 0.9, and decay of 0.00005. We set two stages for fine-tuning. First, we fine-tune the training set of each data set by running weight updates for the connected layers during five epochs. In the

second one, we update the upper layers from *Conv4_1*-layer for twenty epochs. The training objective function is categorical cross-entropy loss function, expressed as

$$\textit{CategoricalLoss}(o) = -\sum_{i=1}^{K} y_{o,c} \log(p_{o,c}) \tag{4.3}$$

where K denotes the class number, y is a binary indicator (0 or 1) that equals 1 if the class label c is the correct class for observation o, and p indicates the probability that observation o is predicted as class c. After the fine-tuning network, we extract 100352-D visual features from the *flatten* layer.

4.2.5 Audio Segment-based Fusion

Since each audio-visual video sample contains a different number of segments, we perform pooling over all the segments and frames. In the case of audio segments, we apply average-pooling, obtaining an audio feature vector of 4096-D length.

4.2.6 Video Frame-based Fusion

For the video frames, we have two visual segment sequences. One of them is composed of CLAHE processed images and the other composed of CT images. Then, for each segment sequence, we compute max-pooling, generating a visual feature vector of 100352 dimensions. Thus, we apply max-pooling on both generated feature vectors, obtaining a final vector of 100352 dimensions.

4.2.7 Fusion and Classification

We concatenated both resulting audio and visual vectors, forming a global feature vector of 104448 dimensions. Therefore, we employ feature dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Thus, we apply PCA over the resulting audio-visual feature vector. Consequently, we apply LDA over the feature space generated by PCA.

This procedure has been proved to enhance accuracy notably in several object recognition problems [24, 29]. Finally, we perform audio-visual classification through K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB) classifiers for comparing the accuracy rates.

4.3 Experimental Results

The proposed methodology has been tested on three audio-visual emotion data sets: the Ryerson Multimedia Lab (RML) [132] data set, the eNTERFACE05 [95] data set and the spontaneous BAUM-1s data set [146].

The RML data set is comprised of 720 audio-visual videos from eight subjects, which were asked to speak six different languages: English, Mandarin, Italian, Urdu, Punjabi, and Persian.

This data set contains six emotions: anger, disgust, fear, happiness, sadness, and surprise. The samples are recorded by an audio sample rate of 22050 Hz with 16-bit resolution and mono channel. Each video frame has a size of 720×480×3 [132]. Figure 4.4(a) presents some samples of the cropped facial images on the RML data set.

The eNTERFACE05 data set contains 1290 audio-visual video samples, labeled as one of six emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise, performed by 44 models from 14 different counties. Among the subjects, 81% is men, while the remaining is women. 31% wear glasses, and 17%, beard. Thus, each subject is asked to listen to a short story in order to stimulate a particular emotion. Each sample is captured at a sampling rate of 48 KHz using a single channel 16-bit digitization. The video frame size is 720×576×3 [95]. Figure 4.4(b) shows some cropped facial image samples on the eNTERFACE05 data set.

The BAUM-1s data set is a spontaneous audio-visual video collection of 31 Turkish subjects, being 17 female, performing twelve target emotions and mental states: happiness, sadness, fear, anger, disgust, surprise, boredom, contempt, unsure, thinking, concentrating, and bothered. This data set is composed of 1222 videos, recorded in a mono channel. For recording spontaneous audio-visual emotions, emotion elicitation using films is applied. The size of the video frames is 720×576×3. The BAUM-1s data set also provides English subtitles per video sample [146]. In this work, we focus on recognizing the six emotions as the other data sets. Thus, we filter the videos containing those emotions, obtaining a total of 521 videos. Figure 4.4(c) gives some samples of the cropped facial regions on the BAUM-1s data set.

Figure 4.4: Cropped facial expression regions from (a) RML, (b) eNTERFACE05, and (c) BAUM-1s data sets. From left to right, the emotions cataloged are anger, disgust, fear, happiness, sadness, and surprise.

We implement our deep models using Keras platform. For training our deep neural networks, we employ an NVIDIA TITAN V GPU with 12GB memory. For performing emotion recognition, we use the scikit-learn library. Moreover, we follow the Leave-One-Subject-Out (LOSO) protocol for validating our proposed methodology over RML and eNTERFACE05 data sets, while the Leave-One-Speakers-Group-Out (LOSGO) strategy with five speaker groups is used for BAUM-1s data set.

We conducted unimodality experiments using dimensionality reduction techniques (PCA and PCA+LDA) and different classifier algorithms (KNN, SVM, LR, and GNB) for comparing recognition rates. The results are shown in Tables 4.1 and 4.2, whose values represent the average recognition accuracy rates of unimodality features from the performed experiments. We can see that PCA+LDA is significantly superior to just using PCA.

Table 4.3 presents the best results on each data set. These results are subject-independent and obtained by means of the audio and visual learned features with the pre-trained AlexNet and the fine-tuned CT based VGG. In order to verify the effectiveness of the learned features, we compare them with the results of other approaches available in the literature that use the same experimental settings, such as subject-independent validation protocols. Table 4.4 and 4.5 summarize the audio and visual recognition rates reached by our approach, including the other methods available in the literature, across the RML, eNTERFACE05 and BAUM-1s data sets.

From Table 4.4, we can observe that our audio features with AlexNet are slightly superior to the other approaches on RML and BAUM-1s data sets, except on the eNTERFACE05 data set. On the spontaneous BAUM-1s data set, our proposed method provides an improvement of about 5% compared to the deep features obtained by Zhang et al.'s approach [149]. On the acted eNTERFACE05 data set, despite that our method is not superior than some other approaches, it is still better and more consistent than other hand-crafted techniques, such as prosody features, Power Normalized Cepstral Coefficients (PNCC), Low-level Descriptors (LLD), Mel-Frequency Cepstral Coefficients (MFCC), Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) [7, 46, 47, 59, 93, 121, 146, 148].

From Table 4.5, we can notice that the recognition rate achieved by our visual learned features outperforms significantly the other literature methods, including the ones based on deep neural networks (VNet [149]) and hand-crafted features (LBP [148], LPQ [146], Gabor wavelet [46], facial points [93] and Quantized Image Matrix (QIM) [7]). This shows that our visual features learned by the fine-tuned CT based on VGG model have more discriminative power, clearly showing that deep features are more powerful than the hand-crafted features. As mentioned in Section 4.2, we employ a pre-trained deep model on a large face data set for face recognition, whereas the VNet introduced by Zhang et al. [149] uses a pre-trained model on the Imagenet data set, which is developed for object recognition. Our starting deep model decision seems to be more suitable for solving facial expression recognition task.

Table 4.1: Subject-independent unimodality average accuracy rates (%) using audio features for each data set.

Data Sets	Method	KNN	SVM	LR	GNB
RML	PCA	37.50	53.75	60.00	43.75
	PCA + LDA	66.25	65.00	64.38	**68.75**
eNTERFACE05	PCA	30.67	45.33	57.33	36.67
	PCA + LDA	54.67	54.67	55.33	**62.00**
BAUM-1s	PCA	31.56	37.75	40.26	38.64
	PCA + LDA	33.94	45.73	40.80	**40.76**

Table 4.2: Subject-independent unimodality average accuracy rates (%) using visual features for each data set.

Data Sets	Method	KNN	SVM	LR	GNB
RML	PCA	35.00	16.25	**75.00**	41.25
	PCA + LDA	63.75	47.50	73.75	65.00
eNTERFACE05	PCA	35.00	16.67	63.33	27.50
	PCA + LDA	65.00	63.33	64.17	**68.33**
BAUM-1s	PCA	34.37	35.36	54.50	25.25
	PCA + LDA	54.17	55.40	55.27	**59.52**

Table 4.3: Subject-independent unimodality average accuracy rates (%) using audio features extracted by the original pre-trained AlexNet and visual features obtained through the CT based VGG for each data set.

Unimodality	RML	eNTERFACE05	BAUM-1s
Audio	68.75	62.00	46.76
Visual	75.00	68.33	59.52

Table 4.4: Comparison of subject-independent audio recognition average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets.

	Approach	Audio features	Accuracy
RML	Gao et al. [59]	Prosody	51.04
	Elmadany et al. [47]	Prosody	56.25
	Elmadany et al. [46]	PNCC	58.33
	Zhang et al. [148]	LLD	61.86
	Zhang et al. [149]	ANet	66.17
	Ours	**AlexNet**	**68.75**
eNTERFACE05	Mansoorizadeh et al. [93]	Prosody	43.00
	Bejani et al. [7]	Prosody, MFCC	54.99
	Zhalehpour et al. [146]	MFCC, RASTA-PLP	72.95
	Schuller et al. [121]	Prosody, MFCC	72.40
	Zhang et al. [149]	ANet	78.08
	Ours	**AlexNet**	**62.00**
BAUM-1s	Zhalehpour et al. [146]	MFCC, RASTA-PLP	29.41
	Zhang et al. [149]	ANet	42.26
	Ours	**AlexNet**	**46.76**

Table 4.5: Comparison of subject-independent visual recognition average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets.

	Approach	Visual features	Accuracy
RML	Zhang et al. [148]	LBP	56.90
	Elmadany et al. [46]	Gabor wavelet	64.58
	Zhang et al. [149]	VNet	68.09
	Ours	**CT based VGG**	**75.00**
eNTERFACE05	Mansoorizadeh et al. [93]	Facial points	37.00
	Bejani et al. [7]	QIM	39.27
	Zhalehpour et al. [146]	LPQ	42.16
	Zhang et al. [149]	VNet	54.35
	Ours	**CT based VGG**	**68.33**
BAUM-1s	Zhalehpour et al. [146]	LPQ	45.04
	Zhang et al. [149]	VNet	50.11
	Ours	**CT based VGG**	**59.52**

Table 4.6 shows the average recognition rates of the fusion of audio and visual features on each data set. From these experiments, we can observe that the feature fusion significantly increased unimodal recognition accuracy rates. We also compared our approach against other methods available in the literature.

Table 4.7 exhibits the best results achieved with our method and the ones obtained with other methods available in the literature for the RML, eNTERFACE05 and BAUM-1s data sets. On the acted RML and eNTERFACE05 data sets, it is observed that our method reached competitive recognition rates to other state-of-the-art results. On the spontaneous BAUM-1s data set, the accuracy rate is much higher by approximately 5% to the recognition rate achieved by Zhang et al. [149], whose method also employs a deep learning approach. The BAUM-1s data set is more challenging and realistic since the subjects express spontaneous emotions.

Table 4.6: Subject-independent multimodality average accuracy rates (%) using the fusion of audio and visual features for each data set.

