Numéro d'ordre : 272/2025-C/GE

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté de Génie Electrique

Thèse de Doctorat

Présentée pour l'obtention du **grade de Docteur**

En : Télécommunications

Spécialité : Télécommunications et Traitement de L'information

Par : FERHAT Roumiassa

Sujet

Reconnaissance des gestes de la main pour la compréhension de la parole

Soutenue publiquement, le 05/02/2025, devant le jury composé de :

M.	HAMZA Abdelkarim	Professeur	à l'USTHB	Président
Mme.	CHELALI Fatma Zohra	Professeur	à l'USTHB	Directeur de thèse
Mme.	MEKSEN Thouraya	Professeur	à l'USTHB	Examinateur
Mme.	OUARZEDDINE Mounira	Professeur	à l'USTHB	Examinateur
M.	KERDJIDJ Oussama	Maître de Recherche	à CDTA	Examinateur

Order Number: 272/2025-C/GE

People's DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

University of Science and Technology Houari Boumediene

Faculty of Electrical Engineering

جامعة هواري بومدين
للعلوم والتكنولوجيا
USTHB

PhD Thesis

Presented for the attainment of the **degree of Doctor**

In : Télécommunications

Specialty : Telecommunications and information processing

By : FERHAT Roumiassa

Title

Hand Gesture Recognition for Speech Comprehension

Publicly defended on February 5th 2025 in front of the jury composed of:

Mr.	HAMZA Abdelkarim	Professor	USTHB	President
Mme.	CHELALI Fatma Zohra	Professor	USTHB	Thesis supervisor
Mme.	MEKSEN Thouraya	Professor	USTHB	Examiner
Mme.	OUARZEDDINE Mounira	Professor	USTHB	Examiner
Mr.	KERDJIDJ Oussama	MRA	CDTA	Examiner

To my dear parents

To my beloved husband

To my precious children, Meriem and Mohammed

To all my family...

Acknowledgments

First and foremost, I would like to express my deepest gratitude to Almighty Allah, for His countless blessings, guidance, and strength throughout this journey. Without His grace, none of this would have been possible.

I am profoundly grateful to my parents, whose unwavering love, encouragement, and support have been a constant source of strength. Thank you for always believing in me and for laying the foundation that allowed me to reach this accomplishment.

To my beloved husband, thank you for your patience, understanding, and continuous motivation. Your support has been my anchor during this journey, and I am forever grateful.

To my wonderful children, you are my greatest joy and source of inspiration. Your smiles and love have kept me going, and I appreciate your understanding and patience throughout this time.

I would also like to express my sincere appreciation to my thesis advisor, Professor Chelali Fatima Zohra. Your insightful guidance, constructive feedback, and support have been invaluable in shaping this thesis. Thank you for your dedication and belief in my work.

I extend my heartfelt thanks to the members of my thesis jury: President Prof HAMZA Abdelkarim , as well as Prof MEKSEN Thoraya, Prof OUARZEDDINE Mounira and Dr. KERDJIDJ Oussama from CDTA. Your time, expertise, and valuable insights have greatly contributed to the quality of this work.

This thesis is the product of the collective support and love from all of you. From the depths of my heart, thank you.

Abstract

Hand gesture recognition has emerged as a crucial area in human-computer interaction, with applications in assistive technologies, virtual reality, and sign language interpretation. This thesis explores a bimodal approach combining speech and gesture recognition to enhance communication, particularly for individuals with hearing impairments. The study focuses on the integration of co-verbal gestures with speech processing, leveraging deep learning techniques to improve recognition accuracy.

A key contribution of this work is the development of a multimodal dataset containing Arabic speech and dynamic hand gestures, recorded in real-world conditions. Advanced machine learning models, including Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs), were employed to classify gestures and speech patterns. A novel descriptor, the Modified Local Optimal-Oriented Pattern (Mi-LOOP), is proposed to enhance feature extraction in gesture recognition.

Experimental results demonstrate the effectiveness of the proposed approach in recognizing speech-gesture correlations, significantly improving communication accessibility for hearing-impaired individuals. The findings also contribute to the broader field of multimodal interaction, paving the way for more intuitive and adaptive human-machine interfaces.

Résumé

La reconnaissance des gestes de la main est devenue un domaine clé dans l'interaction homme-machine, avec des applications dans les technologies d'assistance, la réalité virtuelle et l'interprétation de la langue des signes. Cette thèse explore une approche bimodale combinant la reconnaissance vocale et gestuelle afin d'améliorer la communication, notamment pour les personnes malentendantes. L'étude met l'accent sur l'intégration des gestes co-verbaux avec le traitement de la parole, en s'appuyant sur des techniques d'apprentissage profond pour améliorer la précision de reconnaissance.

L'un des principaux apports de ce travail est le développement d'un jeu de données multimodal contenant des enregistrements de la langue arabe et des gestes de la main dynamiques, captés dans des conditions réelles. Des modèles avancés d'apprentissage automatique, tels que les réseaux de neurones convolutifs (CNN) et les machines à vecteurs de support (SVM), ont été utilisés pour classifier les gestes et les schémas vocaux. Un nouveau descripteur, le Modified Local Optimal-Oriented Pattern (Mi-LOOP), est proposé pour améliorer l'extraction des caractéristiques dans la reconnaissance gestuelle.

Les résultats expérimentaux montrent l'efficacité de l'approche proposée pour reconnaître les corrélations entre la parole et les gestes, améliorant ainsi considérablement l'accessibilité à la communication pour les personnes malentendantes. Ces travaux contribuent également au domaine plus large de l'interaction multimodale, ouvrant la voie à des interfaces homme-machine plus intuitives et adaptatives.

Contents

Acknowledgments **v**

Abstract **vi**

Résumé **vii**

Contents **xi**

Table of figures **xiii**

List of Abbreviations **xv**

Introduction **1**

1 Literature Survey of gesture and speech recognition system **6**

1.1 Introduction 6

1.2 Gesture Recognition System 7

1.3 Speech Recognition System 13

1.4 Bimodal Gesture/Speech Recognition System 16

1.5 Conclusion 17

2 Auditory and visually based input features **19**

2.1 Introduction 19

2.2 Acoustic parametrization 19

2.2.1 Mel-Frequency Cepstral Coefficients (MFCCs) 20

2.2.2 Linear Predictive Coding (LPC) . . . 21
2.2.3 Speech signal variability . . . 22
2.2.3.1 Intra-speaker variability . . . 23
2.2.3.2 Inter-speaker variability . . . 23
2.3 Visual parametrization . . . 23
2.3.1 Global techniques . . . 24
2.3.1.1 Discrete Wavelet Transform (DWT) . . . 24
2.3.1.2 Contourlet transform (CT) . . . 27
2.3.2 Local techniques . . . 29
2.3.2.1 Local Binary Patterns . . . 29
2.3.2.2 Local Directional Pattern (LDP) . . . 30
2.3.2.3 Local Optimal-Oriented Pattern (LOOP) . . . 31
2.3.2.4 Local Gabor Binary Pattern Histogram Sequence (LGPHS) . . . 31
2.3.2.5 Modified$_i$ Local Optimal-Oriented Pattern (M$_i$-LOOP) . . . 32
2.4 Conclusion . . . 37

3 Proposed Systems and Methodologies 38

3.1 Introduction . . . 38
3.2 Speech recognition system using MLP classifier . . . 39
3.2.1 Audio Dataset description . . . 39
3.2.1.1 Arabic dataset . . . 39
3.2.1.2 Continuous Berber dataset . . . 40
3.2.2 Evaluation setup . . . 40
3.3 Hand and sign gesture recognition system using SVM classifier . . . 44
3.3.1 Gesture Dataset description . . . 44
3.3.2 Evaluation setup . . . 44
3.4 Bimodal speech/gesture recognition system using MLP, SVM and CNN classifiers 49
3.4.1 Gesture and speech Dataset description . . . 52

3.4.1.1 Isolated Word Algerian Language dataset 53
3.4.1.2 Algerian Hand Gesture Dynamic Dataset 53
3.4.2 Evaluation setup 53
3.4.2.1 Convolutional Neural Networks 55
3.4.2.2 2D-CNN Implementation for Speech Identification 57
3.4.2.3 3D-CNN Implementation for Dynamic gesture recognition . . . 58
3.5 Conclusion 59

4 Simulation Results and Evaluation 61

4.1 Introduction 61
4.2 Performance Analysis of the first proposed system 61
4.2.1 Results for speaker recognition rate depending on isolated words using different training functions 62
4.2.2 Results for speaker recognition rate depending on continuous speech using different training function 63
4.2.3 Discussion 64
4.3 Performance Analysis of the second proposed system 65
4.3.1 Performance analysis on El Halawani ArSL dataset using SVM classifier . 65
4.3.2 Performance analysis on ArSL2018 ArSL dataset using SVM classifier. . . 65
4.3.3 Performance analysis on Jochen Triesch blue ASL dataset using SVM classifier. 65
4.3.4 Performance analysis on Dynamic dataset (Sebastien Marcel) using SVM classifier 68
4.3.5 Discussion and comparison with state-of-the-art methods 71
4.4 Performance Analysis of the third proposed system 72
4.4.1 Performance analysis on the Arabic audio clips using MLP and SVM classifier 74
4.4.2 Performance analysis on the 3D co-verbal gestures using MLP and SVM classifier. 74

4.4.3 Performance analysis of the bimodal speech/gesture system 76

4.4.4 Discussion 78

4.5 Conclusion 79

Conclusion **80**

Bibliography **96**

List of Figures

1.1 A taxonomy of gesture categories 8

2.1 MFCC feature extraction 20

2.2 Decomposition at one level by wavelet transform. 25

2.3 Wavelet decomposition at three resolution levels, LD and HD are respectively the approximation and detail coefficients at the $i-th$ level. 25

2.4 DWT decomposition, where cA, cH, cV and cD are the Approximation, Horizontal, Vertical and Diagonal image respectively. 26

2.5 Laplacian pyramid. (a) One level of decomposition. The outputs are a coarse approximation a(n) and a difference b(n) between the original signal and the prediction. (b) The new reconstruction scheme for the Laplacian pyramid 28

2.6 Directional filter bank. (a) Frequency partitioning where l = 3 and there are $2^3 = 8$ real wedge-shaped frequency bands. Subbands 0–3 correspond to the mostly horizontal directions, while subbands 4–7 correspond to the mostly vertical directions. (b) The multichannel view of an $l - level$ tree-structured directional filter bank . . 28

2.7 Hand posture decomposed using2 levels contourlet transform 29

2.8 LBP code calculation 29

2.9 LDP code calculation 30

2.10 LOOP code calculation 31

2.11 LGBPHS code calculation with $O = 8$ and $S = 5$ 32

2.12 Usual and proposed Masks used for Modified$_i$ -LOOP 34

2.13 Calculation of the edge responses IM_i and their association to the pixels of the 3×3 square neighborhood. $*$ is the convolution 35

3.1 Diagram of the proposed method. 43

3.2 Some examples from each gesture datasets. 45

3.3 Diagram of the proposed hand and sign recognition system using SVM. 47

3.4 Flowchart of the proposed recognition system 48

3.5 Diagram of the proposed method. 51

3.6 Few frames from hand gesture dataset. 52

3.7 Flowchart of the proposed recognition system 54

3.8 Single 2DCNN-based structure. 57

3.9 Hand gesture 3DCNN-based structure. 58

3.10 Flowchart of the proposed recognition system 60

4.1 Global speaker recognition for Arabic data set: (a) using traingda function, (b) using trainrp function, (c) using trainoss function, (d) using trainscg function. 62

4.2 Global speaker recognition for Berber data set: (a) using traingda function, (b) using trainrp function, (c) using trainoss function, (d) using trainscg function. 63

4.3 Global hand recognition for El Halawani ArSL dataset using SVM classifier 66

4.4 Global hand recognition for ArSL2018 ArSL dataset using SVM classifier. 66

4.5 Global hand recognition for American data set for white and black background using SVM classifier 67

4.6 Global hand recognition for Static Hand Gesture Asl Dataset using SVM classifier 68

4.7 Global hand recognition for Dynamic dataset (Sebastien Marcel) 69

4.8 Global accuracy for speech using MLP and SVM classifier 75

4.9 Global accuracy for gesture using MLP and SVM 75

4.10 Global accuracy for seepch/gesture using MLP classifier 76

4.11 Global accuracy for seepch/gesture using SVM classifier 77

List of Tables

3.1 Datasets description 45

4.1 Recognition rates obtained over the two datasets 64

4.2 Recognition rates obtained over the five datasets using SVM classifier 70

4.3 Processing time of the nine descriptors 72

4.4 Comparison with previous works 73

4.5 Recognition rates obtained fo the different methods 79

List of Abbreviations

ARVM Auto-Regressive Vector Model

ASL Arithmetic-Harmonic Sphericity

AI Artificial intelligence

ASL American Sign Language

ArSL Arabic Sign Language

AISLRSEW Automated Indian Sign Language Recognition System for Emergency Words

BiLSTM Bi-directional long short term memory network

CS-LBP Center-symmetric Local Binary Patterns

CSR Continuous speech recognition

CT Contourlet transform

DFB Directional Filter Bank

CNN Convolutional neural network

DTW Dynamic time warping

DWT Stationary wavelet transform

DWT Discrete wavelet transform

DSHMFCC Difference Mel Frequency Cepstral Coefficients

EnLDP enhanced Local Directional Pattern

FC-DNN Fully connected deep neural network

FBC Filter bank coefficients

GRU Gated Recurrent Unit

GSRS Gesture/speech recognition system

HMM Hidden Markov Model

HCI Human Computer Interaction

HOG Histogram of Oriented Gradients

ISA Indian sign language

ISR Isolated Speech Recognition

KNN K-Nearest Neighbors

LBP Local Binary Patterns

LDP Local Directional Pattern

LGPHS Local Gabor Binary Pattern Histogram Sequence

LDA Linear Discriminant Analysis

LPC Linear Predictive Coding

LP Laplacian Pyramid

LOOP Local Optimal-Oriented Pattern

LSTM Long short term memory network

M$_i$-LOOP Modified$_i$ Local Optimal-Oriented Pattern

MFCC Mel Frequency Cepstral Coefficients

MLP A multi-layer perceptron

NPC Neural Predictive Coding

ORB Oriented Fast and Rotated Brief

PCA Principal Component Analysis

RNN Recurrent neural networks

SARS Static Arabic recognition system

SVD Singular value decomposition

SVD Singular value decomposition

SHMFCC Difference Mel Frequency Cepstral Coefficients

SR Speech Recognition

sEMG Sparse multi-channel

ViT Vision Trans-former

SVM Support Vector Machine

VS-LBP Volumetric spatiograms of local binary patterns

Introduction

Hand gesture recognition is a rapidly evolving field within computer science and artificial intelligence, focused on interpreting and understanding human hand movements. It is considered one of the most complex and fascinating challenges in computer vision, due to the articulated structure of the hand and variations in environmental conditions. These systems aim to translate the intricate language of hand gestures into commands or information that can be understood by computers, enabling seamless human-computer interaction .

Driven by advancements in computer vision, machine learning, and sensor technology, hand gesture recognition systems are finding diverse applications, including human-computer interaction, robotics, gaming, sign language recognition, and more.

Co-verbal gestures, which are hand movements or body language that accompany speech, play a crucial role in enriching communication. Speech and co-verbal gestures work together as a dual system, forming a multi-modal communication approach. Gestures visually reinforce and add nuance to verbal expression, allowing individuals to convey emotions, ideas, and emphasis more effectively than through words alone. In addition, a speaker might use hand gestures to illustrate a concept, highlight a point, or direct the listener's attention.

The importance of this multi-modal approach becomes particularly apparent when considering individuals with hearing impairments, who rely heavily on visual cues for communication. Sign languages, which are formalized systems of gestures, demonstrate the power of gestures as a primary communication tool when spoken language is absent. Furthermore, co-verbal gestures can bridge communication gaps for everyone, providing an extra layer of clarity and understanding.

This research delves into gesture recognition with a specific focus on the relationship between gestures and speech. By examining how co-verbal gestures enhance spoken language, this work aims to understand their role in improving communication, particularly for individuals who rely on non-verbal cues. This involves understanding the harmonized communication between speech

and gestures, where their timing and meaning are synchronized. It also necessitates analyzing the cognitive processes behind gesture production, as they are often spontaneous and linked to speech formulation.

The ultimate goal of this research is to contribute to the development of advanced gesture recognition systems that facilitate better communication, especially in diverse linguistic and multi-modal settings, such as Arabic language communication. These systems will not only recognize gestures but also analyze the interplay between gestures and speech to provide a more comprehensive understanding of human communication. By combining speech and gesture recognition technologies, we can bridge the gap between hearing and non-hearing individuals, fostering more inclusive communication spaces.