Data Sets	Method	KNN	SVM	LR	GNB
RML	PCA	42.50	16.25	77.50	43.75
	PCA + LDA	63.75	52.50	**82.50**	76.25
eNTERFACE05	PCA	41.67	16.67	69.17	37.50
	PCA + LDA	80.83	71.67	72.50	**85.00**
BAUM-1s	PCA	35.37	35.36	55.39	25.18
	PCA + LDA	54.82	56.30	57.10	**59.70**

Table 4.7: Comparison of subject-independent multimodality average accuracy rates (%) on RML, eNTERFACE05 and BAUM-1s data sets.

Data Sets	Approach	Accuracy
RML	Sarvestani et al. [120]	72.03
	Zhang et al. [148]	74.32
	Elmadany et al. [46]	75.00
	Zhang et al. [149]	80.36
	Ours	**82.50**
eNTERFACE05	Sarvestani et al. [120]	70.11
	Mansoorizadeh et al. [93]	71.00
	Zhalehpour et al. [146]	77.02
	Bejani et al. [7]	77.78
	Zhang et al. [149]	85.97
	Ours	**85.00**
BAUM-1s	Zhalehpour et al. [146]	51.29
	Zhang et al. [149]	54.57
	Ours	**59.70**

4.4 Final Considerations

In this work, we proposed a hybrid deep convolutional neural network for audio-visual emotion recognition. We introduced a visual network based on Census-Transform for extracting facial expression features from videos, demonstrating its discriminative power. Experimental results showed that following a PCA+LDA approach can improve accuracy rates significantly. Furthermore, our proposed method outperformed robust and effective methods available in the literature, especially in the spontaneous BAUM-1s data set.

In order to improve the recognition accuracy rates, we intend to explore the pre-trained AlexNet model, i.e., perform fine-tuning, investigate other deep models and fusion strategies. We also plan to perform experiments on other data sets to validate the consistency of our method.

Chapter 5

Down Syndrome Detection based on Facial Features using a Geometric Descriptor

Abstract

Down syndrome is one of the most common genetic disorders caused by chromosome abnormalities in humans. Among other physical characteristics, certain facial features are typically associated in people with Down syndrome. In this work, we investigate the problem of Down syndrome detection from a collection of face images. As main contribution, a compact geometric descriptor is used to extract facial features from the images. Experiments are conducted on an available data set to demonstrate the performance of the proposed methodology.

5.1 Introduction

Dysmorphology is the branch of medicine (more specifically, clinical genetics) in which physicians are involved in the identification of congenital anomalies caused by hundreds of different syndromes [113]. Early and accurate diagnosis of a syndrome is important to the patients and their families, resulting in a more effective care. However, it is worth emphasizing that a precise recognition of dysmorphic features depends upon experienced clinicians and health staff, as well as complex laboratory procedures.

The use of computer-aided tools [42, 79] based on image processing and machine learning techniques can facilitate the recognition of facial dysmorphic features associated with genetic syndromes. By extracting facial points and computing measurements from images, such structural malformations may be automatically identified.

Several genetic disorders present dysmorphic features, for instance, Williams, Fragile X, Cornelia de Lange, Down, Dubowitz, Smith-Magenis, Edwards, Russell-Silver, and Prader-Willi syndromes.

Down syndrome is a genetic abnormality that occurs in approximately one per 1,000 babies born each year [91]. The syndrome was described in details by John Langdon Down in 1866 [66]. Identified by French researchers in 1959, the presence of an extra chromosome 21 causes the genetic disorder [108].

Several physical and intellectual disabilities are typically associated with Down syndrome [53], however, their degree varies from person to person. While some people with

Down syndrome may require much support and care, others can live an ordinary life.

Some common characteristics present in people with the disorder include distinctive facial features such as slanting eyes, small chin, round face, flat nasal bridge, Brushfield spots in the iris, abnormal outer ears, and flattened nose.

There are still few studies available in the literature that investigate the identification of Down syndrome from images. In this work, we propose and analyze a method for Down syndrome detection based on facial features using a geometric descriptor [29]. This compact descriptor is composed of 14 distances calculated between facial fiducial points.

Experimental results demonstrate that the proposed method is effective to discriminate facial differences from individuals affected by Down syndrome and healthy individuals over a collection of images [54]. The data set contains photographs of patients from different age, gender and ethnicity, acquired from different angles.

The remainder of the paper is organized as follows. Section 5.2 briefly describes related work available in the literature. Section 5.3 presents the methodology proposed in this work. Section 5.4 describes and analyzes the experimental results. Section 5.5 concludes the paper with final remarks and directions for future work.

5.2 Background

Landry et al. [78] used mathematical techniques for classifying congenital abnormalities from hand radiographs. Discriminant analysis was used to identify different syndromes from normal condition.

A computer-based recognition of dysmorphic faces based on a graph-matching algorithm for describing facial patterns was described by Loos et al. [85]. Gabor wavelets were used to preprocess the photographs and obtain a description of image regions.

A method for identifying among ten syndromes is presented by Boehringer et al. [10]. Gabor wavelet filter was applied to the landmark coordinates extracted from the faces.

Erogul et al. [49] described a method for detecting Down syndrome in photographs of children. A graph method is constructed from facial points and a neural network is used in the classification process.

Zhao et al. [151] proposed a hierarchical constrained local model based on independent component analysis (ICA), which was applied to detect Down syndrome from facial photographs. Local binary patterns (LBP) were extracted and selected from located landmarks. Different classifiers were used to compare the results of the Down syndrome identification. The LBP descriptor was employed by Burçin and Vasif [11] to identify Down syndrome in face images.

Saraydemir et al. [119] used Gabor wavelet transform as feature extractor, as well as K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) as classifiers to analyze the effect of training set dimension on the recognition of dysmorphic faces.

David and Lerner [37] described a method for genetic syndrome classification based on SVM classifier using fluorescence in situ hybridization (FISH) signals. Different SVM kernels were analyzed, where Gaussian kernel provided more accurate results.

Zhao et al. [152] presented a method for detecting Down syndrome in images through the combination of texture and geometric information. An SVM classifier was employed to

distinguish between normal and abnormal cases.

Table 5.1 presents the results for some Down syndrome recognition methods available in the literature. It is worth mentioning that most of the methods report accuracy rates for private data sets.

Table 5.1: Comparison of some methods for Down syndrome recognition according to accuracy rates (in percentage).

Approach	Strategy	Number of Images	Accuracy (%)
Zhao et al. [152]	Geometric and Texture Features + SVM	48	97.90
Saraydemir et al. [119]	Gabor Wavelet Transform + PCA + LDA + SVM	30	97.34
Kruszka et al. [77]	Geometric and Texture Features + SVM	65	94.30
Burçin and Vasif [11]	Local Binary Patterns (LBP)	107	95.30
Erogul et al. [49]	Elastic Bunch Graph Matching (EBGM)	86	68.70

5.3 Proposed Methodology

The proposed methodology for recognition of Down syndrome in face images is composed of four main stages: (i) facial detection, (ii) feature extraction, (iii) feature reduction, and (iv) classification. Figure 5.1 illustrates these steps, which are described in more details in the following subsections.

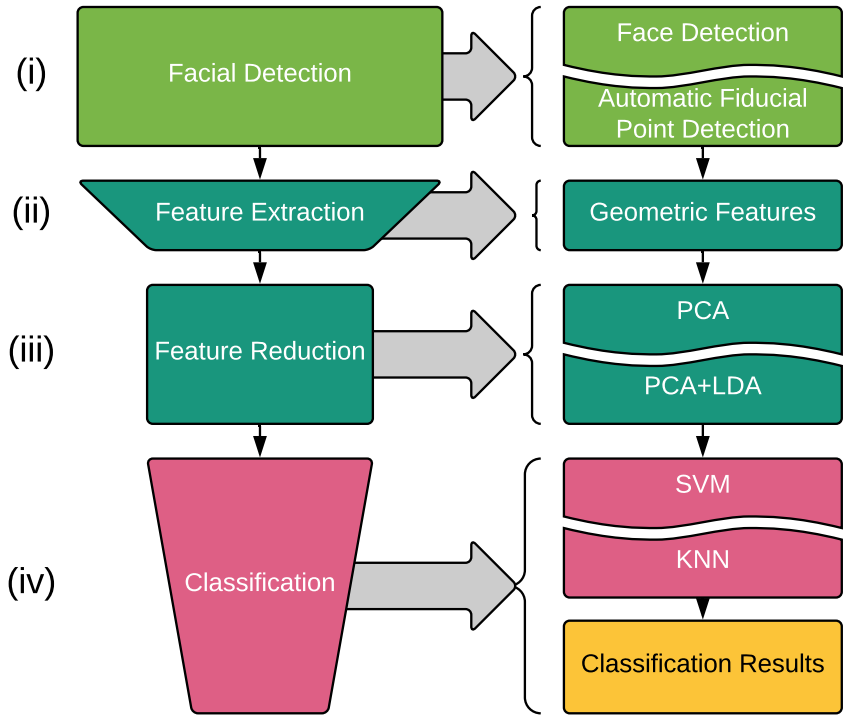
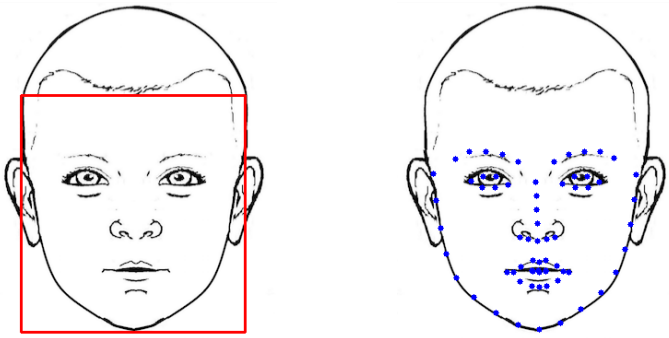


Figure 5.1: Diagram with the main steps of the proposed Down syndrome detection methodology.

5.3.1 Facial Detection

Initially, the Viola-Jones method [129] is applied to detect faces in the image collection. The algorithm is based on a machine-learning approach that uses a cascade function trained with positive and negative images. Haar-like features are extracted from windows at multiple scales over an integral image. Relevant features are selected through a variant of Adaboost learning algorithm, where a cascade classifier is composed of several stages, each one containing a classifier that determines if a window contains a face or not. This cascade structure improves the chance of a face being detected. Figure 5.2(a) illustrates a detected face.

After a face is detected, Dlib library [75] is used to extract facial landmarks. The algorithm employs a ensemble of regression trees trained to estimate the landmark positions. A total of 68 points, corresponding to (x, y) coordinates of regions surrounding each facial structure (eyes, eyebrows, mouth, nose, jaw), are located in the images. Figure 5.2(b) illustrates this stage.



(a) (b)

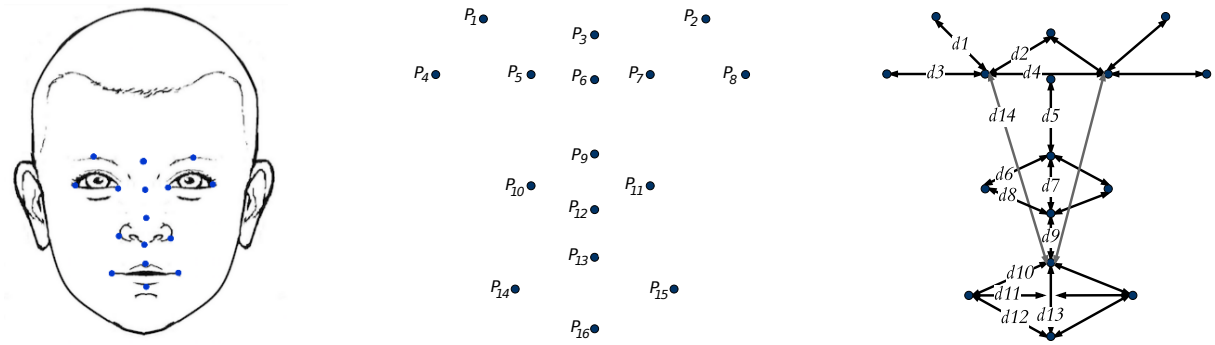
Figure 5.2: Facial detection step. (a) detected face; (b) marked fiducial facial points.

5.3.2 Feature Extraction

In this step, a geometric descriptor is built from the detected facial landmarks. The geometric representation uses sixteen 2D facial fiducial points: two points for the middle of the eyebrows (P_1, P_2), one point for the glabella (P_3), four points for the inner and outer corner of the eyes (P_4, P_5, P_7, P_8), one point for the root of nose (P_6), two points for the alars sidewalls (P_{10}, P_{11}), one point for the supratip (P_9), one point for the columella (P_{12}), two points for the mouth corners (P_{14}, P_{15}), and two points for the top of the upper lip and bottom of the lower lip (P_{13}, P_{16}). Figures 5.3 (a)-(b) show the localization of the sixteen facial points.

Fourteen distances are extracted from the mentioned sixteen points, as shown in Figure 5.3 (c). These distances are normalized to the face width to ensure the features are scale invariant. The nine distances d_1, d_2, d_3, d_6, d_8, d_{10}, d_{11}, d_{12} and d_{14} represent the average values of the two mirrored distances on the left and right sides of the face. The latter distance d_{11} is calculated using the intersection point of the line between the points on the top of the upper lip and bottom of the lower lip, and the line between the left and right corners of the mouth. Hence, the resulting feature vector has 14 dimensions.

For Down syndrome recognition, it is important to consider the intercanthal distance d_4, that is, the distance between the inner corners of the eyes. People with Down syndrome often present telecanthus [33,48], meaning an increasing of the intercanthal distance. Another common



(a) (b) (c)

Figure 5.3: Feature extraction. (a)-(b) sixteen facial landmark detected from the face; (c) fourteen distances extracted from the sixteen fiducial points.

characteristic is the occurrence of small palpebral fissures [33], which is the space between the lateral and medial canthus of the eyes. This feature is captured by distance d_3. Individuals with Down syndrome frequently present symptom of flattened nose [48]. Distances d_5 to d_8 represent this characteristic. Finally, another symptom is the presence of small mouth [48], whose characteristic is captured by distances d_{10} to d_{13}.

As opposed to the extraction of geometric features, an image preprocessing stage is required for the texture feature extraction in order to obtain aligned faces with uniform size and shape. This preprocessing task consists of the following 6 steps: (i) automatic facial point detection, (ii) coordinate extraction of the inner corner of the eyes, (iii) image rotation to align the eye's inner corner coordinates, (iv) image scaling proportionally to the mean distance between the eyes, (v) facial cropping using a proper bounding rectangle, (vi) conversion of the color images to grayscale.

Census Transform (CT) [144] is a non-parametric local transform originally developed for determining correspondences between local patches. This transform compares the intensity value of a pixel with its eight neighboring pixels, such that if its value is higher than or equal to one of its neighbors, a bit 0 is set in the corresponding location or, otherwise, a bit 1 is set. The process of CT encoding for one pixel can be illustrated in the following example:

$$\begin{array}{|c|c|c|} \hline 95 & 27 & 98 \\ \hline 40 & 69 & 82 \\ \hline 30 & 55 & 79 \\ \hline \end{array} \quad \Rightarrow \quad \begin{array}{|c|c|c|} \hline 1 & 0 & 1 \\ \hline 0 & \times & 1 \\ \hline 0 & 0 & 1 \\ \hline \end{array} \quad \Rightarrow \quad 10111000 = (184)_{10} \tag{5.1}$$

The concatenation of the eight bits produced after the intensity comparison is then converted to a decimal number in the $[0, 255]$ interval. It has been demonstrated that CT is robust to illumination changes [111].

The Census Transform Histogram (CENTRIST) [136] is a histogram constructed from the CT values from an image, which has been successfully applied as a visual descriptor to several image classification problems [23, 24]. CENTRIST is able to describe local structures of the image, providing a high generalization for categorization and suppressing detailed textural information.

In our methodology, CENTRIST is applied to each entire image, generating a feature vector of length 256. This vector is normalized by dividing each intensity level r by the total pixels n of the image, that is

$$p(r) = \frac{h(r)}{n} \tag{5.2}$$

where $h(r)$ denotes the occurrence frequency of each intensity level in the image, whereas $p(r)$ represents the probability of occurrence intensity.

5.3.3 Feature Reduction

Two techniques for feature reduction [14], Principal Component Analysis (PCA) [73], and Linear Discriminant Analysis (LDA) [55], are employed to maintain the most representative features, avoiding redundant or noisy information [96]. These approaches are used sequentially and applied individually for each feature set.

PCA is a known technique [73] used to transform the original data into a lower dimensional feature descriptor. New orthogonal bases are obtained by solving the eigenvalue problem through training samples. These orthogonal bases, called eigenfaces, are suitable for image reconstruction [106].

LDA is used to search for a combination of features that discriminate class samples [55]. The method projects the images onto a subspace that minimizes the within-class scatter and maximizes the between-class scatter of the projected data.

The combination of PCA and LDA takes advantages of both techniques, that is, data dimensionality reduction by projecting the data onto the eigenface space followed by LDA to perform class separability.

5.3.4 Classification

Two different classifiers [56] are evaluated in our experiments, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). The classifiers are applied to the feature vector, after its dimensionality reduction, in order to provide a comparison between the achieved recognition rates.

Our methodology is validated through a k-fold cross-validation protocol, that is, the original set of samples is randomly partitioned into k equal sized subsets. A single subset of samples is then kept as the validation data for testing the model, whereas the remaining $k-1$ subsets are used as training data. The process of cross-validation is repeated k times, such that each of the k subsets of samples is used exactly once as the validation data. Finally, the k results from the folds are averaged to generate a single estimation.

A grid search strategy is applied to look for combinations of parameter values and then select the parameters that provide the most accurate classification model.

5.4 Experimental Results

Our method used the data set collected by Ferry et al. [54], which is composed of 199 images from publicly available or scientifically published photographs of patients diagnosed with Down

Table 5.2: Average accuracy, in percentage, using our geometric representation and the CENTRIST descriptor on the evaluated data set.

Recognition Method	Geometric (%)	CENTRIST (%)	Geometric + CENTRIST (%)
PCA + KNN	91.13	90.33	96.78
PCA + LDA + KNN	95.17	94.35	97.59
PCA + SVM	95.17	93.55	97.59
PCA + LDA + SVM	**98.39**	95.17	**98.39**

syndrome. This data set contains a collection of images of patients from different ethnicity, gender and age. The images were captured spontaneously from different backgrounds, angles and distances, showing varying head poses, facial expressions and occlusions.

We selected only images of children, discarding images of adults and at very poor resolution. The latter decision is because it precludes accurate facial fiducial point detection. The resulting image set consisted of 153 images.

We used the Dartmouth Database of Children's Faces [35] for obtaining images of healthy children. This data set contains 40 male and 40 female children between 6 and 16 years of age, who posed eight facial expressions (neutral, contempt, happiness, surprise, anger, sadness, disgust, fear). Models were photographed from various angles and under different lighting conditions. We selected a set of 153 images from the Dartmouth Database of Children's Faces [35], considering males and females showing very distinct facial expressions.

Therefore, our data set is composed of 153 images of each of the above mentioned data sets, that is, 306 images in total to test and verify our proposed method. Then, we conducted experiments using our geometric representation, CENTRIST descriptor and the fusion of both features through four approaches: PCA + KNN, PCA + LDA + KNN, PCA + SVM and PCA + LDA + SVM.

Our methodology was validated through a 10-fold cross-validation protocol. The results are reported in Table 5.2, whose values represent the average detection accuracy rate obtained in the performed experiments.

From our experiments, we can observe that the geometric features are capable of reaching high accuracy rates for Down syndrome detection. We can also notice that the PCA + LDA approach achieves higher detection rates. Furthermore, the CENTRIST descriptor provides a competitive performance with more than 90% accuracy.

Although the fusion of the geometric and texture features provides high accuracy, its result is not superior to the accuracy rate obtained only with geometric features. Geometric features seem to have predominance over textural features.

Table 5.3 shows the best detection accuracy rates achieved with the geometric features proposed in our approach and in the method developed by Zhao et al. [152]. For our method, the best parameters $\gamma = 1.0$ and $C = 1$ of the Radial Basis Function (RBF) kernel SVM were obtained through a grid search in order to improve the classification accuracy rates. The area under the receiver operating characteristic curve (AUC) was 0.9839. For the method by Zhao et al. [152], the best parameters obtained with our implementation were achieved with $\gamma = 0.05555556$ and $C = 1$, generating an AUC = 0.9273. It is worth mentioning that the

Table 5.3: Comparison of average accuracy rates (%) for Down syndrome detection over our formed data set.

Recognition Method	Strategy	Accuracy (%)
Our method	14 geometric distances + PCA + LDA + SVM	**98.39**
Zhao et al. [152]	19 geometric distances + SVM	92.73

original work developed by Zhao et al. [152] was evaluated in a private data set. They achieved an average accuracy rate of 95.80% with the geometric features, 77.10% with textural features and 97.90% with combined features.