Furthermore, the study of the bimodality of speech and gestures is essential to better understanding language development and communication for deaf children. Research in this area can significantly enhance educational and therapeutic practices, promoting access to language and improving the social integration of these children.

Context of the Study

This thesis was carried out at the L.C.P.T.S Laboratory, Faculty of Electrical Engineering, within the Telecommunications Department. The research primarily focuses on the field of audio-gestural communication, emphasizing the critical role of gestures in non-verbal communication, particularly for individuals with hearing impairments. Gestures serve as essential tools for enhancing understanding, providing clarity, and bridging communication gaps when verbal communication is limited or absent. In the case of individuals with hearing difficulties, gestures become a primary mode of interaction, allowing them to express themselves and comprehend others more effectively.

This research explores the intricate relationship between speech and co-verbal gestures, aiming to highlight how these two modalities work in harmony to facilitate communication. Specifically, this study investigates how gestures can enhance spoken language and how they can contribute to improving communication accessibility for the hearing-impaired.

Applied to the Arabic language, this research also introduces a specialized database that was recorded as part of the project. This database, containing both static and dynamic gestures, serves as the foundation for the study's experiments, which aim to improve the recognition of gestures in

an Arabic-language context. The ultimate objective is to contribute to the development of a system that recognizes and interprets gestures, providing greater access to communication for diverse populations, including the hearing-impaired community. The data-set includes Arabic audio clips and 3D co-verbal gestures, specifically focused on the Arabic language sequence. Both datasets were recorded at the Speech Communication and Signal Processing Laboratory at USTHB in Algiers. Data collection involved multiple visits to the school for the deaf in Rouiba to ensure the gestures' authenticity and alignment with the Algerian sign language community's communication needs.

Problematic

The central issue addressed in this thesis involves establishing a comprehensive system for recognizing and interpreting both sound and gestures, focusing on the complex interplay between speech and co-verbal gestures. The research identifies several key challenges within this framework, including the need to understand the precise role of gestures in communication and the development of a robust methodology for their accurate interpretation. To tackle these challenges, the research aims to explore the interaction between verbal and non-verbal elements of communication, which is crucial for advancing sound recognition systems. Additionally, the study extends its focus to the investigation of various modalities, with the goal of creating a system that integrates both acoustic signals and hand-generated gestures. This comprehensive approach necessitates the recording of a specialized database, which will serve as the basis for detailed analysis and experimentation.

Objectives and Contributions

This thesis presents a comprehensive study aimed at addressing a critical need through the integration of innovative techniques in audio data processing and gesture recognition. The primary objective of this study is to design and implement a system capable of recognizing gestures by leveraging both acoustic and gesture cues. To achieve this objective, several key tasks are under-taken, including bibliographic research, creation of a specialized database, extraction of acoustic and visual descriptors related to gestures, system implementation, comparative analysis with existing systems, and discussion of the results. More precisely, we aim to fulfill the following constraints:

1. Creation of a Specialized Arabic Audio-Gesture Database: Develop a comprehensive database

tailored for the hearing-impaired community, which includes both static and dynamic gestures. This resource aims to enhance communication accessibility and provide a valuable tool for research and practical applications in gesture recognition.

2. Development of an Integrated Gesture Recognition System: Create a system that merges acoustic and gesture modalities for recognizing gestures. This system is designed to improve accuracy and robustness in real-world scenarios by leveraging both sound and gesture data.

3. Proposal of a Modified Local Optimal Oriented Pattern Descriptor: Introduce a new textual descriptor, Modified LOOP, to enhance the representation and analysis of gestures. This proposal aims to improve the accuracy and efficiency of gesture recognition systems.

4. Study of Bimodal Speech-Gesture Communication: Investigate the bimodal nature of speech and gestures, with a particular focus on its impact on children with hearing impairments. This study aims to advance our understanding of how combined speech and gestures can benefit the language development and social integration of deaf children.

Overview of the thesis

In this section, we provide an overview of the remaining contents of this thesis, which is organized into four main chapters. Chapter 1: "Literature Survey of gesture and speech recognition system". This chapter provides an up-to-date review of existing research in gesture, speech, and gesture/speech recognition systems. It covers the types of gestures, speech representations, recognition techniques, and their applications, setting the stage for the thesis by presenting the current state of the field.

Chapter 2: "Hand Gestures and Speech characterization". This chapter explores the detailed characterization of hand gestures and speech used in the proposed methodologies, focusing on both auditory and visual modalities. It examines how these features are integrated into the systems to enhance the recognition and interpretation of gestures and speech.

Chapter 3: "Proposed Systems and Methodologies". This chapter meticulously outlines the construction process of the proposed systems. It details the experimental procedures, data collection methods, and parameter configurations. Initially, it covers the development of the speech recognition system. Following this, the chapter introduces the second proposed system, which focuses on

Hand and Sign Gesture Recognition. The chapter culminates with an advanced discussion of the integrated Speech/Gesture Recognition system, highlighting the technological innovations achieved.

Chapter 4: "Simulation Results and Evaluation". The final chapter presents and discusses the results from the simulations and experiments conducted on the proposed systems. It evaluates the performance of each system, providing insights into their effectiveness, strengths, limitations, and areas for improvement. Finally, we conclude our thesis and we present our future work.

Chapter 1

Literature Survey of gesture and speech recognition system

1.1 Introduction

Nonverbal communication is essential for natural and effective face-to-face human-human interaction. It is the process of communicating through sending and receiving wordless (mostly visual, but also auditory) signals between people. Consequently, a natural and effective face-to-face Human Computer Interaction (HCI) requires machines (e.g., robots) to understand and produce such human-like signals. There are many types of nonverbal signals used in this form of communication including, body postures, hand gestures, facial expressions, eye movements and touches. An innovative approach to enhance HCI is the recognition of hand gestures for speech comprehension. By integrating hand gestures with speech recognition systems, it becomes possible to create more intuitive and efficient interfaces for communication. This combined approach bridges the gap between human communication patterns and machine understanding, resulting in a more natural and expressive way of interacting with speech technology. In this chapter, we present an up-to-date review of the state-of-the-art of gesture, speech and gesture/speech recognition system which includes gesture and speech representations, recognition techniques and applications.

1.2 Gesture Recognition System

Gestures are a continuous movement that can be recognized by the person comfortably but challenging to recognize with the machine. The human gesture is described as the bodily motion of hands, fingers, arms, and other body parts that allows humans to communicate meaning and information during interactions with one another.

In HCI, hands are the most intuitive and useful part compared to other human body parts. According to the temporal evolution of the gesture shape, two types of gestures can be distinguished, namely static gestures and dynamic gestures. The information carried in static gesture is expressed by a single posture contrast, whereas in dynamic gesture, It is expressed by two characteristics, configuration of the hand and its position, which corresponds to a movement of the hand in time. All the static gesture uses a single frame with limited information, and it has a little computational cost. On the other hand, dynamic gesture contains more meaningful information with greater complexity, and it is in the form of video and more suitable for real-time application.

A gesture can be also categorized with respect to different criteria [4]. For instance, distinguishes five types of gestures:

1. Emblems: are gestures that can be directly translated into brief verbal communications, like using a goodbye wave to replace words. These gestures are highly culture-specific, meaning their meanings and interpretations can vary significantly across different cultures

2. Illustrators: are gestures that depicts what the communicator is saying verbally (e.g. emphasis a key-point in the speech. These gestures are inherent to the communicator's thoughts and speech. They can be classified into five subcategories as proposed by McNeill [5]

    - Beats: refer to rhythmic and frequently repetitive flicks, typically characterized by short and quick movements of the hand or fingers.
    - Deictic gestures: pointing gestures which can be concrete (pointing to a real location, object or person) or abstract (pointing abstract location, period of time).
    - Iconic gestures: it consists of hand movements that depict a figural representation or an action (hand moving upward with wiggling fingers to depict tree climbing).
    - Metaphoric gestures: gestures depicting abstractions.

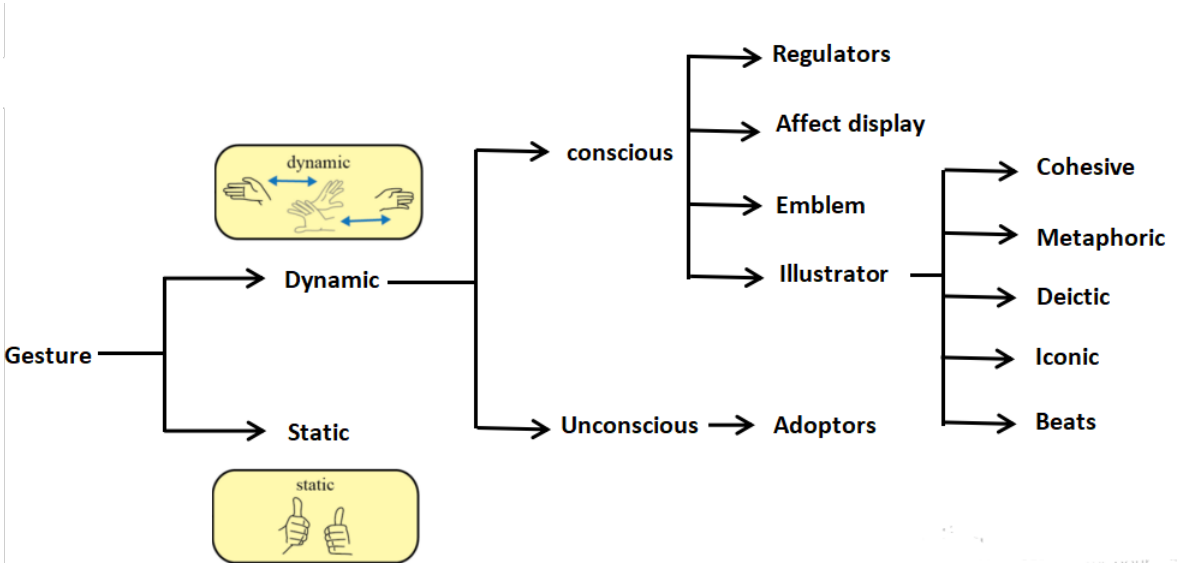


Figure 1.1: A taxonomy of gesture categories

- Cohesive gestures: they are thematically related but temporally separated gestures due generally to an interruption of the current communicator by another one [7].

3. Affect displays: an affect display is a gesture that conveys emotion or communicator's intentions (e.g. if the communicator is embarrassed). This type of gesture is less dependent on the culture.

4. Regulators: a regulator is a gesture that controls interaction (e.g. control turn-taking in conversation) [5].

5. Adaptors: an adaptor is a gesture that enables the release of body tension (e.g. head shaking, quickly moving one's leg). These gestures are not used intentionally during a communication or interaction: they were at one point used for personal convenience and have turned into a habit though [5].

In addition, a gesture can be also conscious (intentional) or non-conscious (reflex, adaptors).Figure 1.1 illustrates a taxonomy of gesture categories which resumes all the criteria except the last one.

Over the past three decades, significant advancements have been made in the field of hand gesture recognition, with a particular focus on recognizing Alphabet Sign used in various sign languages, including American, British, Australian, Japanese, Indian, Arabic, Thai, and more. The primary objective of these research efforts aimed to aid the deaf and hard-of-hearing communities in their communication needs. Sign languages encompass a rich variety of gestures, ranging from

descriptive and non-descriptive signs to alphabet or finger-spelling signs. By developing technologies capable of recognizing and interpreting these gestures, researchers aim to provide a means for individuals with hearing impairments to effectively communicate with others, bridging the communication gap and enhancing their overall quality of life. The advancements in this area have the potential to create more inclusive and accessible societies for the deaf and hard-of-hearing population worldwide.

As a typical problem in pattern recognition, hand gesture recognition has to deal with several of constraints and difficulties. The most prominent challenge in the practical process of hand gesture recognition is how to characterize each hand gesture with features that best describe it. Various architectures and several algorithms have been developed in this area based on local features [6, 7, 8, 9, 10, 11, 12, 13], global features [6, 14, 15, 16, 17, 18] or hybrid features [9, 19, 20, 21, 22, 23, 24] methods. These algorithms can be broadly classified into several categories: shape [25, 26], orientation [8, 27, 28], texture [7, 8, 9, 11, 20, 29, 30], contour [31, 32, 33], motion [34, 35, 36], distance [37, 38], center of gravity [8, 39], to name a few.

Authors in [12] propose a simplified method using Local Binary Patterns (LBP) feature extraction and Linear Discriminant Analysis (LDA) for a hand sign gesture recognition system. To evaluate the effectiveness of their system, the researchers conducted experiments using Chinese and Bangladeshi numeral gesture datasets. These datasets likely contained images or video sequences of hand gestures representing numerical values in Chinese and Bangladeshi sign languages. The effectiveness of their system is 92.4%. Kim et al [40] devised an adaptive LBP approach to effectively detect and track hands in three-dimensional environments. The adaptive nature of their method likely allows it to adjust to varying lighting conditions, hand poses, and complex backgrounds, enhancing its robustness and accuracy.The advantages of their proposed system lie in its simplicity and effectiveness.

A simple yet effective hand detection and tracking system is highly desirable, especially in real-time applications where computational efficiency is crucial. By combining the power of LBP with adaptivity, the authors' approach likely achieves a balance between accuracy and computational speed, making it suitable for real-world applications.Maqueda et al [41] focused on HCI based on visual hand gesture recognition using volumetric spatiograms of local binary patterns Volumetric spatiograms of local binary patterns (VS-LBP). They proposed a new discriminative video descriptor for the recognition stage, presumably intended to improve the system's capability and enhance overall recognition accuracy. The results they obtained were satisfactory, demonstrating the

system's effectiveness in accurately recognizing and interpreting a variety of hand gestures. This suggests its potential suitability for various applications in HCI.

Authors in [11] present an interesting method for a hand gesture recognition system. The authors' method centers around addressing hand pose estimation challenges, crucial for accurate hand gesture recognition, on resource-constrained mobile devices. To achieve this, they combined two popular feature descriptors: Histogram of Oriented Gradients (HOG) and LBP . The reported results indicated a good detection time, suggesting that their method is computationally efficient and well-suited for mobile devices with limited processing capabilities. Additionally, the method achieved an acceptable detection rate, implying that it performs reasonably well in recognizing hand gestures accurately. Houssem et al. [42] developed a hand posture recognition system designed for Android devices, employing the LBP approach. They performed experiments to assess the system's performance under diverse conditions, including variations in lighting, backgrounds, and hand orientations. Impressively, the system achieved an overall accuracy of 88%, which is considered promising for practical applications. Nevertheless, it's crucial to acknowledge that the system's effectiveness may differ based on the complexity and diversity of the hand gestures utilized in the evaluation.

In [15], Priyanka et al. developed a hand gesture recognition system using machine learning that was tested using Sebastian Marcel hand gesture database. The overall classification accuracy of their system is 96.5% with a recognition time of 0.024s using a Support Vector Machine, (SVM) classifier, which is remarkably fast. This real-time performance is valuable for interactive applications that require quick and responsive gesture recognition. Sheena et al.[43] propose a static gesture classification and recognition using Histogram of Oriented Gradients (HOG) feature descriptors ; their experiment was carried out on a database of ten different static gestures.The effectiveness of their system was about 98% using Support Vector Machine (SVM) classifier.This level of effectiveness indicates that the system holds significant promise for a range of applications, including human-computer interaction, sign language recognition, and gesture-based control systems. In [29], a real-time hand gesture recognition system is presented. The system is implemented with a HOG cascade detection step and a HOG-SVM detection step, Satisfactory results were obtained. Authors in [30] developed a static American Sign Language (ASL) using canny-edge and HOG. The experiments were carried out on two datasets, the Massey and Kaggle. The overall accuracy of the system is 97.6% with a recognition time of 0.39s for the first and 99.0% with a recognition time of 0.43s for the second.

On the other hand, deep learning approaches have been also explored in Hand Gesture Recog-

nition System (HGRS) [44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55]. For instance, Sakshi et al. [44] propose a hand gesture recognition system using deep learning for the interpretation of sign language. Two datasets, Indian sign language ISLand American Sign Language (ASL), were used to evaluate the performance. The proposed model obtained accuracies of 99.96% and 100% for the ISL and ASL datasets, respectively. In [54], a sign static hand gesture recognition system based on a convolutional neural network (CNN) with a feature extraction method using an Oriented Fast and Rotated Brief (ORB)descriptor and Gabor filter is introduced. Different datasets were used, obtaining remarkable results. Authors in [52] proposed an hybrid single stage framework for HGRS based on the deep learning model. For this CNN based architecture was used for the classification of a custom dataset, which consists of alphabets, digits, and words. In [53], authors implement an algorithm for HGRS based on neural network. The complete recognition system consists of a hand locating network based on wavelet packet transform for the feature extraction. This method achieves good recognition accuracy. Damaneh et al. [54] introduced a novel deep learning neural network architecture for identifying static hand gestures in sign language, the proposed structure comprises a combination of CNN and classical non-intelligent feature extraction methods.To evaluate the effectiveness of the proposed structure, experiments were conducted on three different databases: Massey, ASL Alphabet, and ASL. These databases consist of a substantial number of hand gesture images. The results indicate impressive mean accuracy for each database's test set: 99.92% for Massey, 99.8% for American Sign Language (ASL), and 99.80% for ASL Alphabet. In [55], authors introduce an Automated Indian Sign Language Recognition System for Emergency Words (AISLRSEW). The system is designed to recognize ISL words that are frequently used in emergency situations. To address the challenge of identifying ISL words with identical hand orientation and multiple viewing angles, the proposed AISLRSEW utilizes a combination of CNN and local handcrafted features.The results of the evaluation demonstrate that the proposed system achieves an average accuracy of 94.42%.

In [56], the authors introduced a static hand gesture recognition system based on the Vision Trans-former (VIT) using a self-made dataset and two publicly available American Sign Language da-tasets. Hand gesture images are captured, and the background is filtered using depth information from the Microsoft Kinect camera. The eight-connected discrimination algorithm and the distance transformation algorithm are employed to eliminate unnecessary arm information. The validation accuracy of the proposed model on the three datasets achieves 99.44%, 99.37%, and 96.53%, respectively. Chun et al [57] proposed a hand gesture recognition using the Stacked Distilled Vision

Transform-ers (SDViT) model. The approach starts with a pretrained VIT that uses a self-attention mechanism to capture complex relationships among image patches, thereby addressing the challenge of similar hand gestures To improve model generalization and compress the VIT, knowledge distillation is applied. The stacking of multiple distilled VITs results in enhanced predictive performance and reduced overfitting. The model achieved classification performance ranging from 91% to 98%. Zabihi et al. [58] introduced a hybrid transformer-based architecture for hand gesture recognition system using sparse multi-channel sEMG. The architecture consists of two parallel paths followed by a linear layer that acts as a fusion center, integrating the advantages of each module. The pro-posed architecture was evaluated using the widely-used second Ninapro dataset (DB2). This meth-od effectively classifies a substantial number of gestures (49) with high accuracy.

In order to improve the discrimination capability, robustness and simplicity of local descriptors, a vast number of other variants have been proposed including center-symmetric LBP (CS-LBP) [59], VS-LBP, LDP [60, 61], multi-block local binary pattern (MBLBP) [62], NCDB-LBP [63], enhanced LDP (EnLDP) [64] and LOOP [65].

A considerable amount of research has been devoted to ASL in the literature, research began in 1996 with the work of Starner and Pentland [66]. Many feature extraction methods have been implemented such as scale invariant transform (SIFT), histogram of oriented gradient HOG, wavelet moments and Gabor filters [67]. In contrast to ASL, ArSL has received minimal attention and only a few research studies have been published. As described in [68], there is no standardized language coordination for ArSL, which makes learning as well as translating a difficult task for deaf Arab communities.

Over the past 20 years, many studies have been developed for ArSL to recognize the Arabic alphabet. in 2001 Al-Jarrah and Halawani [69] developed a neuro-fuzzy system using static Arabic manual alphabets, the authors used 60 examples for each sign with a total of 1800 samples for 30 Arabic alphabet characters, the system achieved an accuracy of 93.55%. In [70], Al-Rousan and Hussain built an adaptive neuro-fuzzy interference system for letter recognition. A colored glove was used to ease the process of segmenting the regions of the hand. The recognition accuracy achieved was 95.5%. Assaleh and Al-Rousan [71] used polynomial classifiers to perform alphabet recognition. Second-order polynomial classifier per class were created for the training stage, and the feature vectors were generated from the dataset. The error rates are 1.6% for the training set and 6.59% for the testing data.

Al-Jarrah, Shatnawi, and Halawani [72] present a feedforward neural network and a probabilistic neural network for the recognition of sign language alphabets. Feature vectors such as the orientation and position of the hand within the image are fed into the neural networks. An accuracy of 94.4 % was obtained with ANN, whereas 91.3 % was achieved with probabilistic neural network. Authors in [71] present various spatiotemporal feature extraction techniques with applications in online and offline recognition of isolated ArSL gestures.Spatial-domain feature extraction was applied, where K nearest neighbor (KNN), Bayesian, likelihood ratio and classifiers were used. Classification performance ranging from 97 to 100% was obtained.

1.3 Speech Recognition System

Speech is the verbal expression of language through the articulation and production of sounds, words, and sentences using the vocal organs. It is a fundamental form of communication, allowing individuals to convey thoughts, ideas, emotions, and information to others, enabling social interactions, exchange of knowledge, and expression of individuality [73].

The goal of Speech Recognition (SR)is to develop techniques and systems that allow computers to be able to "hear","understand" and "act upon" spoken information. The problem of SR has been actively studied since the 1950s, and it is natural to wonder why one should continue researching this area. SA technology aims to convert spoken language into written text or commands, allowing for seamless human-computer interaction. It has garnered significant attention from scientists and researchers across various disciplines due to its immense potential and practical applications, such as voice assistants, transcription services, voice-controlled devices, language translation, and accessibility tools for individuals with speech impairments.

SR systems can be classified in various ways, and an important classification is based on whether the system is speaker-independent or speaker-dependent. Speaker-dependent systems recognize a particular person's utterances only if that person has previously stored examples of his or her speech in the system.On the other hand, speaker independent systems recognize speech without prior experience with a particular person.

Additionally, there are two primary types of SR : Isolated speech recognition (ISR) and Continuous speech recognition (CSR). Isolated speech recognition involves distinct and separate utterances of individual words with clear pauses between them, treating each word as an independent entity

for processing and analysis. On the other hand, Continuous speech recognition involves the seamless flow of spoken language by a speaker, where sounds and words are produced without explicit breaks or pauses. Unlike isolated-word speech, continuous speech includes connected phrases and sentences, mirroring natural everyday communication.

SR systems encompasses various approaches and algorithms for deciphering spoken language. The acoustic-phonetic approach employs algorithms like zero crossing, energy measurement, and feature extraction to analyze the acoustic properties of speech sounds and their corresponding phonetic representations. These methods aid in distinguishing different phonemes and speech sounds within the signal.

On the other hand, the pattern recognition approach utilizes algorithms like template matching. This involves comparing the input speech signal with predefined templates representing known speech units, such as phonemes, words, or sentences. By finding the closest match, the system recognizes the spoken content.

The artificial intelligence (Artificial intelligence (AI)) approach to SA relies on algorithms that rely on knowledge sources, statistical methods for dealing with the stochastic nature of speech signals, and neural networks. These techniques enable the system to decode and understand spoken language by leveraging linguistic knowledge, probabilistic models, and advanced neural network architectures. This AI-based approach allows for more sophisticated language understanding and context analysis.

In practice, speech recognition systems often combine multiple approaches, combining the strengths of each method to achieve more accurate and robust results. These diverse approaches play a crucial role in advancing the field of SA, making speech recognition a powerful tool with numerous practical applications, including voice assistants, transcription services, voice-controlled devices, and more. Authors in [74] have proposed a novel scoring method based on distance calculation for similarity measurement. The study probably includes empirical findings derived from evaluations carried out on two publicly available collections of speaker recognition data. these experiments likely demonstrate that the new distance-based scoring method outperforms the existing state-of-the-art approaches. Barhoush et al [75] developed a novel approach for end-to-end speaker identification and localization using a simple fully connected deep neural network (FC-DNN), coupled with input data from two microphones.The model uses the new proposed Mel Frequency Cepstral Coefficients (MFCC) based feature called Shuffled MFCC (SHMFCC) and its variant Difference Shuffled MFCC (DSHMFCC) features. This innovative approach consistently demonstrates

superior accuracy compared to both baseline techniques and conventional methods across a spectrum of scenarios and conditions.

Chetouani et al [76] introduced a new characterization algorithm based on nonlinear prediction, which is an extension of the classical LPC parameters for speaker identification. The parameters performances are estimated by two different methods: the Arithmetic-Harmonic Sphericity (AHS) and the Auto-Regressive Vector Model (ARVM). Two different methods are proposed for the parameterization based on the Neural Predictive Coding (NPC) classical neural networks initialization and linear initialization, the effectiveness of their system was about 100%. Zhao et al [77] proposed a feature fusion method for speaker recognition based on embedding mechanism. The fusion features adopted in there approach are filter bank coefficients Fbank and MFCC. long short term memory network (LSTM) and bi-directional long short term memory network (BiLSTM), are used as testbed system . Results show that, by using the proposed feature fusion approach, the performance of both models are improved. Sreehari and Mary [78] proposed a system based on LP residual for text independent speaker recognition. Discrete wavelet transform (DWT) and stationary wavelet transform SWT are experimented to parameterize the linear prediction residual. Effectiveness of the system is evaluated by using 10 s–10 s task of NIST speaker recognition evaluation 2010 database. This method achieves good recognition accuracy.

One the other hand, Mandar and Dharmraj [79] introduced a real-time speech recognition system using Convolutional neural network (CNN) networks for the English language. The system focuses on MFCC techniques with cepstral-based features. Various techniques like CNN and dynamic time warping (DTW) are employed for feature vector matching, a recognition of 100% is achieved. Hong et al [80] proposed a hybrid acoustic model of deep convolutional neural network, long short-term memory, and deep neural network for Chinese speech recognition, acceptable results is obtained. Authors in [81] implemented a speaker recognition system using CNN and LSTM Techniques. The experiments are carried out on two different datasets, good recognition rate is achieved. In [82] authors presented a speech recognition system using an enhancement of CNN, the effectiveness of their system was about 96.17%. Qasim and Abdulbaqi [83] presented an Arabic speech recognition system using CNN. The proposed system is evaluated on dataset content 12000 samples Arabic spoken digits and commands from various dialects. The proposed system has achieved a recognition accuracy of 100%. In [84], a CNN model is used to improve the recognition accuracy of 17 mathematical words which are the most commonly used in mathematical expression. The recognition accuracy of 83.33% is achieved.

Azim et all [85] develop a new CNN based spoken digits recognition system for the Arabic digits, 88 different speakers with different genders and of distinct age-groups are used. The system achieved an accuracy of 99%.In [86] a novel method for improving speech emotion recognition using ViT model is proposed. From Mel spectrogram input fed into the model the authors leverage the ViT model's capabilities to capture spatial dependencies and high-level features in images which are adequate indicators of emotional states. Two benchmark speech emotion datasets are used. The results of the proposed model on the Two datasets achieved 98%, 91%, and 93% accuracy, respectively. Authors in [87] designed a visual speech transformer architecture for audio-visual speech enhancement, which was tested on the Grid and AVSpeech datasets. The results show that the new model outperforms several stat-of-the-art models by a large margin.

1.4 Bimodal Gesture/Speech Recognition System

Although the recognition of speech and gesture systems have been studied extensively, currant attempts of combining them constitute the most important modalities in human-computer interaction. The conducted literature survey of multi-modal fusion provides an insight into the different methods and techniques that can be adopted and implemented to facilitate user interaction. Traditionally, gesture/speech recognition systems (GSRS) consist of two processing stages: feature extraction from audio and visual information followed by modality fusion and recognition [88, 89].For traditional methods, features are usually extracted from visual gestures and audio wave-forms and then concatenated [90, 91]. Following this fusion step, the resulting features are fed to a classifier such as SVM, hidden markov model (HMM), coupled HMM, etc. In recent years, due to advancements in deep learning technology, numerous deep learning methods have been introduced, replacing the traditional feature extraction step with deep bottleneck architectures. The first CNN image classifier designed to distinguish visemes was trained in reference [92]. Deep bottleneck features have been employed for word recognition, allowing for the full utilization of deep convolutional layers and the exploration of highly abstract features.

In [93] authors proposed a hybrid network called "DMS-SK/BLSTM-CTC" to improve the performance human–robot interaction. They use 10 gestures in the ASL data-set and 10 speech commands to build the gesture/speech hybrid data-set. The experimental results show that the recognition accuracy of the gesture-speech hybrid network improves by 2% and 12%, respectively, compared to the gesture or pure speech networks. Haodong et all [94] develop a real-time, multi-

model human–robot collaboration system using speech and gestures, An improved open-source speech with 16 dynamic gestures data-set are used. The experimental results validate the feasibility and effectiveness of the proposed CNN algorithms. Authors in [95], developed a multi-modal speech/gesture recognition network using 3D CNN to extract visual features and Gated Recurrent Unit (GRU) to extract speech features, good results were obtained. Authors in [96] proposed a hybrid Multidimensional CNN for different type of multi-modal recognition (speech, gesture, etc). Several variants of the fusion operation have been developed, the effectiveness of their system was about 86%. Salah et all [97] developed a co-speech gesture generation system by DSI team for the GENEA challenge 2022.This system sounds intriguing with its unique hybrid encoder-decoder architecture based on transformer networks and recurrent neural networks (RNN). Chen et al [98] developed a real-time, multi-modal HRC (Human-Robot Collaboration) system that use speech and gestures. They implemented a CNN to recognize dynamic gestures in real-time using motion history images and deep learning methods. Real-time speech recognition from the human worker is achieved through the use of an improved open-source speech recognizer. The paper also introduces an integration strategy to combine the results of gesture and speech recognition. Furthermore, a software interface has been designed for system visualization. The experimental results validate the feasibility and effectiveness of the proposed algorithms and the HRC system. Qiu et al [99] introduced a deep learning-based multi-modal fusion framework. This architecture incorporates two modalities: speech commands and hand gestures. Initially, user inputs for speech and gesture commands are subjected to recognition using CNN for speech command identification and LSTM for hand gesture recognition, respectively. The experiments demonstrate the superiority of the proposed multi-modal fusion model over the single-modal fusion model.

1.5 Conclusion

In this chapter, we have offered a concise summary of diverse methodologies aimed at tackling the challenges of hand gesture recognition, speech recognition, and the integration of speech and gestures. Our review encompasses a substantial body of research conducted over the past two decades, highlighting key advancements, prevailing trends, and emerging challenges in the field. While significant progress has been made through innovations in machine learning, sensor technology, and multimodal fusion, challenges such as robustness, adaptability, and real-time processing still require further exploration. These insights set the foundation for future research aimed at enhancing natural

and seamless human-computer interaction.

Chapter 2

Auditory and visually based input features

2.1 Introduction

Feature extraction plays a critical role in numerous processes related to computer vision, object detection and localization, image processing, image retrieval, speech recognition, data mining, pattern recognition, machine learning, and bio-informatics. It is used to extract the most distinct features present in a data-set (image, text, voice), which serve to represent and describe the underlying information. Various methods are used to extract features, often based on geometric, statistical, texture, color, and other characteristics. There is no universally optimal technique for describing input parameters; consequently, a range of global and local techniques have been developed to address these challenges effectively.

In this chapter, we present a comprehensive analysis of the auditory and visual input features utilized in our systems. Detailed information is provided for each descriptor employed.

2.2 Acoustic parametrization

The speech signal is a highly complex acoustic vector carrying a wealth of linguistic, para-linguistic, and emotional information. Understanding and unraveling this complexity is crucial for developing advanced speech processing and recognition systems that can accurately capture the nuances of human communication. For this purpose, we will focus on the following features known as

Mel frequency cepstral Coefficient MFCC [100], first and second derivatives ΔMFCC, $\Delta\Delta$MFCC as well as linear predictive coding LPC [101] from the Arabic/Berber audio clips.These three acous-

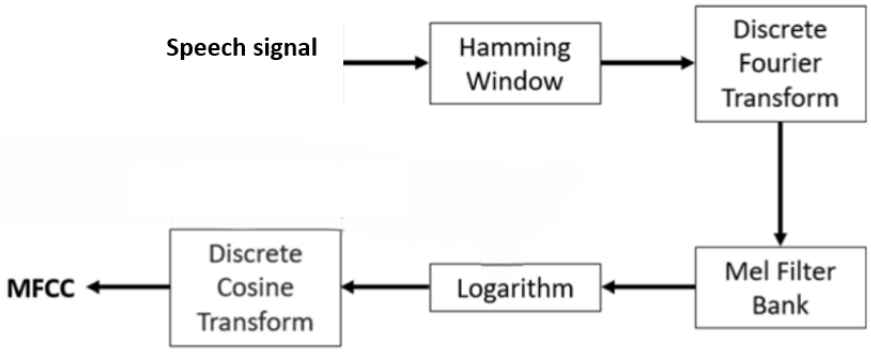


Figure 2.1: MFCC feature extraction

tical analysis methods perform modeling of perception and hearing mechanisms,which is expected to provide more robustness to the speech recognition system.