Table 5.4 shows the confusion matrix obtained with the proposed method, that is, the geometric descriptor composed of 14 geometric distances combined with the PCA+LDA+SVM strategy. Table 5.5 shows the confusion matrix obtained with the method developed by Zhao et al. [152], that is, the geometric descriptor composed of 19 geometric distances combined with the SVM classifier.

Table 5.4: Confusion matrix for the proposed method.

		Actual Classes	
		Down	Healthy
Predicted Classes	Down	60	2
	Healthy	0	62

Table 5.5: Confusion matrix for the method developed by Zhao et al. [152].

		Actual Classes	
		Down	Healthy
Predicted Classes	Down	58	4
	Healthy	5	57

The Wilcoxon rank-sum test [92] was performed to compare the results obtained between the proposed method and the approach developed by Zhao et al. [152]. This nonparametric statistical test [21], which does not require the assumption of normal distributions, determined that there were significant differences between the methods (p-value = 0.000157, that is, less than 0.05).

There are still few works related to Down syndrome detection available in the literature. Most of these approaches are evaluated on unavailable data sets and consider less than 50 images for each class. These data sets use frontal faces well positioned images without the presence of any occlusion, that is, captured under constrained conditions.

It is worth mentioning that, due to the impossibility of having access to the data sets used in other methods and in order to perform a fair comparison, we implemented the geometric-based approach that presents the highest accuracy in the literature [152].

Our method used only 16 facial fiducial points to calculate 14 distances, whereas the other approach employed 19 fiducial points, resulting in 19 geometric distances. Additionally, our

method was able to perform Down syndrome detection using images captured under uncontrolled conditions, while the other approach worked with frontal images.

From the experimental results, it is possible to state that our geometric descriptor demonstrated to be very discriminative for computing a set of measurements over the facial points and capturing dysmorphic facial features present in individuals affected by Down syndrome.

5.5 Final Considerations

A method for recognizing Down syndrome in face images was proposed and analyzed in this work, which consists facial fiducial point detection, feature extraction, feature reduction, and classification.

A geometric descriptor was used to extract and represent a set of facial features. Different classifiers were employed in the classification process to generate the final results.

Experimental results have shown that a geometric representation can achieve high accuracy rates for Down syndrome recognition in the wild. Thus, geometric features have proven to be robust and independent on gender, race and age.

As directions for future work, we intend to analyze other set of visual features and apply the developed system to different genetic disorders.

Chapter 6

Recognition of Genetic Disorders Based on Deep Features and Geometric Representation

Abstract

In this work, we analyze facial abnormalities in people diagnosed with different genetic disorders through deep features and anthropometric measurements. Based on the assumption that patients with distinct genetic conditions present significant differences in facial morphology, we conjecture that such facial patterns and geometric distances could help in the detection of certain syndromes. Experiments conducted on an available data set demonstrate the effectiveness of the proposed recognition methodology.

6.1 Introduction

There is currently a notable population of near 8% of people with genetic disorders due to mutations in genes, which can affect any part of the body and its functionality. Approximately a third of people with genetic disorders present more serious symptoms that compromise their physical and mental well-being. About 3 to 6% of babies will be born with a genetic disease or a birth disability. In addition, 1 to 3% of people worldwide have an intellectual disability. More than 20% of infant deaths are caused by genetic conditions or congenital defects. Therefore, genetic disorders can be lethal or require major medical care. Genetic conditions affect people of all ages, genders and ethnic groups [31].

Furthermore, 30 to 40% of genetic syndromes present facial and cranial abnormalities, which help physicians diagnose certain disorders, such as Angelman syndrome, Down syndrome, Williams syndrome, among others. Although there are more than 6,000 known genetic disorders, only a few people with a suspected syndrome receive a clinical diagnosis [31]. In this work, we propose and evaluate a strategy for recognizing patterns of facial abnormalities associated with different genetic disorders.

Several studies have been conducted in the literature to investigate facial abnormalities in images of patients with genetic disorders. Loos et al. [85] presented a computer-based recognition of dysmorphic faces to describe facial patterns among five types of syndromes. Their method

extracted a set of features through Gabor wavelet transformations. Similarly, Boehringer et al. [10] applied a set of Gabor wavelet filters at facial landmarks to identify facial abnormalities of ten types of syndromes. Vollmar et al. [130] presented an analysis to determine the impact on recognition accuracy when increasing the number of syndromes. They also described the improvements in the use of geometric features and their combination with texture features in accuracy rates. Ferry et al. [54] proposed an approach to extracting phenotype information using a combination of shape and texture features to recognize eight syndromes. They performed syndrome recognition through supervised and unsupervised learning methods.

The remainder of the paper is organized as follows. Section 6.2 describes the methodology proposed in this work, composed of preprocessing, feature extraction, feature reduction and classification stages. Section 6.3 presents and evaluates the experimental results. Section 6.4 concludes the paper with final remarks.

6.2 Methodology

The proposed methodology for genetic syndrome recognition is composed of four main stages: preprocessing, feature extraction, feature reduction, and classification. These steps are illustrated in Figure 6.1 and explained as follows.

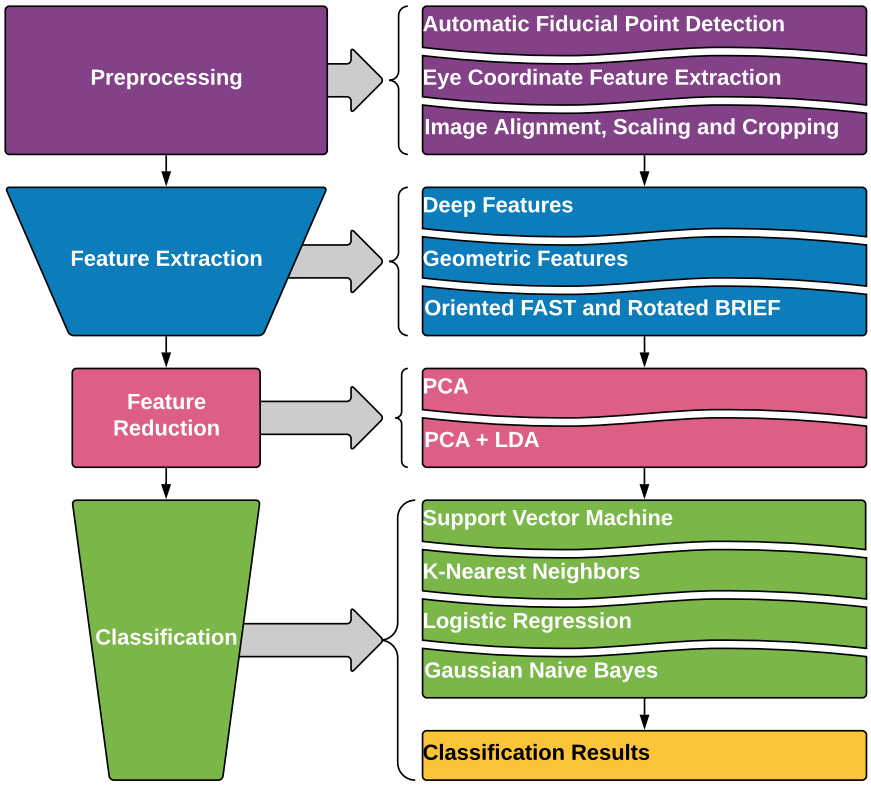


Figure 6.1: Main steps of the genetic syndrome recognition methodology.

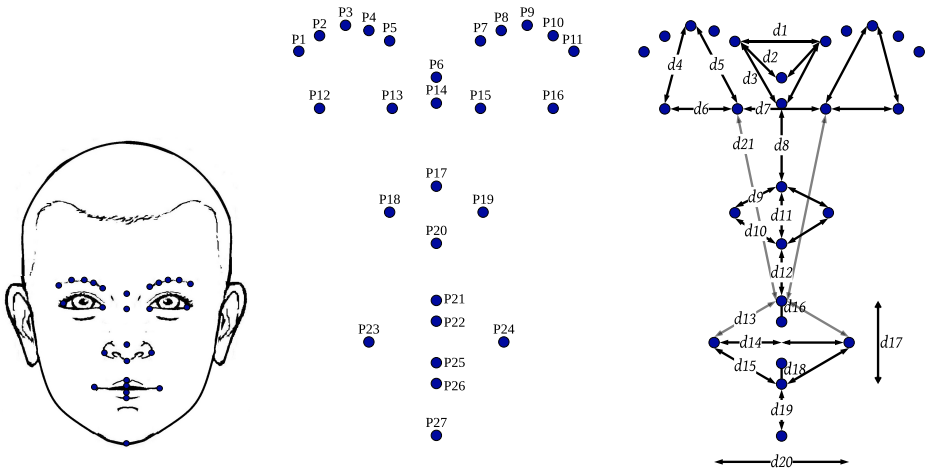
6.2.1 Preprocessing

The image preprocessing procedure is crucial for the genetic syndrome recognition task, whose primary purpose is to provide aligned and cropped faces. This preprocessing stage consists of the following five steps: (i) automatic landmark detection by Dlib library [75]; (ii) eye coordinate feature extraction; (iii) face alignment due to eye coordinates; (iv) cropping the face region applying a fitting bounding rectangle; and (v) and scaling the images to 224×224 pixels.

6.2.2 Feature Extraction

Three feature extraction techniques – Deep Features, Geometric Representation and Oriented FAST and Rotated BRIEF (ORB) [115] – were extracted and fused for genetic syndrome recognition.

We employed a Deep Convolutional Neural Network (CNN) architecture based on the very deep network with the triplet loss function [107], which was trained on 2622 identities of 2.6 million images. The CNN model receives images with size of 224×224 pixels as input. The model consists of 3×3 convolution kernels with stride 1, which are followed by non-linear rectification layers (ReLU), and 3 fully connected layers. This model does not use local contrast normalization. Then, the deep features are extracted from the cropped facial images through this CNN model, forming a feature vector of length 2622.



(a) (b) (c)

Figure 6.2: Feature extraction. (a)-(b) twenty seven facial fiducial points detected; (c) twenty one distances extracted from the twenty seven landmarks.

A geometric representation is created from the detected facial fiducial points. The geometric descriptor employs twenty seven 2D facial landmarks: two points for the outer corner eyebrows (P_1, P_2), two points for the inner corner eyebrows (P_5, P_7), six points for the middle of the eyebrows (P_2, P_3, P_4, P_8, P_9, P_{10}), one point for the glabella (P_6), four points for the inner and outer corner of the eyes (P_{12}, P_{13}, P_{15}, P_{16}), one point for the root of nose (P_{14}), one point for the supratip (P_{17}), two points for the alars sidewalls (P_{18}, P_{19}), one point for the subnasale (P_{20}), two points for the mouth corners (P_{23}, P_{24}), two points for the top and bottom of the upper lip (P_{21}, P_{22}), two points for the top and bottom of the lower lip (P_{25}, P_{26}), and one point for the

gnathion (bottom of chin) (P_{27}). Figures 6.2 (a)-(b) shows the localization of the twenty-seven facial landmarks.