2.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are frequency-domain features that have demonstrated their effectiveness in speech recognition [100, 102, 103] Indeed, MFCC is one of the most commonly used frame-based features in speech processing. It assumes that the speech signal can be considered wide-sense stationary over short frames with typical time lengths ranging from 10 to 25 milliseconds.The extraction process of MFCC follows several steps, which are illustrated in Figure 2.1. These steps are designed to approximate the human auditory system's response and capture relevant spectral information on the Mel scale. Considering its linear behavior at low frequencies and logarithmic behavior at frequencies above 1 kHz, The Mel scale is defined by the following equation [104]:

$$Mel(f) = xlog(1 + L); L = \frac{f}{y} \tag{2.1}$$

Several values are used for x and y, the most commonly used values are $x = 2595$ and $y = 700$.

The analysis using MFCC coefficients is based on cepstral analysis, which involves taking the logarithm of the Mel spectrogram and then applying a Discrete Cosine Transform (DCT) to transform the Mel spectrum into the cepstral domain, generating the features known as MFCCs [104].

The MFCC coefficients are calculated through a series of mathematical transformations, with each coefficient representing a specific characteristic of the vocal signal's power spectrum on a perceptual frequency scale (the Mel scale). Here is the detailed formula for calculating an MFCC

coefficient:

$$C_n = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_m \cos\left(\frac{n\pi\left(m - \frac{1}{2}\right)}{M}\right) \tag{2.2}$$

C_n is the n-th MFCC coefficient, M is the total number of Mel filters, X_m is the log-Mel energy obtained for the m-th frequency band after applying the Mel filter bank, n is the index of the MFCC coefficient for $n = 1, 2, \ldots, N$ (where N is often set to 12 or 13).

To gain a better understanding of the dynamics of the speech signal, the first derivative (delta), which represents the variation of the MFCC coefficients over time, is also considered, as well as the second derivative (delta-delta), which describes their acceleration. The delta coefficients are calculated using the following formula.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \tag{2.3}$$

The coefficients denoted as delta $(d(t))$ for a frame (t) are computed based on the static coefficients $c(t-1)$ à $c(t+1)$. Similarly, the coefficients denoted as deltadelta are calculated from the delta coefficients.

2.2.2 Linear Predictive Coding (LPC)

Linear predictive coding(LPC) is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. It was first proposed as a method for encoding human speech by the United States Department of Defence in federal standard 1016, published in 1984 [101]. It is a powerful technique used for feature extraction purpose of speech signals since it characterizes the vocal tract well [101].

The LPC analysisa, also known as LPC-10, estimates the vocal tract resonance from a signal's waveform, removing their effects from the speech signal in order to get the source signal. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples at a particular time, k, the speech sample $s(k)$ is represented as a linear sum of the n previous samples. This can be represented by the equation:

$$s(k) = \sum_{i=1}^{n} a_i s(k-i) \tag{2.4}$$

Where $S(k)$ is the value of the signal at time k. The coefficients a_i are called the Linear Predictive Coding Coefficients. The coefficients can be analyzed to provide insight to the nature of the signal. Another important feature of LPC is that it minimizes the sum of the squared differences between the original speech and estimated speech signal over a finite duration. It produces a unique set of predictor coefficients which are normally estimated with every frame which is of usually 20 ms to 30 ms long [101].

The input signal is sampled at a rate of 8000 samples per second. This input signal is then broken up into segments or blocks which are each analysed and transmitted to the receiver. The 8000 samples in each second of the speech signal are divided into segments of m samples, typically containing between 160 and 240 samples per segment. The predictor coefficients are represented by a_k . Another important parameter is the gain (G). The transfer function of the time-varying digital filter is given by equation 2.5

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.5}$$

Summation is computed starting at $k = 1$ up to p. Here LPC 10 is used. This means that only the first 10 coefficient are transmitted to the LPC synthesizer. Several methods, such as the autocorrelation method and the covariance method, have been developed to determine the parameters that minimize the sum of the squared error. In this case, the autocorrelation method is used to compute the coefficients [101].

The sequence of LPC coefficient vectors generated from all the frames represents the speech signal in a compressed form while retaining essential spectral information. These coefficients can be efficiently quantized and transmitted or stored, making LPC a popular technique for low-bit-rate speech encoding, speech synthesis, and other speech-related applications.

2.2.3 Speech signal variability

The speech signal is a phenomenon rich in complexity and variation, encompassing a variety of articulatory and acoustic features specific to spoken language. Its nature is influenced by a multitude of factors. Two types of variability can be distinguished based on their origins and characteristics:

inter-speaker and intra-speaker variability.

2.2.3.1 Intra-speaker variability

Intra-speaker variability reflects the natural diversity in the speech of a single person and can result from multiple physiological, psychological, linguistic, and contextual factors. This variability arises from various sources. Among these are variations in the phonetic realization of phonemes, influenced by phonetic or prosodic contexts, the level of formality in speech, the nature of the social relationship with listeners, and even specific listener needs, especially in noisy environments. These factors are often grouped under the term "style effects." Another type of factor affecting how a speaker expresses themselves is related to their cognitive state, including elements such as emotions, attention, and fatigue [105].

2.2.3.2 Inter-speaker variability

Inter-speaker variability is influenced by a range of physiological, demographic, regional, and social factors unique to each speaker. These individual differences are reflected in various aspects of speech and serve as cues that enable the distinction between speakers, or even between groups of speakers [105].

2.3 Visual parametrization

Visual parametrization typically refers to the process of representing visual information, such as images or visual patterns, using a set of parameters or features that capture relevant characteristics of the visual data. It involves reducing the complexity of the visual data while retaining essential information, making it easier to analyze, process, or compare images. In this section, we present the techniques used for characterizing the hand form and extracting vector descriptors. To achieve this objective, various types of global and local techniques are employed.

2.3.1 Global techniques

2.3.1.1 Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is an effective technique for extracting distinctive features from an image, as it enables analysis at multiple resolution levels. Using low-pass and high-pass filters, DWT decomposes the original image, with the low-pass filter providing an approximation of the image and the high-pass filter revealing its details. This approximation can then be further decomposed into additional levels of approximation and detail, depending on the specific application requirements [106]. The DWT is defined as:

$$W_\psi(m, n, s) = \sum_x \sum_y f(x, y) \cdot \psi_{m,n,s}(x, y) \tag{2.6}$$

where:

- $W_\psi(m, n, s)$ is the wavelet coefficient at position (m, n) and scale s,
- $f(x, y)$ is the original image signal (pixel intensities),
- $\psi_{m,n,s}(x, y)$ is the scaled and shifted wavelet function,

$$\psi_{m,n,s}(x, y) = \frac{1}{\sqrt{s}} \psi \left(\frac{x - m}{s}, \frac{y - n}{s} \right) \tag{2.7}$$

- x, y are the spatial coordinates of the image.
- m, n are the translation indices.
- s is the scale parameter (determines the level of decomposition).

From an image $f(x, y)$, two sets of coefficients are calculated: approximation coefficients LD and detail coefficients HD. These coefficients are obtained by convolving $f(x, y)$ with the low-pass filter $h_0(n)$ for approximation and with the high-pass filter $h_1(n)$ for detail, followed by a decimation operation (subsampling by 2, typically denoted by $\downarrow 2$, as shown in Figure2.2 .

From the obtained coefficients LD and HD, the original signal can be reconstructed using inverse filters $g_0(n)$ and $g_1(n)$, preceded by an interpolation operation (upsampling by 2, typically denoted by $\uparrow 2$.

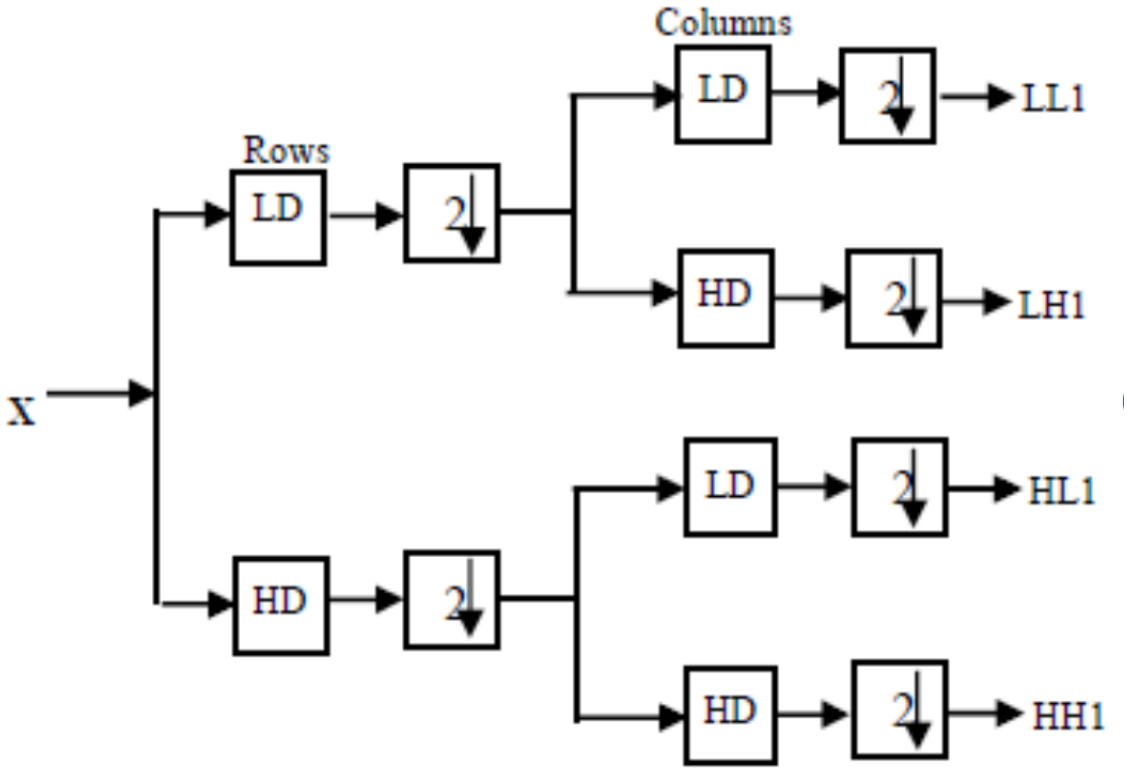


Figure 2.2: Decomposition at one level by wavelet transform.

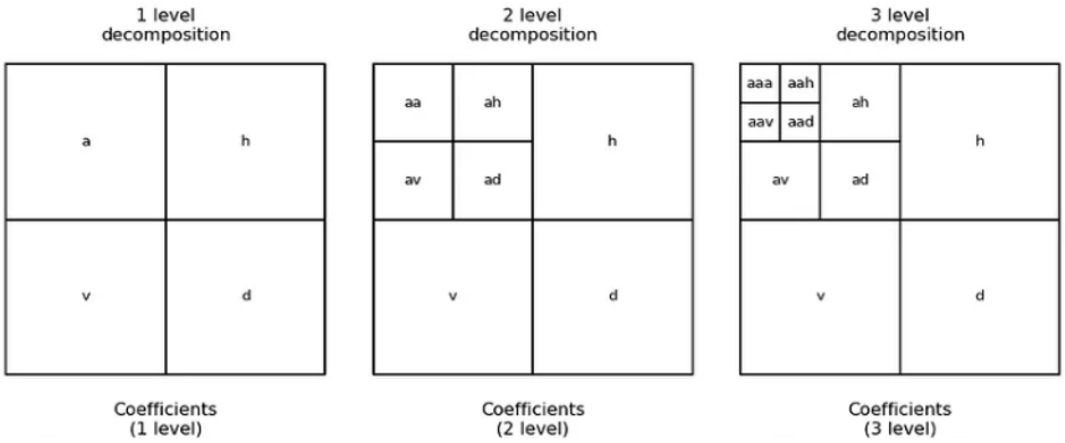


Figure 2.3: Wavelet decomposition at three resolution levels, LD and HD are respectively the approximation and detail coefficients at the $i-th$ level.

Depending on the targeted application, the approximation coefficients can also be decomposed into two sets of high and low-frequency coefficients. These successive decompositions of the approximations form a tree structure known as Mallat's decomposition tree [107]. Figure 2.3 illustrates the decomposition of f(x,y) into three resolution levels. At each decomposition level, the temporal resolution is reduced by half, and the frequency resolution is doubled. In this way, the

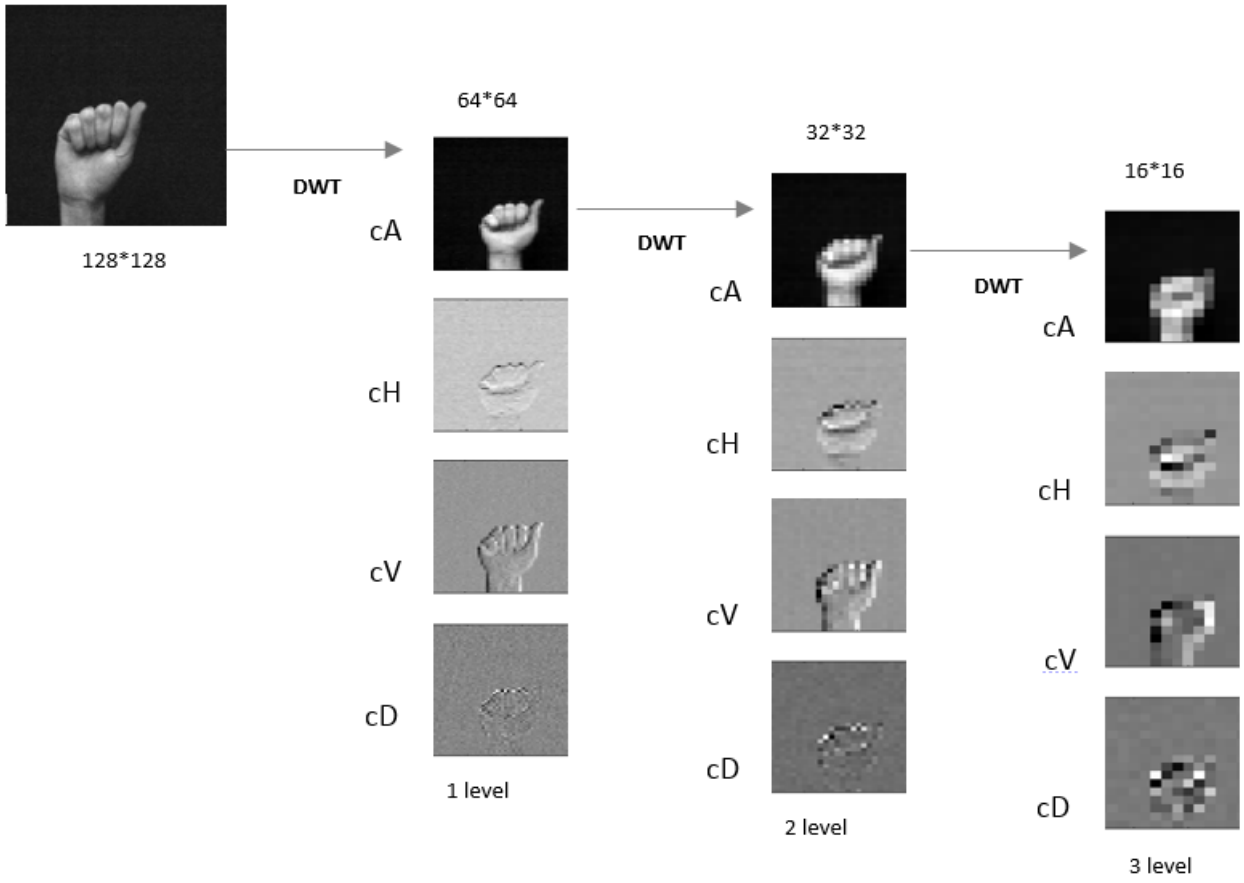


Figure 2.4: DWT decomposition, where cA, cH, cV and cD are the Approximation, Horizontal, Vertical and Diagonal image respectively.

DWT provides good frequency resolution for low frequencies and good temporal resolution for high frequencies.

In the proposed system, 3 levels of DWT are applied as shown in figure 2.4, where the approximated image is used in each subsequent levels of decomposition with Haar wavelet. The coefficients derived from the wavelet transform will be used as feature vectors that represent the gesture shape to be analyzed for all participants and the six dynamic gestures.