Thirty one geometric features are extracted from the specified twenty seven fiducial points. From these thirty one geometric features, we calculated twenty one distances, as illustrated in Figure 6.2 (c), which are normalized to the face width in order to ensure the features to be scale invariant. The eleven distances d_2, d_3, d_4, d_5, d_6, d_9, d_{10}, d_{13}, d_{14}, d_{15} and d_{21} represent the average values of the two mirrored distances on the left and right sides of the face. The distance d_{14} is computed using the intersection point of the line between the points on the top of the upper lip and bottom of the lower lip, and the line between the left and right corners of the mouth. From the upper lip thickness d_{16} and the lower lip thickness d_{18}, we calculated the ratio between them ($R_{lips} = d_{16}/d_{18}$).

We also computed the curvature from the fiducial points of both eyebrows (P_1, P_2, P_3, P_4, P_5, P_7, P_8, P_9, P_{10}, P_{11}). For each eyebrow, we obtained three coefficients and their discriminant curve value, forming eight geometric features. Moreover, we calculated the subnasale angle $\alpha_{subnasale}$, generated by three points (P_{18}, P_{20}, P_{19}), where P_{20} is the central point, and P_{18} and P_{19} are the extreme points of the angle. Therefore, we form a geometric feature vector of total length 31.

Individuals with Cornelia de Lange syndrome frequently present a long philtrum, that is, the distance between the subnasale and the upper lip. Distance d_{12} represents this feature. They also have thin lips, which are described by the distance d_{16} and the ratio R_{lips} [39]. They usually exhibit a short upturned nose that is captured by the angle $\alpha_{subnasale}$ and nose distances d_9 to d_{11}. On the other hand, in the case of people with Progeria syndrome, they show a beak-shaped nose, which is also described by the previously mentioned nose features. Another common symptom of Progeria syndrome is a shrunken chin, being represented by the distance d_{19} [65].

People with Down syndrome often feature telecanthus, referring to the increased intercanthal distance, that is, the space between the inner corners of the eyes. The distance d_7 can represent this symptom. Another frequent symptom is a flattened nose, whose characteristic is covered by the subnasale angle and nose distances. Another common sign is a small mouth, which is described using the mouth distances d_{13}, d_{14}, d_{15}, d_{17} and d_{20}. The presence of upslanting palpebral fissures, that is, the distance between the lateral and medial canthus of the eyes, is also a symptom captured by the distance d_6 [30, 48].

For Treacher Collins syndrome recognition, it is essential to consider the chin length d_{19}. The most common clinical feature is the presence of a small lower jaw and chin. Another typical characteristic is a small upper jaw, which is represented by the philtrum distance d_{12}. Moreover, as well as Down syndrome symptom, a Treacher Collins symptom is the presence of short and down-slanting palpebral fissures (d_6) [19]. This symptom also characterizes Apert syndrome. Individuals with this syndrome also present a broad and short nose with a bulbous tip, which can be described using the subnasale angle and nose distances [20].

People with Angelman syndrome have a prominent chin, whose feature is captured by the distance d_{19}. They also have a wide mouth, featured by the mouth distances [18]. People with Williams syndrome also present a wide mouth. However, they show a small chin. Another Williams syndrome feature is a short nose with broad nasal tip, which can be represented by the nose distances. As Cornelia de Lange syndrome, William syndrome features a long philtrum (d_{12}) [50]. Moreover, regarding Fragile X syndrome, people with this syndrome have a prominent

forehead, a long and thin face, and a large jaw [60], which are covered by several facial distances.

ORB is a feature descriptor based on Features from Accelerated Segment Test (FAST) keypoint detector and Robust Independent Elementary Features (BRIEF) descriptor, which appeared as a fast and efficient alternative to Scale Invariant Feature Transform (SIFT), Speed up Robust Feature (SURF) in computation cost and matching performance. For generating ORB features, we selected fifteen facial keypoints previously detected: two points for the middle of the eyebrows (P_1, P_2), four points for the inner and outer corner of the eyes (P_3, P_4, P_6, P_7), one point for the root of nose (P_5), one point for the supratip (P_8), two points for the alars sidewalls (P_9, P_{10}), one point for the columella (P_{11}), two points for the mouth corners (P_{13}, P_{14}), two points for the top of the upper lip and bottom of the lower lip (P_{12}, P_{15}), and one point for the bottom of chin (P_{16}). Figures 6.2 (a)-(b) show the localization of the sixteen facial points. Then, an ORB feature vector is computed for each keypoint. The resulting ORB descriptor is formed by the concatenation of the generated vector for each keypoint.

For several computational problems, it has been demonstrated that recognition accuracy can enhance by fusing feature descriptors [150]. After extracting the deep, geometric and ORB features, the combined feature vector is set to 3069 features and then scaled.

6.2.3 Feature Reduction and Classification

We followed two approaches, PCA and PCA+LDA, for performing feature reduction. Firstly, Principal Component Analysis (PCA) was employed over the resulting feature vector set, obtaining the principal feature vectors. Finally, Linear Discriminant Analysis (LDA) was applied over the PCA reduced feature vector, forming a new reduced feature space.

For the classification stage, we used Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR) and Gaussian Naïve Bayes (GNB) classifiers to compare the accuracy rates for the genetic syndrome recognition. After performing dimensionality reduction, the classifiers are trained with reduced feature vectors. We applied the same validation protocol as the one used in the syndrome data set from Ferry et al. [54], splitting the training and testing sets with a 4:1 ratio and obtaining a classification average from 10 repeats.

6.3 Experimental Results

The proposed methodology was tested on the Diagnostically Relevant Facial Gestalt Information from Ordinary Photos Database, collected by Ferry et al. [54]. This data set is composed of 1499 ordinary and spontaneous photographs of different patients diagnosed into one of eight genetic syndromes: Angelman (205), Apert (200), Cornelia de Lange (250), Down (197), Fragile X (163), Progeria (150), Treacher Collins (103), and Williams-Beuren (231). This data set comprises facial pose variations, illumination deviations, low-resolution photographs, different backgrounds, and occlusions, such as glasses, hair, scarves, hand gestures, among others. The subjects present in this data set are of different ethnicities, genders, and ages, including children, adolescents, and adults.

For obtaining images of healthy controls, we employ the Dartmouth Database of Children's Faces [35]. This data set contains 40 male and 40 female children between 6 and 16 years of age. Models were photographed from different angles and lighting conditions, and also performing

eight facial expressions: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. From this image collection, we randomly selected 187 images, that is, the average number of images per genetic syndrome class, considering females and males who exhibit distinct pose deviations, facial expressions, and illumination variations. It is worth mentioning that we made this decision due the fact that children present the same facial musculature as adults [44].

Therefore, our data set is composed of 1686 (= 1499+187) images in total to verify our recognition method. Then, we conducted experiments using our geometric representation, deep features, ORB features and the fusion of both features through the following approaches: PCA+KNN, PCA+LDA+KNN, PCA+SVM, PCA+LDA+SVM, PCA+LR, PCA+LDA+LR, PCA+GNB, and PCA+LDA+GNB.

We validated our methodology following the same protocol used in the Diagnostically Relevant Facial Gestalt Information from Ordinary Photos Database [54], that is, we randomly select 80% of samples of each class for the training set and the remaining 20% for the testing set. The results are reported in Table 6.1, whose values represent the average recognition accuracy rates obtained through the execution of ten repeats.

Table 6.1: Average accuracy (%) using our geometric, ORB and deep features, and the fusion between them on the evaluated data set.

Recognition Method	Geometric (%)	ORB (%)	Deep Features (%)	Geometric + ORB + Deep Features (%)
KNN	50.62	46.86	71.76	73.20
PCA+KNN	50.78	48.43	72.81	74.87
PCA+LDA+KNN	58.33	62.65	85.10	**88.33**
SVM	64.48	65.23	85.42	88.69
PCA+SVM	64.57	66.34	85.62	88.56
PCA+LDA+SVM	65.03	66.24	85.65	**89.08**
LR	65.23	62.97	84.18	88.53
PCA+LR	65.52	65.03	85.95	89.80
PCA+LDA+LR	64.87	67.29	86.31	**90.16**
GNB	51.73	54.44	63.01	62.97
PCA+GNB	57.94	53.66	45.75	46.76
PCA+LDA+GNB	65.82	67.29	86.86	**90.29**

From our experiments, we can observe that the fusion of deep features with the geometric and ORB representation provides a high accuracy rate for genetic syndrome recognition. We can also notice that the individual use of deep features enables reaching an accuracy of about 85%, which is much superior to just using geometric or ORB representation independently. It is also shown that following a PCA+LDA approach provides increasing recognition rates. Table 6.2 shows the best detection accuracy rates achieved with the fusion feature set proposed in our approach and the methodology developed by Ferry et al. [54]. The accuracy rates were obtained using the validation protocol by Ferry et al. [54]. We can see that the proposed approach reaches

competitive results.

Table 6.2: Comparison of average accuracy rates (%) for genetic syndrome recognition.

Recognition Method	Strategy	Accuracy (%)
Our method	Deep, geometric and ORB representation + PCA+LDA+GNB	90.29
Ferry et al. [54]	Appearance and shape descriptors + SVM	93.10

6.4 Final Considerations

Experimental results demonstrated that the use of geometric and ORB representation allowed to enhance the discriminative power of the deep features. Our approach also proved to be robust for recognizing genetic syndromes in ordinary photographs in the presence of different occlusions, for instance, facial expressions, glasses, facial pose deviations, among others. Although the geometric and ORB representation did not provide a higher recognition rate individually, their fusion achieved a higher accuracy rate with deep features. Furthermore, PCA and LDA approaches, as well as the reduction and selection of discriminative features, allowed to increase the recognition rates significantly.

Chapter 7

Conclusions and Future Work

In this doctoral work, we investigated facial expression patterns to perform emotion recognition tasks. We focused on developing techniques for improving accuracy rates in automatic recognition of occluded facial expressions, dynamic facial expressions, and audio-visual emotions. Moreover, we studied how emotion recognition can be applied to address other problems, in particular, recognition of Down syndrome and other genetic disorders.