The DWT offers several advantages for signal and image processing. One of its key benefits is its ability to analyze data at multiple resolutions. Allowing for both coarse and fine detail extraction. This multi-resolution approach is particularly useful in tasks such as image compression, denoising, and feature extraction. It also provides localization in both time and frequency, which is crucial for analyzing images that have varying characteristics over time. Additionally, DWT is computationally efficient, especially when compared to traditional methods like the Fourier Transform, as it reduces the amount of data required while preserving important features. It is also robust to noise, making it valuable in environments where signal clarity is compromised.

However, there are also limitations to consider. DWT can introduce redundancy in the data, particularly at higher decomposition levels, which may result in larger storage requirements if not carefully managed. Edge effects, where the wavelet fails to accurately represent data at the boundaries of an image, can also lead to inaccuracies. Another limitation is the dependence on the choice of wavelet function—different wavelets can yield different results, so selecting the optimal one for a specific application can be challenging. Finally, while DWT is efficient for 1D signals, applying it to higher-dimensional data (like 2D images) can increase computational complexity, requiring more memory and processing power.

2.3.1.2 Contourlet transform (CT)

CT is a directional multi-resolution image representation constructed by applying two successive decomposition stages, multi-directional decomposition followed by multi-scale decomposition [108] . The basic idea of the CT is use a multi-scale decomposition to detect the singularity points of edges firstly, then, make the singularity points of position close together into contour section according to the information of direction.The CT adopts the multi-resolution structure of the Laplacian Pyramid, which was previously introduced by Burr and Adelson in 1983 for capturing singularity points in images[109, 110].

- Laplacian Pyramid (LP) :The basic idea of The LP decomposition is to decompose an image into a series of band-pass filtered images at multiple scales. The decomposition process is typically carried out in a pyramidal structure, where each level of the pyramid represents a different scale of the original image. The levels at the bottom of the pyramid contain high-frequency details, while the levels at the top contain low-frequency components [109, 110]. Figure2.5 depicts this decomposition process, where H and G are called (low-pass) analysis and synthesis filters, respectively, and M is the sampling matrix.

- Directional Filter Bank (DFB): Bamberger and Smith constructed a two-dimensional DFB that can be maximally decimated while achieving perfect reconstruction.This efficient reconstruction is achieved through an $l - level$ binary tree decomposition, generating 2^L sub-bands with wedge-shaped frequency partitioning, as illustrated in Figure 2.6 . The original construction of the DFB involves modulating the input image and using quincunx filter banks with diamond-shaped filters [111, 112] . To obtain the desired frequency partition, a complicated tree expanding rule has to be followed for finer directional subbands .The DFB can be

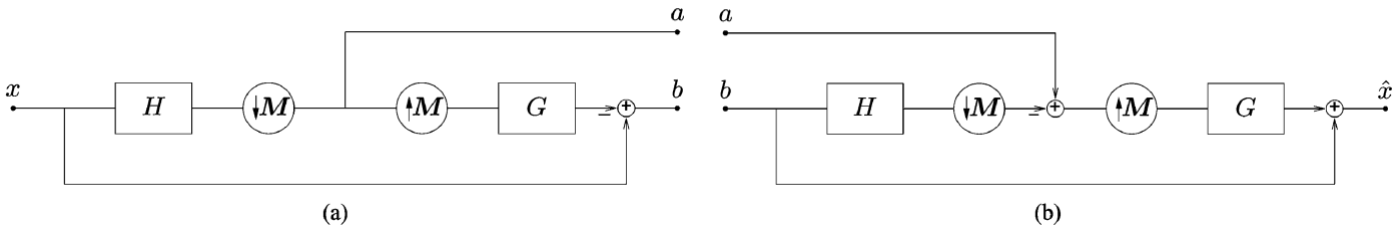


Figure 2.5: Laplacian pyramid. (a) One level of decomposition. The outputs are a coarse approximation $a[n]$ and a difference $b[n]$ between the original signal and the prediction. (b) The new reconstruction scheme for the Laplacian pyramid [1]

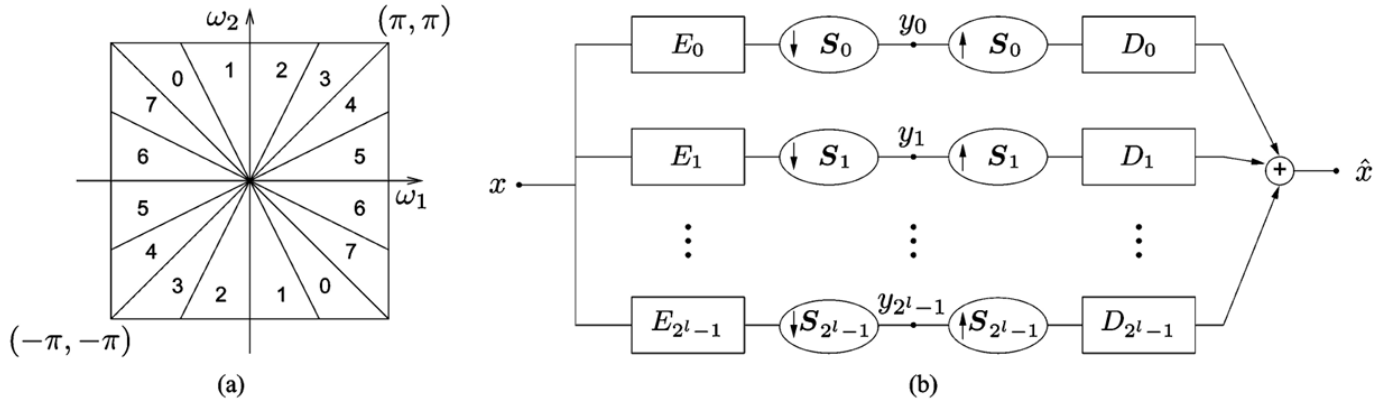


Figure 2.6: Directional filter bank. (a) Frequency partitioning where l = 3 and there are $2^3 = 8$ real wedge-shaped frequency bands. Subbands 0–3 correspond to the mostly horizontal directions, while subbands 4–7 correspond to the mostly vertical directions. (b) The multichannel view of an $l - level$ tree-structured directional filter bank [1]

seen as a set of filters that are designed to be sensitive to different directional features present in the input signal.

Figure 2.7 shows a multiscale and directional decomposition using a combination of a LP and DFB. Bandpass images from the LP are fed into a DFB so that directional information can be captured. The scheme can be iterated on the coarse image. The combined result is a double iterated filter bank structure, named contourlet filter bank, which decomposes images into directional subbands at multiple scales.

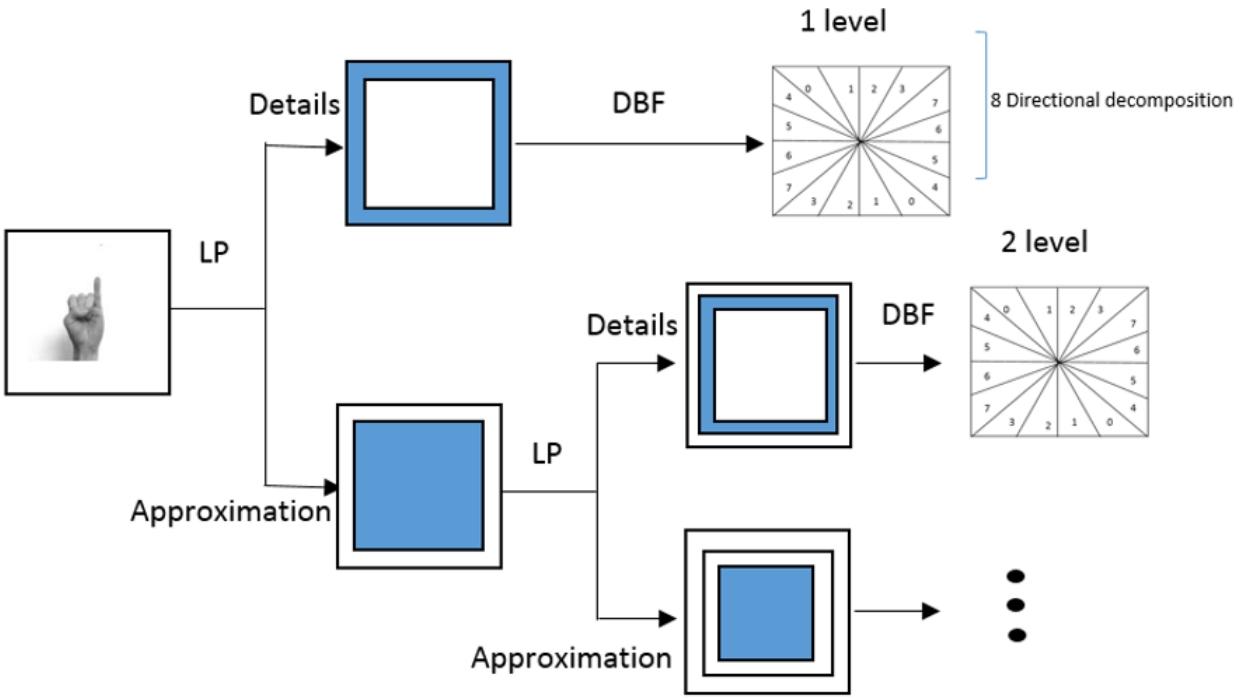


Figure 2.7: Hand posture decomposed using2 levels contourlet transform

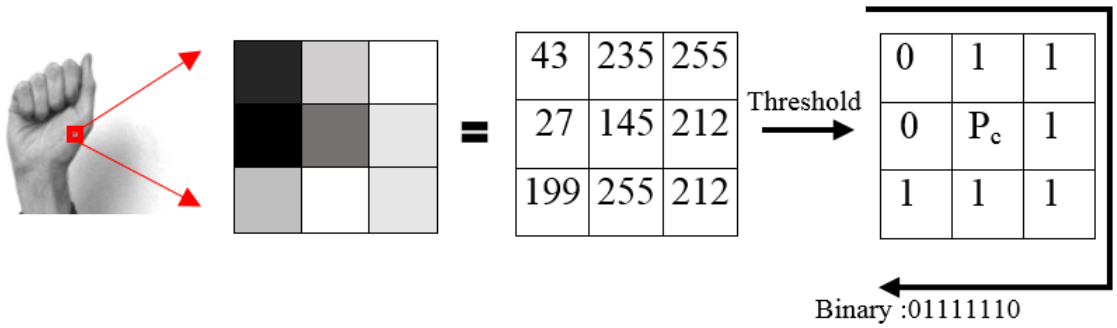


Figure 2.8: LBP code calculation [2]

2.3.2 Local techniques

2.3.2.1 Local Binary Patterns

LBP [113] is a simple technique defined as a grayscale invariant texture measure. The LBP operator assigns a label for each pixel by the relative gray levels of its neighboring pixels. Fig. 2.8 shows an illustration with $3 * 3$ neighborhood.

This label is calculated following a comparison of the gray levels between the central pixel P_c and its neighbors P_n, which is given as follows:

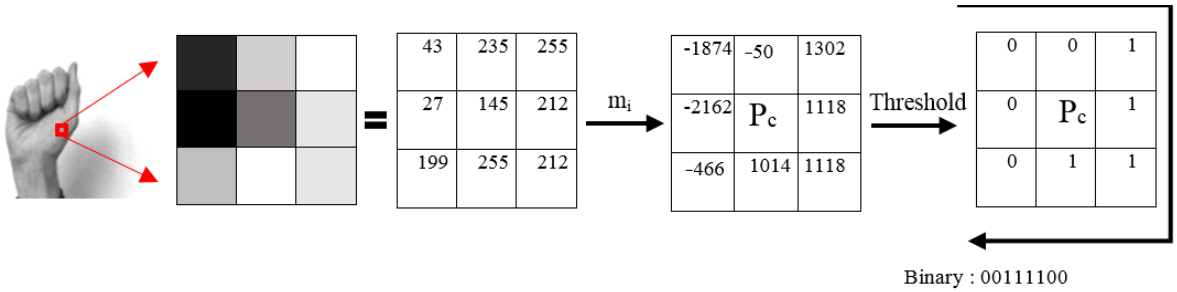


Figure 2.9: LDP code calculation [2]

$$LBP(x_c, y_c) = \sum_{i=0}^{n=7} s(P_n - P_c).2^n,$$
$$\text{with } s(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

where (x_c, y_c) are the coordinates of P_c. The value is set to one if the gray level of the neighboring pixel is higher or equal to P_c. Otherwise, it is set to zero. The LBP code of the current pixel is generated by concatenating the results binomially to form a decimal number.

2.3.2.2 Local Directional Pattern (LDP)

LDP [60, 61] is a gray-scale texture pattern calculated by comparing the relative edge response value of a current pixel in different directions. An eight-bit binary code is assigned to this pixel. The binary code derived by formula (2.9) is obtained by convolving the $3 * 3$ neighborhood with eight Kirsch masks $(M_0 \dots M_7)$, which are shown in Fig 2.9.

$$LDP_k(x_c, y_c) = \sum_{i=0}^{n=7} s(m_n - m_i).2^n$$
$$\text{with } s(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

The top k response values $m_i (i = 0, 1, \dots, 7)$ of these eight directional edges are selected and set to 1 while the remaining $(8 - k)$ responses are set to 0.

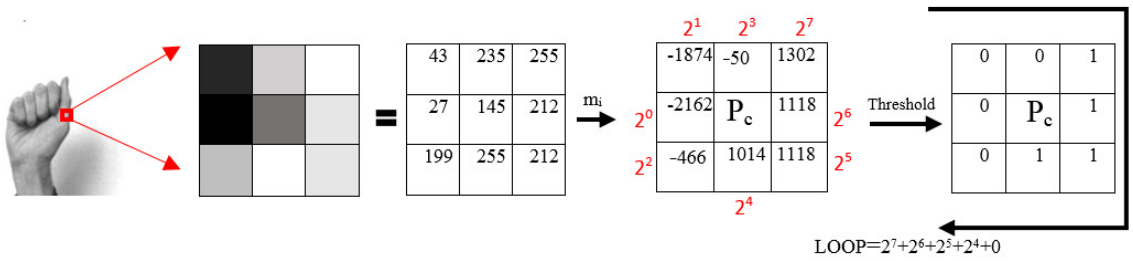


Figure 2.10: LOOP code calculation [2]

2.3.2.3 Local Optimal-Oriented Pattern (LOOP)

The LOOP descriptor is considered as a nonlinear texture descriptor, taking advantage of LBP and LDP [65]. The incorporation of these two descriptors is affected by assigning the binarization weight to each neighboring pixel corresponding to the strength of Kirsch output that adds dependency to the orientation of that pixel. The LOOP value for the pixel (x_c, y_c) is calculated as:

$$\begin{aligned} LOOP(x_c, y_c) &= \sum_{i=0}^{n=7} s(P_n - P_i).2^{W_n} \\ \text{with } s(x) &= \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{2.10}$$

where W_n each pixel according to the rank of the magnitude of P_n among the eight directional edge response of Kirsch masks. Moreover, the LOOP algorithm negates the empirical assignment of the value of the parameter k in the traditional LDP descriptor. Fig. 2.10 depicts LOOP code calculation.

2.3.2.4 Local Gabor Binary Pattern Histogram Sequence (LGPHS)

The LGBPHS operator is a novel non-statistic feature descriptor, which is introduced by W. Zhang et al [114] for face representation and recognition. The basic idea of LGBPHS is to describe the input feature as a "histogram sequence" by the following procedures:

- The input hand image is convolved with a group of multi-scale and multi-orientation Gabor filters to obtain multiple Gabor Magnitude Pictures (GMPs). The Gabor filters are defined as follow:

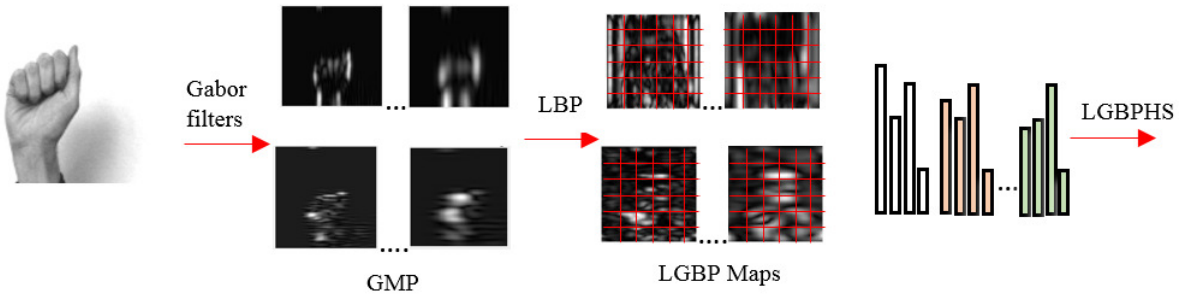


Figure 2.11: LGBPHS code calculation with $O = 8$ and $S = 5$ [2]

$$\psi_{o,s} = \frac{\|k_{o,s}\|}{\sigma^2} e^{-\frac{\|k_{o,s}\|^2 \|z\|^2}{2\sigma^2}} \left[e^{ik_{o,s}z} - e^{-\frac{\sigma^2}{2}} \right], \tag{2.11}$$

where O and S define, respectively, the orientation and scale of the Gabor filters and $\|.\|$ denotes the norm operator. $k_{o,s} = k_s e^{i\phi o}$ is the wave vector, where $k_s = k_{max}/\lambda^s$ and $\phi_o = \frac{\pi o}{8}$. λ Is the spacing factor between filters in the frequency domain.

- Each GMP is converted to Local Gabor Binary Pattern (LGBP) maps by the LBP operation (equation 2.8).
- Each LGBP map is divided into non-overlapping rectangle regions with specific size and histogram which is extracted from each region.
- All the histograms estimated from the regions are concatenated into a histogram sequence as the final hand representation. Fig. 2.11 depicts LGBPHS code calculation with $O = 8$ and $S = 5$.

2.3.2.5 Modified$_i$ Local Optimal-Oriented Pattern (M$_i$-LOOP)

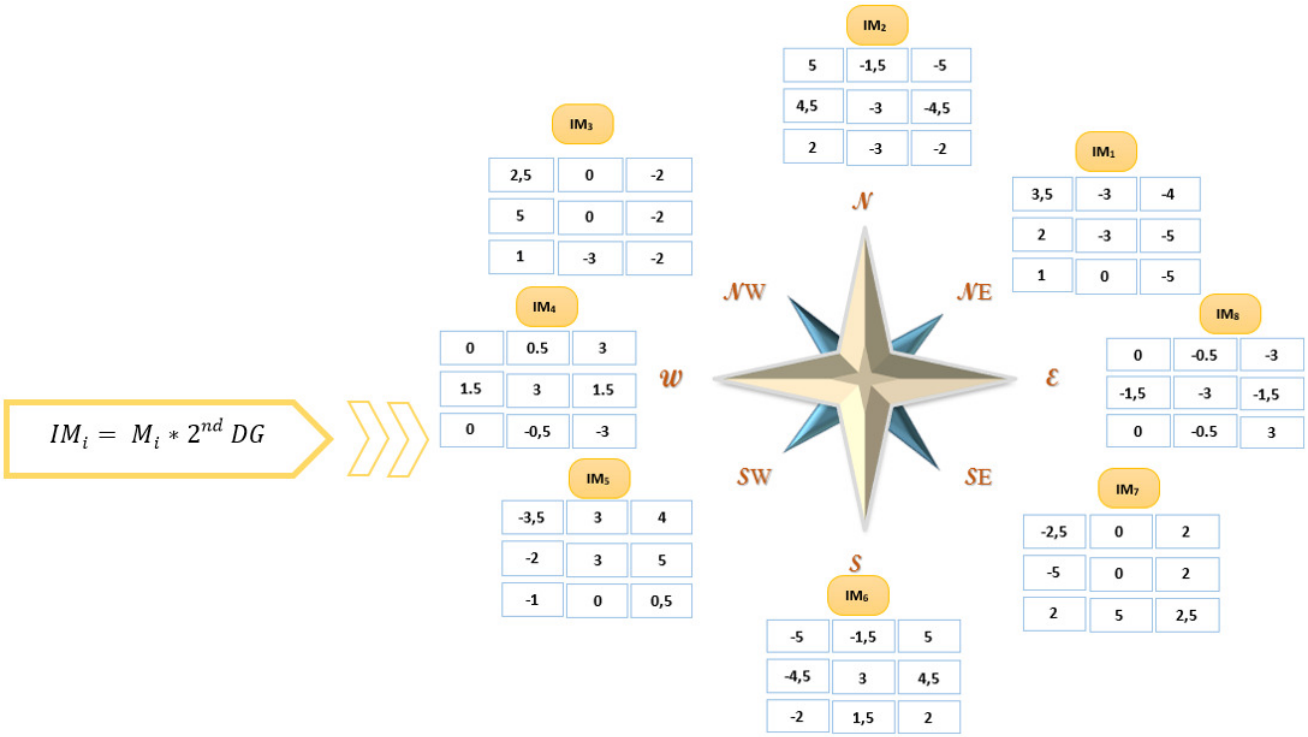
The proposed method consists of addressing the feature extraction technique by introducing a modified binary version of the loop descriptor called M$_i$-LOOP. The rotation invariant from LBP and the directional features from LDP are combined to form the proposed M$_i$-LOOP. In contrast to the conventional LOOP which use only the eight directional edge response of Kirsch masks, the M$_i$-LOOP embedded two different types of compass masks:2nd derivative- Gaussian mask to avoid noise perturbation and $M_i(i = 1, \ldots .5)$ mask to increase the encoded structural information.

We propose in our work two new masks called proposed-1 (M$_4$) and proposed-2 (M$_5$) as shown in Fig 2.12. Consequently, M$_1$-LOOP, M$_2$-LOOP, M$_3$-LOOP, M$_4$-LOOP, M$_5$-LOOP correspond

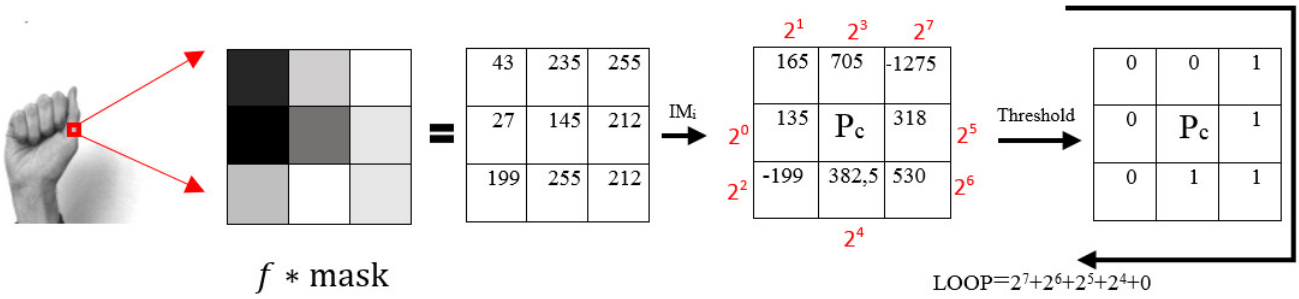
to the Modified-Local Optimal Oriented Pattern using the Robinson, five level simple, Prewitt, proposed-1- and proposed-2, respectively.

Direction of edge	Direction of gradient	M_1 Robinson	M_2 Five level simple	M_3 Prewitt	M_4 Proposed-1-	M_5 Proposed-2-	2nd derivative of Gaussian filter
0	North	$\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ 0 & 0 & 0\\ -1 & -1 & -1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 1\\ -1 & -1 & -1\end{bmatrix}$	$\begin{bmatrix}0 & -1 & 0\\ -1 & 8 & -1\\ 0 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & 1\\ -1 & 0 & 1\\ -1 & 0 & 1\end{bmatrix}$	
1	Northwest	$\begin{bmatrix}0 & 1 & 2\\ -1 & 0 & 1\\ -2 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & -1\\ 0 & -1 & -1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ 1 & -2 & -1\\ 1 & -1 & -1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & -1\\ 0 & 8 & 0\\ -1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ -2 & -2 & -2\\ 1 & 1 & 1\end{bmatrix}$	
2	West	$\begin{bmatrix}1 & 2 & 1\\ 0 & 0 & 0\\ -1 & -2 & -1\end{bmatrix}$	$\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}1 & 1 & -1\\ 1 & -2 & -1\\ 1 & 1 & -1\end{bmatrix}$	$\begin{bmatrix}0 & -1 & 0\\ -1 & 8 & -1\\ 0 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & -1\end{bmatrix}$	$\frac{1}{4} * \begin{bmatrix}1 & -2 & 1\\ 2 & -4 & 2\\ 1 & -2 & 1\end{bmatrix}$
3	Southwest	$\begin{bmatrix}2 & 1 & 0\\ 1 & 0 & -1\\ 0 & -1 & -2\end{bmatrix}$	$\begin{bmatrix}0 & -1 & -1\\ 1 & 0 & -1\\ 1 & 1 & 0\end{bmatrix}$	$\begin{bmatrix}1 & -1 & -1\\ 1 & -2 & -1\\ 1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & -1\\ 0 & 8 & 0\\ -1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}1 & -2 & 1\\ 1 & -2 & 1\\ 1 & -2 & 1\end{bmatrix}$	
4	South	$\begin{bmatrix}1 & 0 & -1\\ 2 & 0 & -2\\ 1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}-1 & -1 & -1\\ 0 & 0 & 0\\ 1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & -1 & -1\\ 1 & -2 & 1\\ 1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}0 & -1 & 0\\ -1 & 8 & -1\\ 0 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & 1\\ 2 & 0 & -2\\ -1 & 0 & 1\end{bmatrix}$	
5	Southeast	$\begin{bmatrix}0 & -1 & -2\\ 1 & 0 & -1\\ 2 & 1 & 0\end{bmatrix}$	$\begin{bmatrix}-1 & -1 & 0\\ -1 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & -1 & 1\\ -1 & -2 & 1\\ 1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & -1\\ 0 & 8 & 0\\ -1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}1 & -2 & 1\\ 0 & 0 & 0\\ 1 & -2 & 1\end{bmatrix}$	
6	East	$\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & 1\\ -1 & 0 & 1\\ -1 & 0 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & 1 & 1\\ -1 & -2 & 1\\ -1 & 1 & 1\end{bmatrix}$	$\begin{bmatrix}0 & -1 & 0\\ -1 & 8 & -1\\ 0 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$	
7	Northeast	$\begin{bmatrix}-2 & -1 & 0\\ -1 & 0 & 1\\ 0 & 1 & 2\end{bmatrix}$	$\begin{bmatrix}0 & 1 & 1\\ -1 & 0 & 1\\ -1 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 1 & 1\\ -1 & -2 & 1\\ -1 & -1 & 1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & -1\\ 0 & 8 & 0\\ -1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}-1 & -1 & -1\\ 0 & 0 & 0\\ 1 & 1 & 1\end{bmatrix}$	

Figure 2.12: Usual and proposed Masks used for Modified$_i$ -LOOP [2]



(a) ***Eighth edge responses IM$_i$***



(b) ***M$_i$-LOOP calculation***

Figure 2.13: Calculation of the edge responses IM_i and their association to the pixels of the 3×3 square neighborhood. $*$ is the convolution [2]

To avoid the above weaknesses of the traditional masks, we investigate a new coding scheme that integrates different concepts. For the first proposed filter-1, the coefficients are obtained through the integration of gaussian smoothing and laplacian filter [115, 116]. The origin of this mask is always the central pixel. The impulse response of an image convolved with the suggested filter satisfy the following conditions:

- High positive value $\rightarrow$ white dot on black background
- High negative value $\rightarrow$ black dot on white background

The direction of the eight responses labeled $\mathrm{M}_4(D_e, D_g)$, D_e (Direction of edge)=$0 \ldots 7$, D_g (Di-

rection of gradient: North, Northwest, Northeast, West,East, South, Southwest,Southeast) obtained by taking a single mask and rotating out to eight major compass orientation cf Fig.2.12).

For the second proposed-2 mask, the coefficients are obtained through the following stages:

- $M_5(0, N)$ = rotation of $\frac{\pi}{2}$ of $M_5(0, N)$
- $M_5(1, NW)$ = rotation of $\frac{\pi}{2}$ of average filter $+2^{nd}$ derivation
- $M_5(2, W)$ = gradient filter + prewitt$(2, W)$
- $M_5(3, SW)$ = rotation of $\frac{\pi}{2}+ M_4(1, NW)$
- $M_5(4, S)$ = filtre gradient+ robinson$(4, S)$
- $M_5(5, SE)$ = rotation of $\frac{\pi}{2}$ of $\mathrm{M}_4(4, S)$
- $M_5(6, E)$ = matrix(1)
- $M_5(7, NE)$ = rotation of $\frac{\pi}{2}$ of $\mathrm{M}_4(0, N)$

Given $f(x, y)$ that represents the structure of the hand and its intensity transitions. The proposed M_i-LOOP value for a pixel is summarized as follows:

1. Apply the 2^{nd} derivative-Gaussian mask on the overlapping 3×3 grayscale image f by formula (2.12)

$$f(x, y)' = f(x, y) * filter = \sum_{i}^{M}\sum_{j}^{N} f(x - i, y - j)filter(i, j), \tag{2.12}$$

   where (M,N) size of the image and filter=2^{nd} derivative-Gaussian(cf.2.12).

2. Calculate the eight responses labeled $IM_0, IM_1, \ldots IM_7$ by convolving of $M_i, (i = 1, \ldots .5)$ mask and 2^{nd} derivative-Gaussian mask.

$$IM_i = M_i * 2^{nd}DGmask. \tag{2.13}$$

   The resulting IM_i represent the new 8 masks used later in step.

3. The proposed M_i-LOOP value for a pixel is given by equation (2.14).

$$M_i - LOOP(x_c, y_c) = \sum_{i=1}^{n=7} s(P_n - P_i).2^{W_n},$$
$$\text{with } s(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.14}$$

where (x_c, y_c) are the coordinates of P_c with an intensity i_n and w_n an exponential weight w_n ranging between 0 and 7 according to the rank of the magnitude of P_n among the eight directional relative edge responses. The eight P_n neighboring pixels (that represents the structure of the hand f' and its intensity transitions in a 3×3 square neighborhood) are oriented by the direction of the eight responses labeled $IM_0, IM_1, \ldots IM_7$ (corresponding to pixels with intensity $I_0, (n = 0, ..., 7)$. Figures 2.13(a) and 2.13(b) illustrate step by step the process of calculation of edge responses and their association to the pixels of the 3×3 neighborhood.

2.4 Conclusion

In this chapter, we focused on the acoustic and visual characterization of our data-set.Several feature extraction have been presented in the context of hand/speech recognition systems.

In the realm of acoustic parametrization, we focused on three essential techniques: MFCC,first and second derivatives and LPC. by applying these techniques to the audio clips in our data-set, we were able to transform raw audio signals into compact and informative feature vectors.

Continuing our exploration, we delved into visual parametrization, where we investigated various global and local techniques for analyzing and representing visual information. Global techniques allowed us to obtain a holistic understanding of entire images, capturing high-level features that are vital in tasks like object recognition and scene classification. On the other hand, local techniques focused on extracting distinctive patterns and descriptors from smaller regions within an image. We also explored specific descriptors like LBP, LDP, LOOP, and M_i-LOOP, which provided valuable insights into the visual characteristics of our data-set.

Chapter 3

Proposed Systems and Methodologies

3.1 Introduction

Audiogestural speech recognition, which combines gestural information (such as hand movements) with acoustic signals, is a promising approach to improving speech recognition, especially in noisy environments. The key idea is that gestures associated with speech, although independent of the sounds produced, can convey complementary information useful for identifying spoken words or phrases.

The main advantage of this method is that gestural signals are not affected by ambient acoustic noise. Therefore, they can be used to enhance or supplement the performance of traditional speech recognition systems. By combining the two sources of information (acoustic and gestural), it becomes possible to design more robust systems capable of operating effectively in conditions where background noise makes acoustic recognition alone difficult.

In this chapter, we present our comprehensive work on bimodal recognition systems, starting with a detailed exploration of speech recognition. We developed a speech recognition system using LPC and MFCC as will as first and second derivative descriptors for characterization, coupled with a classification based on an MLP network.

In addition, we proposed a gesture recognition system for static shapes and dynamic sequences of hand gestures, using various descriptors for characterization and classification methods based on SVM networks. The highlight of our research lies in the integration of these modalities, resulting in a bimodal system combining speech and gestures. We also describe the integration models that characterize the fusion of audio and gestural information, leveraging advanced approaches

such as MLP networks, SVM, and deep learning. This system seamlessly combines spoken language and non-verbal communication, demonstrating the potential of bimodal interaction to enhance human-computer interfaces and promote accessibility, particularly for individuals with hearing impairments.