In Chapter 2, we introduced the Census Transform Histogram as an accurate visual descriptor for solving the emotion recognition problem in static images, performing well to recognize emotions over occluded and non-occluded facial expressions. This approach focused on identifying six emotions: anger, disgust, fear, happiness, sadness, and surprise. Our experiments for facial expression recognition with occlusions achieved higher accuracy rates on both CK+ and JAFFE data sets using the CENTRIST operator, as well as the PCA+LDA approach. The highest accuracy rates for CK+ and JAFFE data sets were 90.30% and 92.00%, respectively. For non-occluded facial expression experiments, the CENTRIST operator also demonstrated to be robust on both data sets (CK+: 94.10%, and JAFFE: 92.00%). However, LBP was slightly superior on the JAFFE data set, achieving 93.00%. This descriptor proved its strength, as well as its agile implementation and computational speed. In this work, we also noticed the importance of following a PCA+LDA approach as it significantly increases the recognition rates.

In Chapter 3, in order to address persistent known issues when working with 2D images, that is, lack of temporal and other information mentioned in Section 1.1, we explored and introduced new techniques for video representation: Visual Rhythms and Motion History Images. Such approach also classified the same classes as the static emotion recognition method. Our experiments demonstrate superior results by fusing the WLD features extracted from the Motion History Image and Visual Rhythms, reaching a recognition rate of about 76.40% on the BU-4DFE data set. This result may be less accurate compared to other related works in the literature, however, this method differs from others because the entire process is performed automatically, while many of the existing ones during their process execute tasks manually, such as selecting keyframes or facial landmarks.

In Chapter 4, we proposed a deep hybrid neural network to perform audio-visual emotion recognition, since speech and facial expressions represent more natural and effective channels for revealing emotions. This approach recognized the same categories as the other recognizing emotion methods. We employed an image representation based on log Mel-spectrogram to extract audio features, whereas two image representations based on CLAHE and Census Transform were

used to extract visual features. These visual representations served as input to 2D Convolutional Neural Networks, separately. Experimental results exhibited competitive accuracy rates for unimodality emotion recognition based on visual features, achieving 75.00% for RML, 68.33% for eNTERFACE05, and 59.52% for BAUM-1s data set. Unimodality results using audio segments achieved 64.38%, 62.00%, and 40.76% for RML, eNTERFACE05, and BAUM-1s data sets, respectively. By merging audio and visual features, there is a significant improvement in the emotion recognition rates (RML: 82.50%, eNTERFACE05: 85.00%, and BAUM-1s: 59.70%). We could observe that adding audio features to facial expression recognition helped enhance multimodal emotion recognition, which answers our first research question. Also, it is important to denote that our approach achieved higher audio-visual emotion recognition accuracy in the spontaneous BAUM-1s dataset.

In Chapter 5, we introduced a method for recognizing Down syndrome in facial photos. We extracted features based on our emotion recognition approaches, explicitly using the CENTRIST and geometric descriptors, and the PCA+LDA procedure. Our experiments achieved an accuracy rate of 98.39% using samples of patients diagnosed with Down Syndrome from the *Diagnostically Relevant Facial Gestalt Information from Ordinary Photos Database*, and healthy controls from the *Dartmouth Database of Children's Faces*. Our method was shown to be robust and superior to other methods in the literature due to its small geometric feature set and its simple implementation to recognize Down syndrome through photographs in the wild.

In Chapter 6, we proposed a methodology for genetic disorder recognition through facial photographs based on deep features. Our approach recognized eight different genetic disorders: Angelman, Apert, Cornelia de Lange, Down, Fragile X, Progeria, Treacher Collins, and Williams-Beuren. We employed the *Diagnostically Relevant Facial Gestalt Information from Ordinary Photos Database* to extract images of people diagnosed with genetic syndromes, while *Dartmouth Database of Children's Faces* was used for healthy controls. We achieved a recognition rate of 90.93% by merging the ORB, geometric, and deep features. Deep features were extracted from a 2D CNN fine tuned and pre-trained on a large facial data set. The experimental result is very competitive compared to the approach developed by Ferry et al. [54].

Our approaches proposed in Chapters 5 and 6 to recognize genetic disorders constitute a useful and valuable tool in the medical field, where they could help specialists to provide an early diagnosis, as well as patients who do not have easy access to specialists to receive a diagnosis and seek out an expert as soon as possible.

We specifically used the same facial landmarks and geometric features employed in the facial expression recognition as the basis for constructing the feature space to perform the genetic disorder recognition, in conjunction with the CENTRIST application. This answers our second research question.

Regarding the third research question, we could observe that deep features clearly improved the genetic disorder recognition task than just using hand-crafted features. Deep features are complementary to other features. In addition, deep and audio features allowed higher accuracy rates than the use of hand-crafted and visual features for the emotion recognition problem.

As proposals for future work, we aim to explore new data sets, such as 4D data (3D images + temporal information), to perform dynamic emotion recognition. With this kind of data, there is no need to deal with issues regarding lighting and facial pose variations. It brings some advantages, for example, a machine learning algorithm does not require to learn viewpoint

invariance, and as a result, it demands less data during the training stage. In addition, since 3D images allow to represent geometry in terms of three-dimensional coordinates, this offers some advantages, for example, the size of an object in a scene is computed directly from the 3D coordinates, being a faster and more accurate calculation.

Furthermore, we expect to seek other feature extraction techniques and neural network architectures to allow a considerable reduction in computational time in order to define directions for performing online emotion recognition. For audio-visual emotion recognition, we also intend to explore lightweight convolutional neural networks for the development of an automatic system for recognizing emotions in real time.

Bibliography

[1] K. Aldridge, I. D. George, K. K. Cole, J. R. Austin, T. N. Takahashi, Y. Duan, and J. H. Miles. Facial Phenotypes in Subgroups of Prepubertal Boys with Autism Spectrum Disorders are Correlated with Clinical Phenotypes. *Molecular Autism*, 2(1):15, 2011.

[2] M. Alhussein. Automatic Facial Emotion Recognition using Weber Local Descriptor for e-Healthcare System. *Cluster Computing*, 19(1):99–108, 2016.

[3] N. Ambady and R. Rosenthal. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological Bulletin*, 111(2):256, 1992.

[4] American Pscychological Association (APA). Reading Facial Expression of Emotion, 2011. `http://www.apa.org/science/about/psa/2011/05/facial-expressions.aspx`.

[5] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava. 4-D Facial Expression Recognition by Learning Geometric Deformations. *IEEE Transactions on Cybernetics*, 44(12):2443–2457, 2014.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental Face Alignment in the Wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[7] M. Bejani, D. Gharavian, and N. M. Charkari. Audiovisual Emotion Recognition using ANOVA Feature Selection Method and Multi-Classifier Neural Networks. *Neural Computing and Applications*, 24(2):399–412, 2014.

[8] S. Berretti, A. Del Bimbo, and P. Pala. Real-Time Expression Recognition from Dynamic Sequences of 3D Facial Scans. In *5th Eurographics Conference on 3D Object Retrieval*, pages 85–92. Eurographics Association, 2012.

[9] A. F. Bobick and J. W. Davis. The Recognition of Human Movement using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[10] S. Boehringer, T. Vollmar, C. Tasse, R. P. Wurtz, G. Gillessen-Kaesbach, B. Horsthemke, and D. Wieczorek. Syndrome Identification based on 2D Analysis Software. *European Journal of Human Genetics*, 14(10):1082–1089, 2006.

[11] K. Burçin and N. V. Vasif. Down Syndrome Recognition using Local Binary Patterns and Statistical Evaluation of the System. *Expert Systems with Applications*, 38(7):8690–8695, 2011.

[12] S. Canavan, Y. Sun, X. Zhang, and L. Yin. A Dynamic Curvature based Approach for Facial Activity Analysis in 3D Space. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 14–19. IEEE, 2012.

[13] E. Candés, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(3):1–37, May 2011.

[14] L. Cao, K. S. Chua, W. Chong, H. Lee, and Q. Gu. A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in Support Vector Machine. *Neurocomputing*, 55(1):321–336, 2003.

[15] J. Chen, Z. Chen, Z. Chi, and H. Fu. Facial Expression Recognition in Video with Multiple Feature Fusion. *IEEE Transactions on Affective Computing*, 9(1):38–50, 2018.

[16] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. WLD: A Robust Local Image Descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, 2010.

[17] M. Chung, J. Lee, H. Kim, S. Song, and W. Kim. Automatic Video Segmentation based on Spatio-Temporal Features. *Korea Telecom Journal*, 4(1):4–14, 1999.

[18] J. Clayton-Smith and L. Laan. Angelman Syndrome: A Review of the Clinical and Genetic Aspects. *Journal of Medical Genetics*, 40(2):87–95, 2003.

[19] A. R. Cobb, B. Green, D. Gill, P. Ayliffe, T. W. Lloyd, N. Bulstrode, and D. J. Dunaway. The Surgical Management of Treacher Collins Syndrome. *British Journal of Oral and Maxillofacial Surgery*, 52(7):581–589, 2014.

[20] M. M. Cohen and S. Kreiborg. A Clinical Study of the Craniofacial Features in Apert Syndrome. *International Journal of Oral and Maxillofacial Surgery*, 25(1):45–53, 1996.

[21] G. W. Corder and D. I. Foreman. *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons, 2014.

[22] J. Cornejo and H. Pedrini. Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks. In *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 111–117. IEEE, 2019.

[23] J. Y. R. Cornejo and H. Pedrini. Automatic Fruit and Vegetable Recognition based on CENTRIST and Color Representation. In *Iberoamerican Congress on Pattern Recognition*, pages 76–83. Springer, Cham, 2016.

[24] J. Y. R. Cornejo and H. Pedrini. Recognition of Occluded Facial Expressions based on CENTRIST Features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1298–1302. IEEE, 2016.

[25] J. Y. R. Cornejo and H. Pedrini. Emotion Recognition Based on Occluded Facial Expressions. In *International Conference on Image Analysis and Processing*, pages 309–319. Springer, 2017.

[26] J. Y. R. Cornejo and H. Pedrini. Emotion Recognition From Occluded Facial Expressions using Weber Local Descriptor. In *25th International Conference on Systems, Signals and Image Processing*, pages 1–5, Maribor, Slovenia, June 2018.

[27] J. Y. R. Cornejo and H. Pedrini. Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3396–3402, Oct. 2019.

[28] J. Y. R. Cornejo and H. Pedrini. Dynamic Facial Expression Recognition Based on Visual Rhythms and Motion History Images. In *15th International Conference on Machine Learning and Data Mining (MLDM)*, volume II, pages 599–613, New York, NY, USA, 2019.