3.2 Speech recognition system using MLP classifier

We propose a speaker identification system for the Arabic and Berber languages, both of which are widely spoken in Algeria. Our approach focuses on speaker identification using a dependent mode of isolated words for both languages, employing Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC) as will as first and second derivative for feature extraction. To reduce the dimensionality of the extracted features, Principal Component Analysis Principal Component Analysis (PCA) is applied. Two datasets are utilized: one dedicated to the Arabic language and the other to Algerian Berber language. The practical implementation, including detailed experimental procedures, data collection methods, and parameter configurations, will be presented subsequently.

3.2.1 Audio Dataset description

We use two datasets in our work. The first dataset consists of isolated words in the Arabic language, recorded at the Speech Communication and Signal Processing Laboratory, University of Science and Technology Houari Boumediene in Algiers. The second dataset comprises continuous speech signals in the Berber language, known as CSB1.0 [117, 118].

3.2.1.1 Arabic dataset

This dataset contains isolated words comprising six Arabic words spoken by six speakers, consisting of three males and three females. Each speaker repeated the sequence, including simple, geminate, and elongated versions, ten times, maintaining an average speed to ensure clear articulation while avoiding any external disturbances. All speakers repeated each sequence ten times. The total number of analyzed words amounts to 1080 files. The recording frequency is 16 kHz, with a recording duration of 2 seconds [117, 118].

3.2.1.2 Continuous Berber dataset

Continuous speech Berber is a set of sentences representing Berber language. We present six sentences of duration 3 and 4s. The dataset includes speech signals from six different subjects. The speech signals are acquired during 2 or 3s with different sessions to consider all variations at a sampling rate 16KHz. During the recording, each repetition has been analyzed to ensure that the entire sequences have been properly stored and avoiding any external interference. The chosen sequences are as follows: “assagui yelha lhal”; “azeka anrouh arthemourth”; “anzoum Remthane g nevdhou”; “yetchak izeme yellozen”; “yeghine wagroud aghour yemasse” and“thelha thazalith gakhem rebi”. The first four sentences have duration of 2 s while the last sequences have equal duration of 3s. The segmentation has been performed manually by using the recording software Wave-surfer. The total files used for training and test are 432, concerning 12 repetitions of each sequences for the six speakers [117, 118].

3.2.2 Evaluation setup

We apply several approaches in our identification system . The aim of our analysis is to choose and investigate the appropriate characteristic feature vector descriptors for speaker recognition. Figure 3.1 illustrates the proposed method’s diagram. Our analysis will involve describing and comparing the recognition rates achieved for each characterization, namely LPC, LPC-PCA, MFCC, and MFCC-PCA techniques, using an MLP classifier for the two datasets.

PCA is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the data, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps. The quality of the PCA model can be evaluated using cross-validation techniques.Mathematically, PCA depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition Singular value decomposition (SVD) of rectangular matrices.

The A multi-layer perceptron (MLP) model is one of the most widely applied and researched Artificial Neural Network model. MLP networks are normally applied to performing supervised learning tasks, which involve iterative training methods to adjust the connection weights within the network. This is commonly formulated as a multivariate non-linear optimization problem over a very high-dimensional space of possible weight configurations [119].

An MLP is a network of simple processing units arranged into a hierarchical model of layers. The units (neurons or nodes) in the first (input) layer are connected to nodes in the following layer(s) (hidden layers), to the final layer (the output layer). Numerical input vectors of (patterns) are presented at the input layer, and activity flows through the network to the output layer. Connections have a numerical weight value associated with them, and the signal transmitted via a connection is multiplied by the weight value. Each unit computes some function of the sum of its weighted inputs, and transmits the result through its output connections. In this way an MLP implements a mapping from input space to output space [119].

In our case, the input layer receives the training matrix, where each column represents the LPC, LPC-PCA, MFCC, and MFCC-PCA feature vectors describing our speech dataset.

The hidden layer contains hidden nodes, which are chosen by the user. In our work, we iteratively select the appropriate number of hidden layers that achieve the best performance, ranging from 1 to 300.

The output layer contains the corresponding outputs. Our dataset is both trained and tested using the MLP network, employing various architectures in our speaker identification system.The first and second architectures explore the advantages of LPC and LPC reduced by PCA, respectively. The third and fourth architectures are dedicated to MFCC and MFCC reduced by PCA, respectively.

For each individual speaker and for the total segmented words, LPC, LPC-PCA, MFCC, and MFCC-PCA feature extraction techniques are applied. We allocate 50% of the data for training and the remaining 50% for testing. The MLP classifier is employed to identify the total number of speakers.

In this context, the input layer accepts a matrix of size (N*36) for the Berber dataset and (N*30) for the Arabic dataset, where N represents the length of our feature vector used for speech characterization. The target matrix used for training contains correct output positions marked as 1, and 0 for the remaining positions.

Different transfer function were tested in hidden and output layer.The log-sigmoid transfer function gives the best result in term of accuracy.The log-sigmoid transfer function is applied to both the hidden and output layers. Convergence is achieved using a learning rule known as the Conjugate Gradient (CG). TRAINSCG is a network training function responsible for updating weight and bias values based on the scaled conjugate gradient method.

We evaluated various training algorithms for our classification task, including Gradient De-

scent algorithms (trainrp, traingda), Conjugate Gradient algorithms (trainscg), and Quasi-Newton algorithms (trainoss). These network training functions update weight and bias values using different methods, such as gradient descent with adaptive learning rate (for traingda); resilient back propagation algorithm (for trainrp); scaled conjugate gradient method (for trainscg) and one-step approximation method for trainoss. These training algorithms can train any network as long as its weight, net input and transfer functions, which have derivative functions [120]. To evaluate our approaches, the standard accuracy is used, defined by:

$$Accuracy = \frac{\text{Number of files correctly classified}}{\text{Total number of files}} \times 100 \tag{3.1}$$

It's important to note that determining the fastest training algorithm for a given problem can be extremely challenging. It depends on several factors, including the problem's complexity, the size of the training dataset, the number of weights and biases in the network, the specified error goal, and whether the network is being utilized for pattern recognition (discriminant analysis) or function approximation (regression).

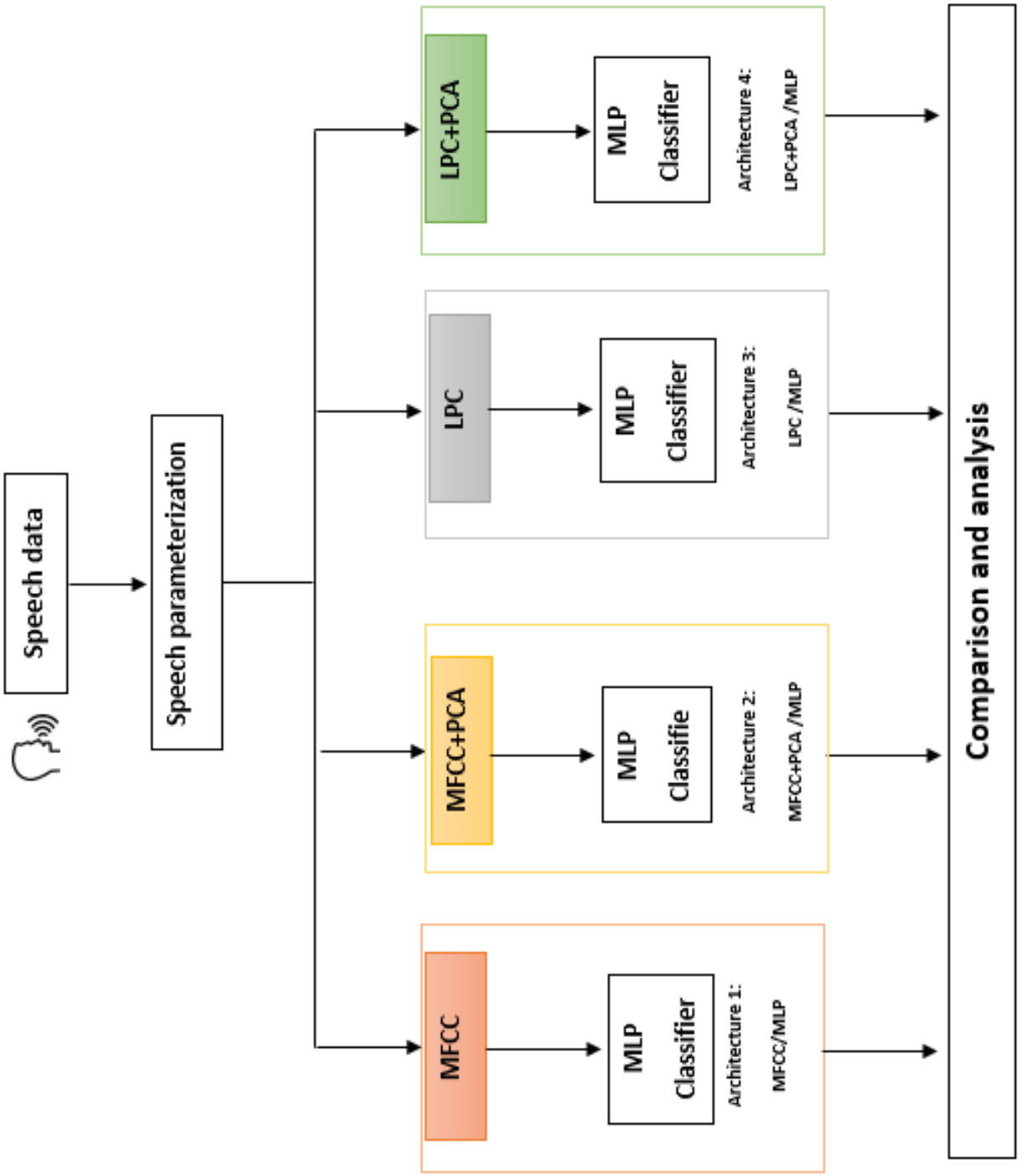


Figure 3.1: Diagram of the proposed method.

3.3 Hand and sign gesture recognition system using SVM classifier

We propose a hand gesture recognition system (HGRS) and sign gesture recognition system (SGRS) that recognizes Arabic, American Sign alphabet and dynamic gesture. Figure 3.1 describes the diagram of the proposed method. Our main motivation is to investigate the efficiency of the proposed textural descriptors for hand parameterization. Our analysis should describe and compare the accuracy obtained for each descriptor, namely LBP, LGBPHS, LDP, LOOP and M_i-lOOP where oneagainst all SVM algorithm is used for hand gesture recognition phase. For this purpose, Five architectures are proposed for designing our HGRS and SGRS, named respectively, LBP/SVM, LDP/SVM, LOOP/SVM LGBHS/SVM and M_i-LOOP/SVM.

3.3.1 Gesture Dataset description

Experiments are conducted on five datasets. Two datasets were devoted for ArSL and two others for ASL, whereas the fifth one for dynamic gestures. The first dataset called “EL Halawani dataset”, This database includes 30 class of Arabic alphabet, where each class contains 60 posture, made by 60 person. The second dataset called “ArSL2018 dataset”. It consists of 54,049 images for the 32 Arabic sign language and alphabets collected from 40 persons. The third datastet called “Jochen Triesch dataset” consists of gray-scale images for all of the 10 gestures (a, b, c, d, g, h, i, l, v, y), made by 24 peoples with different backgrounds (white, black, complex), where only white and black background are used in our work. The fourth one, called IEEE Asl dataset, was collected by capturing the static gestures of the ASL alphabet, made by 8 people. The fifth dataset called ”Sebastien Marcel dataset,which is a dynamic dataset, contains four different gestures (clic, no, rotate, stop). Each gesture is represented by 12 video sequences, and each sequence includes 55 frames made by different person. Table 3.1 describes each datasets and figure 3.2 shows some examples from each one.

3.3.2 Evaluation setup

In order to evaluate the performance of the proposed SGRS and HGRS and to assess the performance of the proposed descriptors, we carried out a set of classification experiments using five different datasets as described in section 3.3.1. We used standard evaluation protocol for the five datasets

Table 3.1: Datasets description [2].

Named	Type	Class number	Images number	collected by (Person)	category
El Halawani	ArSL	30	1800	60	Static
ArSL2018 Arabic	ArSL	32	54,049	40	Static
Jochen Triesch	ASL	10	240	12	Static
IEEE	ASL	26	1040	8	Static
Sebastien Marcel	Diffrent	6	55 F/s	12	Dynamic

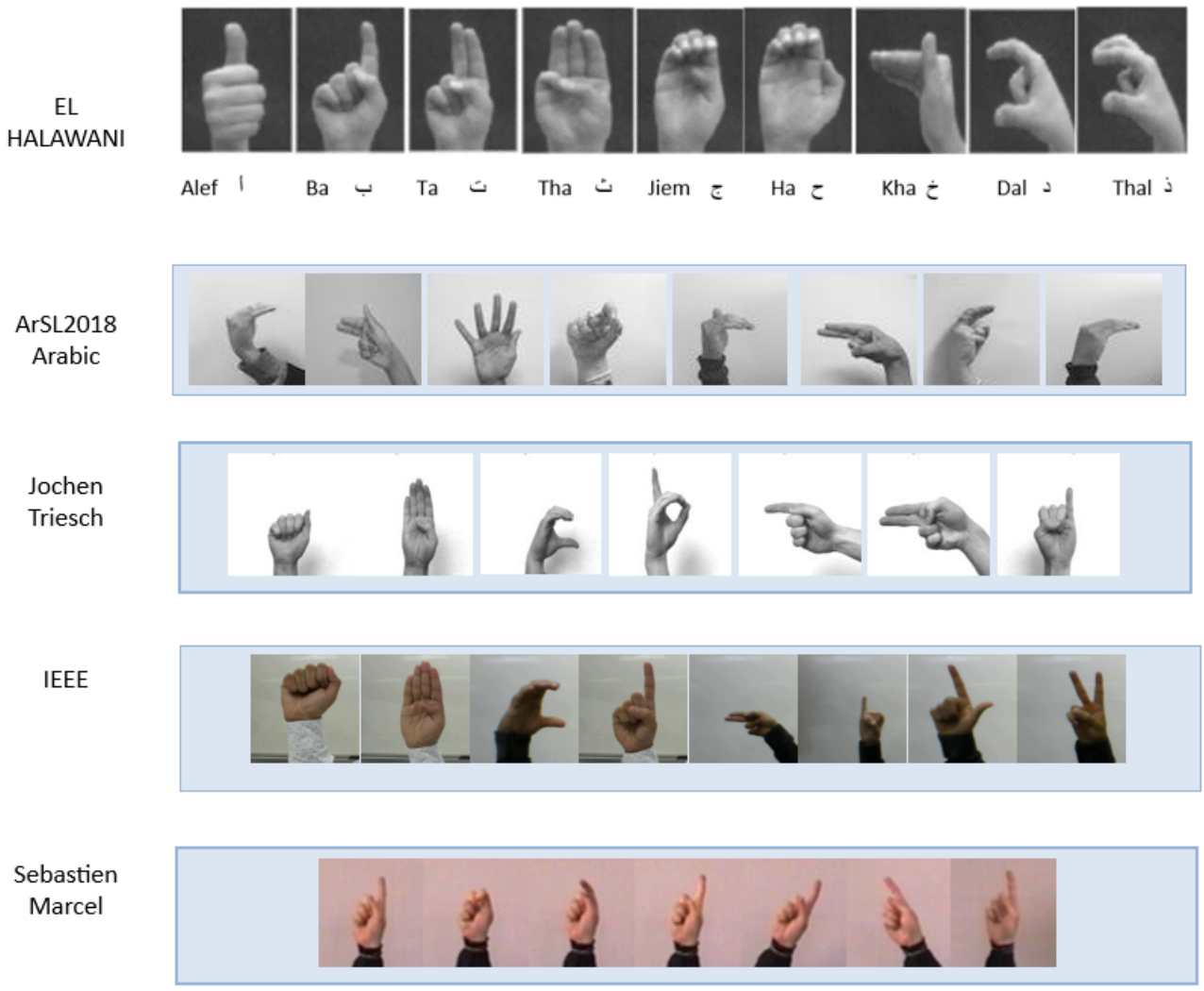


Figure 3.2: Some examples from each gesture datasets.

where for each class, 50% is assigned for training and 50% for the test, except for static hand gesture Asl dataset where 60% of the postures were used to train our classifiers and 40% for the tests. The recognition rate as represented in equation3.2 is used as a metric to evaluate the performance of the fifth architectures.

$$Accuracy = \frac{\text{Gesture correctly recognized}}{\text{Total gestures}} \times 100 \tag{3.2}$$

One against all SVM is applied for recognizing the total hand posters, where the input training consists of a matrix of size $(M * N)$; where M is the length of our feature vector used for hand characterization and N is the length of the total class vector. Each column contains the H_R histograms of LBP, LGBPHS, LDP, LOOP and M_i-LOOP, respectively.