[29] J. Y. R. Cornejo, H. Pedrini, and F. Flórez-Revuelta. Facial Expression Recognition with Occlusions Based on Geometric Representation. In *20th Iberoamerican Congress on Pattern Recognition*, volume 9423, pages 263–270, Montevideo, Uruguay, Nov. 2015. Springer.

[30] J. Y. R. Cornejo, H. Pedrini, A. Machado-Lima, and F. L. Santos Nunes. Down Syndrome Detection based on Facial Features using a Geometric Descriptor. *Journal of Medical Imaging*, 4(4):044008_1–044008_6, 2017.

[31] A. Correa, J. D. Cragan, J. E. Kucik, C. J. Alverson, S. M. Gilboa, R. Balakrishnan, M. J. Strickland, C. Duke, L. A. O'Leary, and T. Riehle-Colarusso. Reporting Birth Defects Surveillance Data 1968-2003. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 79(2):65, 2007.

[32] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[33] A. L. Creavin and R. D. Brown. Ophthalmic Abnormalities in Children with Down Syndrome. *Journal of Pediatric Ophthalmology and Strabismus*, 46(2):76–82, 2009.

[34] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, Washington, DC, USA, 2005. IEEE.

[35] K. A. Dalrymple, J. Gomez, and B. Duchaine. The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. *PloS One*, 8(11):e79131, 2013.

[36] C. Darwin. *The Expression of Emotion in Man and Animals*. New York: Oxford University Press, 1872-1998.

[37] A. David and B. Lerner. Support Vector Machine-based Image Classification for Genetic Syndrome Diagnosis. *Pattern Recognition Letters*, 26(8):1029–1038, 2005.

[38] J. W. Davis and A. F. Bobick. The Representation and Recognition of Human Movement using Temporal Templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928–934. IEEE, 1997.

[39] M. A. Deardorff, S. E. Noon, and I. D. Krantz. *Cornelia de Lange Syndrome*. University of Washington, Seattle, 2016.

[40] H. Deng, L. Jin, L. Zhen, and J. Huang. A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA. *International Journal of Information Technology*, 11(5):86–96, 2005.

[41] O. Déniz, G. Bueno, J. Salido, and F. De la Torre. Face Recognition using Histograms of Oriented Gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.

[42] K. Doi. Computer-aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Computerized Medical Imaging and Graphics*, 31(4):198–211, 2007.

[43] P. Ekman. Facial Expressions. *Handbook of Cognition and Emotion*, 16:301–320, 1999.

[44] P. Ekman and H. Oster. Facial Expressions of Emotion. *Annual Review of Psychology*, 30(1):527–554, 1979.

[45] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*, 44(3):572–587, 2011.

[46] N. E. D. Elmadany, Y. He, and L. Guan. Multiview Emotion Recognition via Multi-Set Locality Preserving Canonical Correlation Analysis. In *IEEE International Symposium on Circuits and Systems*, pages 590–593. IEEE, 2016.

[47] N. E. D. Elmadany, Y. He, and L. Guan. Multiview Learning via Deep Discriminative Canonical Correlation Analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2409–2413. IEEE, 2016.

[48] C. J. Epstein, J. R. Korenberg, G. Annerén, S. E. Antonarakis, S. Aymé, E. Courchesne, L. B. Epstein, A. Fowler, Y. Groner, and J. L. Huret. Protocols to Establish Genotype-Phenotype Correlations in Down Syndrome. *American Journal of Human Genetics*, 49(1):207–235, 1991.

[49] O. Erogul, M. E. Sipahi, Y. Tunca, and S. Vurucu. Recognition of Down Syndromes using Image Analysis. In *14th National Biomedical Engineering Meeting*, pages 1–4. IEEE, 2009.

[50] A. K. Ewart, C. A. Morris, D. Atkinson, W. Jin, K. Sternes, P. Spallone, A. D. Stock, M. Leppert, and M. T. Keating. Hemizygosity at the Elastin Locus in a Developmental Disorder, Williams Syndrome. *Nature Genetics*, 5(1):11, 1993.

[51] X. Fan and T. Tjahjadi. A Dynamic Framework based on Local Zernike Moment and Motion History Image for Facial Expression Recognition. *Pattern Recognition*, 64:399–406, 2017.

[52] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris. 4D Facial Expression Recognition. In *IEEE International Conference on Computer Vision Workshops*, pages 1594–1601. IEEE, 2011.

[53] R. Faragher and B. Clarke. *Educating Learners with Down Syndrome: Research, Theory, and Practice with Children and Adolescents*. Routledge, 2013.

[54] Q. Ferry, J. Steinberg, C. Webber, D. R. FitzPatrick, C. P. Ponting, A. Zisserman, and C. Nellåker. Diagnostically Relevant Facial Gestalt Information from Ordinary Photos. *Elife*, 3, 2014.

[55] R. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[56] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2013.

[57] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11):964–971, 1987.

[58] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast Algorithms for Recovering a Corrupted Low-Rank Matrix. In *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 213–216, Dec. 2009.

[59] L. Gao, L. Qi, and L. Guan. Information Fusion based on Kernel Entropy Component Analysis in Discriminative Canonical Correlation Space with Application to Audio Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2817–2821. IEEE, 2016.

[60] K. B. Garber, J. Visootsak, and S. T. Warren. Fragile X Syndrome. *European Journal of Human Genetics*, 16(6):666, 2008.

[61] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, J. E. Garrido, and E. L.-C. Law. Multimodal Affective Computing to Enhance the User Experience of Educational Software Applications. *Mobile Information Systems*, pages 1–10, 2018.

[62] V. Gay, P. Leijdekkers, J. Agcanas, F. Wong, and Q. Wu. CaptureMyEmotion: Helping Autistic Children Understand their Emotions Using Facial Expression Recognition and Mobile Technologies. *Studies in Health Technology and Informatics*, 189:71–76, 2013.

[63] C. Guyon, T. Bouwmans, and E. Zahzah. Robust Principal Component Analysis for Background Subtraction: Systematic Evaluation and Comparative Analysis. In P. Sanguansat, editor, *Principal Component Analysis*, pages 223–238. INTECH, Mar. 2012.

[64] R. M. Haralick and K. Shanmugam. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

[65] R. Hennekam. Hutchinson–Gilford Progeria Syndrome: Review of the Phenotype. *American Journal of Medical Genetics Part A*, 140(23):2603–2624, 2006.

[66] F. Hickey, E. Hickey, and K. L. Summar. Medical Update for Children with Down Syndrome for the Pediatrician and Family Practitioner. *Advances in Pediatrics*, 59(1):137–157, 2012.

[67] C.-P. Huang, C.-H. Hsieh, K.-T. Lai, and W.-Y. Huang. Human Action Recognition using Histogram of Oriented Gradient of Motion History Image. In *First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pages 353–356. IEEE, 2011.

[68] Z. Huang. Pedestrian Detection Algorithm in Video Analysis Based on Centrist. In *International Conference on Intelligent Transportation, Big Data & Smart City*, pages 117–120. IEEE, 2016.

[69] Y.-S. Hwang, J.-C. Kwak, and K.-Y. Lee. Implementation of a Pedestrian Detection Device based on CENTRIST for an Embedded Environment. *Advanced Science and Technology Letters*, 46:123–127, 2014.

[70] T. Jabid, M. H. Kabir, and O. Chae. Robust Facial Expression Recognition Based on Local Directional Pattern. In *27th Conference on Image and Vision Computing*, pages 464–468, New Zealand, 2012.

[71] L. A. Jeni, A. Lőrincz, T. Nagy, Z. Palotai, J. Sebők, Z. Szabó, and D. Takács. 3D Shape Estimation in Video Sequences Provides High Precision Evaluation of Facial Expressions. *Image and Vision Computing*, 30(10):785–795, 2012.

[72] B. Jiang and K. Jia. Research of Robust Facial Expression Recognition under Facial Occlusion Condition. In *Active Media Technology, Lecture Notes in Computer Science*, volume 6890, pages 92–100. Springer-Verlag Berlin Heidelberg, 2011.

[73] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2002.

[74] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion. *Image and Vision Computing*, 65:66–75, 2017.

[75] D. E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[77] P. Kruszka, A. R. Porras, A. K. Sobering, and F. A. Ikolo. Down Syndrome in Diverse Populations. *American Journal of Medical Genetics Part A*, 173(1):42–53, 2017.

[78] D. Landry, D. Raiside, and J. Vanhoutte. On the Classification of Congenital Abnormalities from Hand Radiographs. *Pattern Recognition*, 11(4):289–296, 1979.

[79] Q. Li and R. M. Nishikawa. *Computer-aided Detection and Diagnosis in Medical Imaging*. Taylor & Francis, 2015.

[80] W. Lim, D. Jang, and T. Lee. Speech Emotion Recognition using Convolutional and Recurrent Neural Retworks. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE, 2016.

[81] K. C. Lin, T.-C. Huang, J. C. Hung, N. Y. Yen, and S. J. Chen. facial rmotion recognition towards affective computing-based learning.

[82] Z. Lin, M. Chen, L. Wu, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. UIUC Technical Report UILU-ENG-09-2215, University of Illinois at Urbana-Champaign, Oct. 2009.

[83] S. Liu, Y. Zhang, and K. Liu. Facial Expression Recognition under Partial Occlusion based on Weber Local Descriptor Histogram and Decision Fusion. In *33rd Chinese Control Conference*, pages 4664–4668. IEEE, 2014.

[84] S. Liu, Y. Zhang, and K. Liu. Facial Expression Recognition under Random Block Occlusion based on Maximum Likelihood Estimation Sparse Representation. In *International Joint Conference on Neural Networks*, pages 1285–1290, July 2014.

[85] H. S. Loos, D. Wieczorek, R. P. Würtz, C. von der Malsburg, and B. Horsthemke. Computer-based Recognition of Dysmorphic Faces. *European Journal of Human Genetics*, 11(8):555, 2003.

[86] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *DARPA Image Understanding Workshop*, pages 121–130. Vancouver, BC, Canada, 1981.

[87] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A Complete Facial Expression Dataset for Action Unit and Emotion-Specified Expression. In *Third IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis*, San Francisco, CA, USA, June 2010.

[88] M. Lyons, M. Kamachi, and J. Gyoba. Japanese Female Facial Expressions (JAFFE), 1997. Database of Digital Images.

[89] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir. Audio-Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach. *Information Fusion*, 46:184–192, 2019.

[90] A. Majumder, L. Behera, and V. K. Subramanian. Emotion Recognition from Geometric Facial Features using Self-Organizing Map. *Pattern Recognition*, 47(3):1282–1293, 2014.

[91] E. Malt, R. Dahl, T. Haugsand, I. Ulvestad, N. Emilsen, B. Hansen, Y. Cardenas, R. Skøld, A. Thorsen, and E. Davidsen. Health and Disease in Adults with Down Syndrome. *Tidsskrift for den Norske Laegeforening: Tidsskrift for Praktisk Medicin, ny Raekke*, 133(3):290–294, 2013.

[92] H. B. Mann and D. R. Whitney. On a Test of whether one of two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar. 1947.

[93] M. Mansoorizadeh and N. M. Charkari. Multimodal Information Fusion Application to Human Emotion Recognition from Face and Speech. *Multimedia Tools and Applications*, 49(2):277–297, 2010.

[94] X. Mao, Y. L. Xue, Z. Li, K. Huang, and S. Lv. Robust Facial Expression Recognition based on RPCA and AdaBoost. *10th Workshop on Image Analysis for Multimedia Interactive Services*, pages 113–116, May 2009.

[95] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE'05 Audio-Visual Emotion Database. In *22nd International Conference on Data Engineering Workshops*, pages 8–8. IEEE, 2006.

[96] A. M. Martínez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[97] T. Moreira, D. Menotti, and H. Pedrini. First-Person Action Recognition Through Visual Rhythm Texture Description. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2627–2631, New Orleans, LA, USA, Mar. 2017.

[98] M. Morningstar, E. E. Nelson, and M. A. Dirks. Maturation of Vocal Emotion Recognition: Insights from the Developmental and Neuroimaging Literature. *Neuroscience & Biobehavioral Reviews*, 2018.

[99] M. Nazir, Z. Jan, and M. Sajjad. Facial Expression Recognition using Weber Discrete Wavelet Transform. *Journal of Intelligent & Fuzzy Systems*, 33(1):479–489, 2017.

[100] M. C. L. Neves, F. Tremeau, R. Nicolato, H. Lauar, M. A. Romano-Silva, and H. Correa. Facial Emotion Recognition Deficits in Relatives of Children with Autism are not Associated with 5HTTLPR. *Brazilian Journal of Psychiatry*, 33(3):261–267, 2011.

[101] V. Nieratschker, C. Brückmann, and C. Plewnia. CACNA1C Risk Variant Affects Facial Emotion Recognition in Healthy Individuals. *Scientific Reports*, 5:17349, 2015.

[102] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-Visual Emotion Recognition in Video Clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2017.

[103] T. Ojala, M. Pietikhenl, and D. Harwood. Performance Evaluation of Texture Measures with Classification based on Kullback Discrimination of Distributions. *International Conference on Pattern Recognition*, pages 582–585, 1994.

[104] W. H. Organization. *International Statistical Classification of Diseases and Related Health Problems*, volume 1. World Health Organization, 2004.

[105] M. Pantic. Facial Expression Recognition. In *The Encyclopedia of Biometrics*, pages 1–8. Springer, 2nd edition, 2014.

[106] M. S. Park, J. H. Na, and J. Y. Choi. PCA-based Feature Extraction using Class Information. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 341–345. IEEE, 2005.

[107] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *British Machine Vision Conference*, volume 1, pages 1–12, 2015.

[108] D. Patterson. Molecular Genetic Analysis of Down Syndrome. *Human Genetics*, 126(1):195–214, 2009.

[109] R. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[110] A. Pinto, W. Schwartz, H. Pedrini, and A. Rocha. Using Visual Rhythms for Detecting Video-based Facial Spoof Attacks. *IEEE Transactions on Information Forensics and Security*, 10(5):1025–1038, Oct. 2015.

[111] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the Limits of Stereo using Variational Stereo Estimation. In *IEEE Intelligent Vehicles Symposium*, pages 401–407. IEEE, 2012.

[112] M. Reale, X. Zhang, and L. Yin. Nebula Feature: A Space-Time Feature for Posed and Spontaneous 4D Facial Behavior Analysis. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2013.

[113] W. Reardon and D. Donnai. Dysmorphology Demystified. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 92(3):F225–F229, 2007.

[114] J. A. Rojas, A. Ramirez, and O. Chae. Facial Expression Recognition Based On Local Sign Directional Pattern. In *19th IEEE International Conference on Image Processing*, pages 2613–2616, 2012.

[115] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An Efficient Alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571, Nov. 2011.

[116] S. Ryoo and J.-K. Chang. Emotion Affective Color Transfer Using Feature Based Facial Expression Recognition. *Advanced Science and Technology Letters*, 39:131–135, 2013.

[117] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3D Facial Expression Dynamics. *Image and Vision Computing*, 30(10):762–773, 2012.

[118] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-Temporal Covariance Descriptors for Action and Gesture Recognition. In *IEEE Workshop on Applications of Computer Vision*, pages 103–110, 2013.

[119] Ş. Saraydemir, N. Taşpınar, O. Eroğul, H. Kayserili, and N. Dinçkan. Down Syndrome Diagnosis based on Gabor Wavelet Transform. *Journal of Medical Systems*, 36(5):3205–3213, 2012.

[120] R. R. Sarvestani and R. Boostani. FF-SKPCCA: Kernel Probabilistic Canonical Correlation Analysis. *Applied Intelligence*, 46(2):438–454, 2017.

[121] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 552–557. IEEE, 2009.

[122] C. Shan, S. Gong, and P. W. McOwan. Facial Expression Recognition based on Local Binary Patterns: A Comprehensive Study. *Image and Vision Computing*, 27(6):803–816, May 2009.

[123] K. G. Smitha and A. P. Vinod. Facial Emotion Recognition System for Autistic Children: A Feasible Study based on FPGA Implementation. *Medical & Biological Engineering & Computing*, 53(11):1221–1229, 2015.

[124] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian Detection using Infrared Images and Histograms of Oriented Gradients. In *IEEE Intelligent Vehicles Symposium*, pages 206–212. IEEE, 2006.

[125] Y. Tong, R. Chen, and Y. Cheng. Facial Expression Recognition Algorithm using LGC based on Horizontal and Diagonal Prior Principle. *Optik - International Journal for Light and Electron Optics*, 125(16):4186–4189, 2014.

[126] B. Torres and H. Pedrini. Detection of Complex Video Events through Visual Rhythm. *The Visual Computer*, 34(2):145–165, Feb. 2018.

[127] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5200–5204. IEEE, 2016.

[128] M. Valstar, M. Pantic, and I. Patras. Motion History for Facial Action Detection in Video. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 635–640. IEEE, 2004.

[129] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2001.

[130] T. Vollmar, B. Maus, R. P. Wurtz, G. Gillessen-Kaesbach, B. Horsthemke, D. Wieczorek, and S. Boehringer. Impact of Geometry and Viewing Angle on Classification Accuracy of 2D based Analysis of Dysmorphic Faces. *European Journal of Medical Genetics*, 51(1):44–53, 2008.

[131] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech Emotion Recognition using Fourier Parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, 2015.

[132] Y. Wang and L. Guan. Recognizing Human Emotional State from Audiovisual Signals. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008.

[133] Z. Wang, Q. Ruan, and G. An. Facial Expression Recognition using Sparse Local Fisher Discriminant Analysis. *Neurocomputing*, 174:756–766, 2016.

[134] Z. Wang, S. Wang, and Q. Ji. Capturing Complex Spatio-Temporal Relations among Facial Muscles for Facial Expression Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[135] J. Wu, H. I. Christensen, and J. M. Rehg. Visual Place Categorization: Problem, Dataset, and Algorithm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4763–4770. IEEE, 2009.

[136] J. Wu and J. M. Rehg. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.

[137] H. Xue and I. Gertner. Automatic Recognition of Emotions from Facial Expressions. In *Proceedings of SPIE - 9090, Automatic Target Recognition XXIV*, pages 90900O–90900O–12, Baltimore, MD, USA, 2014.

[138] M. Xue, A. Mian, W. Liu, and L. Li. Automatic 4D Facial Expression Recognition using DCT Features. In *IEEE Winter Conference on Applications of Computer Vision*, pages 199–206. IEEE, 2015.

[139] C. Yan, F. Coenen, and B. Zhang. Driving Posture Recognition by Joint Application of Motion History Image and Pyramid Histogram of Oriented Gradients. *International Journal of Vehicular Technology*, 2014:1–11, 2014.

[140] X. Yan, T. J. Andrews, and A. W. Young. Cultural Similarities and Differences in Perceiving and Recognizing Facial Expressions of Basic Emotions. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):423, 2016.

[141] N.-M. Yao, H. Chen, Q.-P. Guo, and H.-A. Wang. Non-Frontal Facial Expression Recognition Using a Depth-Patch Based Deep Neural Network. *Journal of Computer Science and Technology*, 32(6):1172–1185, 2017.

[142] Y. Yao, D. Huang, X. Yang, Y. Wang, and L. Chen. Texture and Geometry Scattering Representation-Based Facial Expression Recognition in 2D+3D Videos. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1s):18:1–18:23, Mar. 2018.

[143] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A High-Resolution 3D Dynamic Facial Expression Database. In *8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[144] R. Zabih and J. Woodfill. Non-Parametric Local Transforms for Computing Visual Correspondence. In *European Conference on Computer Vision*, pages 151–158. Springer, 1994.

[145] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[146] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2017.

[147] L. Zhang, D. Tjondronegoro, and V. Chandran. Random Gabor based Templates for Facial Expression Recognition in Images with Facial Occlusion. *Neurocomputing*, 145:451–464, 2014.

[148] S. Zhang, S. Zhang, T. Huang, and W. Gao. Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition. In *ACM International Conference on Multimedia Retrieval*, pages 281–284. ACM, 2016.

[149] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian. Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2018.

[150] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between Geometry-based and Gabor-Wavelets-based Facial Expression Recognition using Multi-Layer Perceptron. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–459. IEEE, 1998.

[151] Q. Zhao, K. Okada, K. Rosenbaum, D. J. Zand, R. Sze, M. Summar, and M. G. Linguraru. Hierarchical Constrained Local Model using ICA and its Application to Down Syndrome Detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 222–229. Springer, 2013.

[152] Q. Zhao, K. Rosenbaum, R. Sze, D. Zand, M. Summar, and M. G. Linguraru. Down Syndrome Detection from Facial Photographs using Machine Learning Techniques. In *SPIE Medical Imaging*, pages 867003–867003. International Society for Optics and Photonics, 2013.

[153] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu. Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[154] Q. Zhen, D. Huang, Y. Wang, and L. Chen. Muscular Movement Model-Based Automatic 3D/4D Facial Expression Recognition. *IEEE Transactions on Multimedia*, 18(7):1438–1450, 2016.

[155] K. Zuiderveld. Contrast Limited Adaptive Histogram Equalization. In P. Heckbert, editor, *Graphics Gems IV*, pages 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.