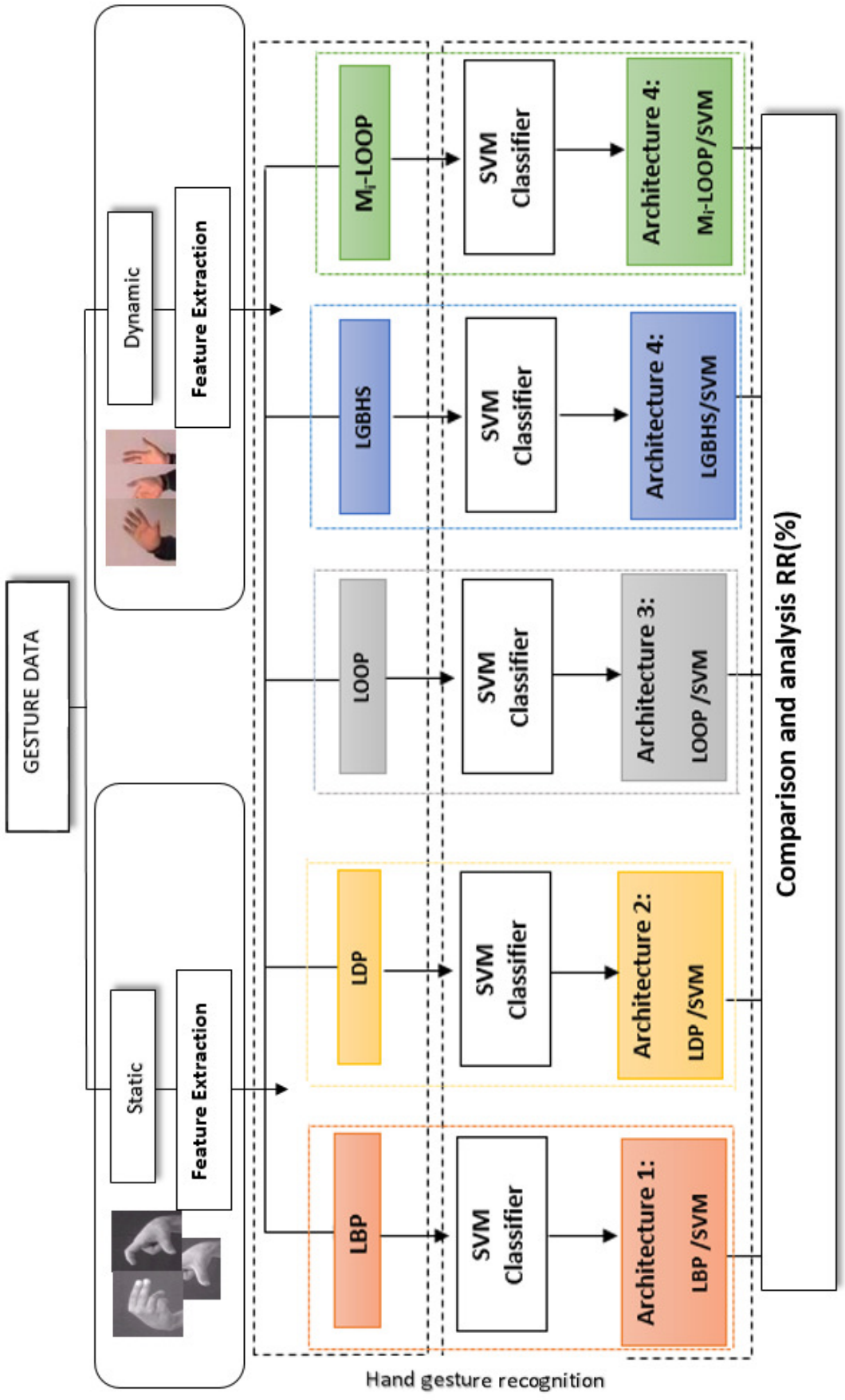


Figure 3.3: Diagram of the proposed hand and sign recognition system using SVM.

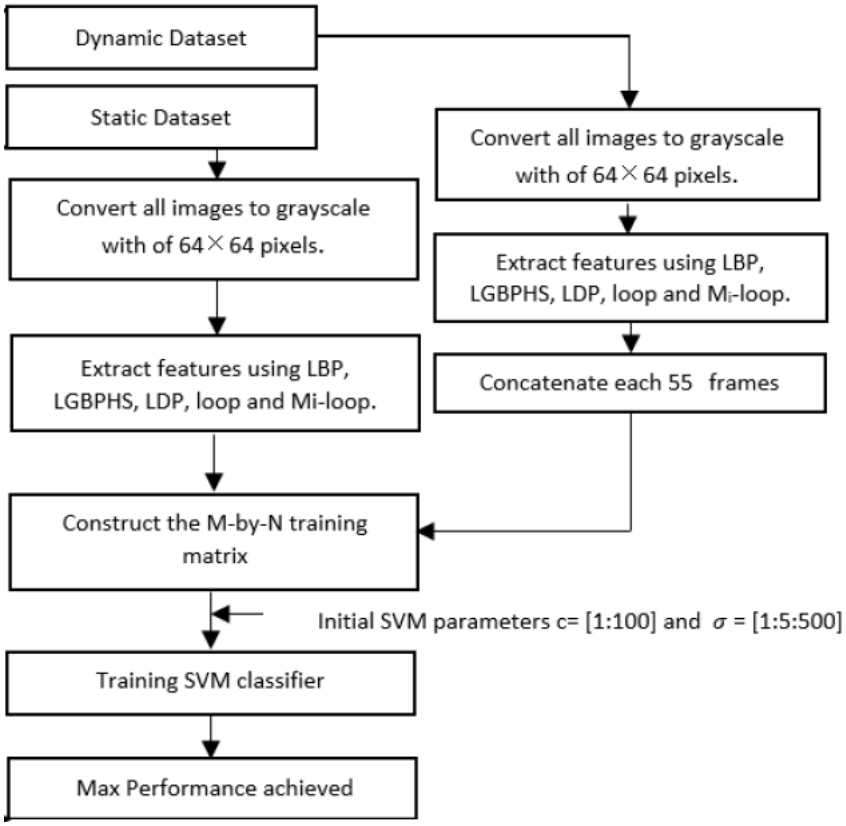


Figure 3.4: Flowchart of the proposed recognition system

SVM is a supervised learning method used for classification and regression introduced by Vapnik [121]. The basic training principle SVM is to determine the optimal linear hyperplane between two classes, such that the expected classification error for unseen training set is minimized. Suppose a training set $(x_i, y_i) \in R^M \times \{\pm 1\}$, $i = 1, \ldots, m$, x_i and y_i are the training sample and its class label, respectively [122].The m-dimensional input vector are mapped into a an inner product space via a kernel function such that: $k(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$. Then, data are classified according to

$$f(x) = \sum_{i}^{M} \alpha_i c_i k(x, x_i) + b \tag{3.3}$$

Where M is the number of support vectors, which are training data whose positive Lagrange multipliers α_i, are not zero, and $c_i \in \{-1, 1\}$ is the class label of training sample . Furthermore, several kernel functions, such as, Polynomials and Radial Basis Functions, etc, which satisfy Mercer's conditions are eligible to be a SVM kernel [123]. For pattern recognition, the Radial Basis Function (RBF) kernel provides commonly the best performances [124], which is defined as:

$$k(x, x_i) = e^{-\frac{D(x, x_i)}{2\alpha^2}} \tag{3.4}$$

α is the standard deviation of the Gaussian function and $D(x, x_i)$ generally is the Euclidean Distance.

We present different experiments carried out to evaluate the performance of the proposed hand recognition system. The flowchart of the proposed recognition system is depicted in Fig.3.4.

Since there is no rule that could be followed to find the best selection of SVM parameters, and to achieve better accuracy, each parameter was varied considering the largest meaningful range of values. The suitable values of the regularization parameter C and σ are chosen optimally by varying C in the range [1:1:100] and σ in [1:5:500].

3.4 Bimodal speech/gesture recognition system using MLP, SVM and CNN classifiers

Information fusion technology is one of the emerging technologies of data processing. Among the three levels of information fusion (data level, feature level, score and decision level) [125] , the feature level fusion, delegated by multi-data combination, has been one of the hot research fields on pattern recognition, and has achieved successful application such as authentication, medical diagnosis, speech/gesture recognition, and remote sensing.

Feature-level fusion combines features from multiple data sources or modalities into a single, unified representation. This approach exploits the complementary nature of features from different domains to improve system performance. By integrating feature vectors into a higher-dimensional space, the fused representation captures richer and more diverse information than any single modality alone.

For our proposed system, we implement feature-level fusion by concatenating the feature sets extracted from multiple information sources. Specifically, we propose a modality fusion approach that combines acoustic and visual features in a speech/gesture recognition system. This integration leverages different architectural designs to effectively merge these modalities. Figure 3.5 illustrates the various architectures used to achieve this fusion, highlighting how the combination of acoustic and visual information enhances the system's overall performance.

The overall system is mainly divided into three main parts; firstly, a dataset of 3D co-verbal gestures with Arabic audio clips collected from 18 person has been well-structured and well-annotated. Secondly, our recognition system can accommodate various acoustic and visual features as input

parameters, addressing both speech and gesture. Our analysis aims to delineate and compare the accuracy achieved through different modality fusions, specifically MFCC/DWT, ΔMFCC/DWT, $\Delta\Delta$MFCC/DWT, LPC/DWT, MFCC/CT, ΔMFCC/CT, $\Delta\Delta$MFCC/CT, LPC/CT and MFCC/M_5-LOOP, ΔMFCC/M_5-LOOP, $\Delta\Delta$MFCC/M_5-LOOP and LPC/M_5-LOOP.

Furthermore, for comparison and validation purposes, we also developed a baseline hand and speech recognition system based on SVM and MLP classifiers that is similar to the system described in Sections 3.2 and 3.3.

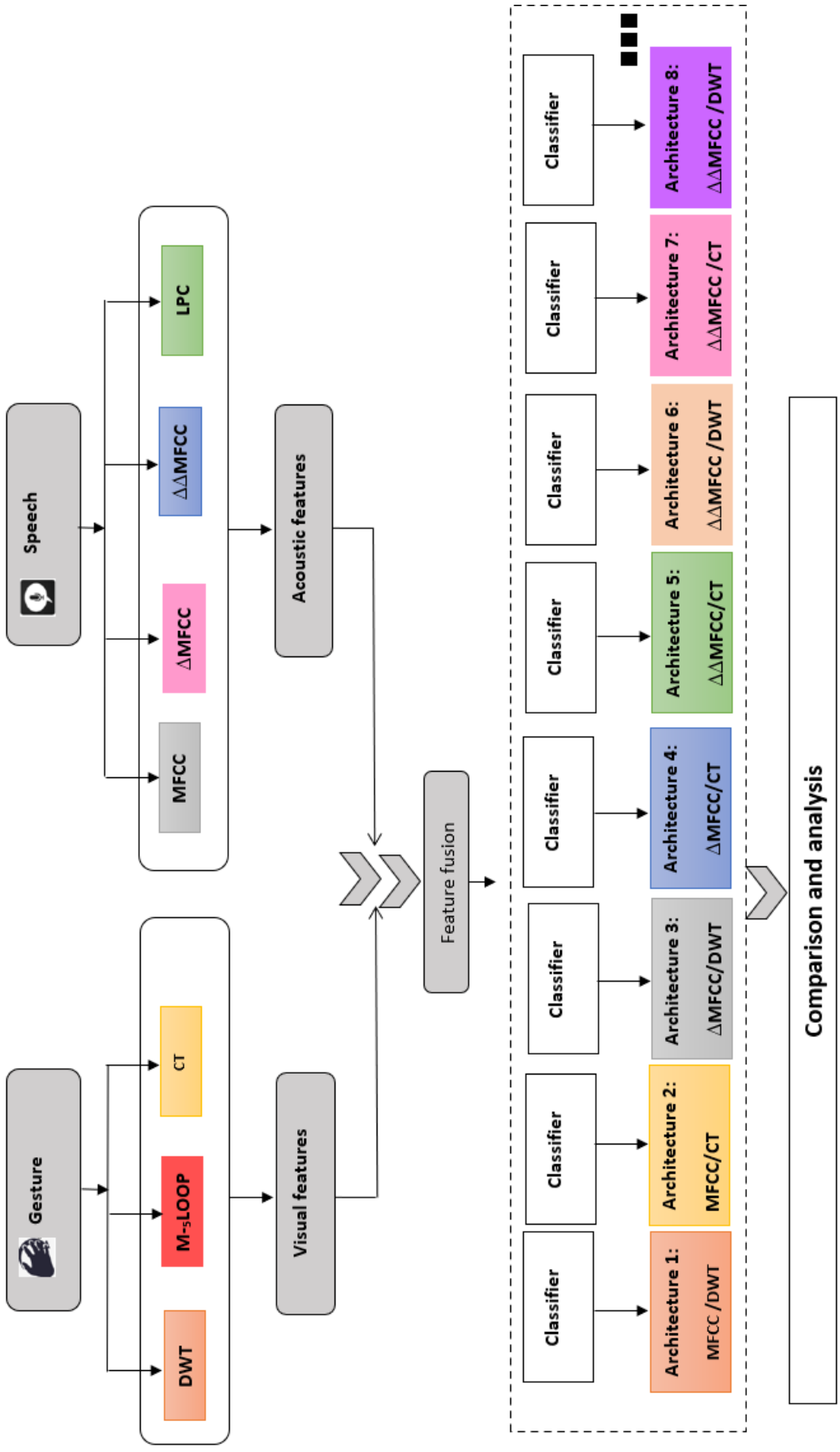


Figure 3.5: Diagram of the proposed method.

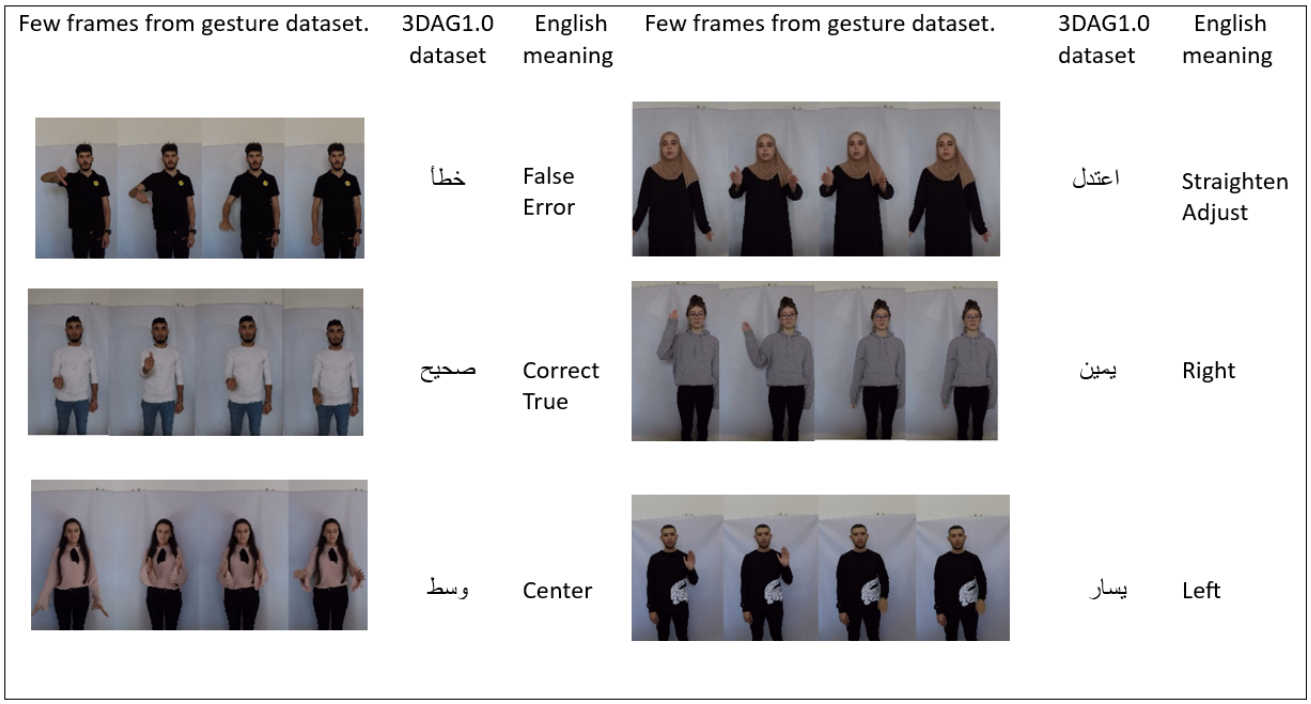


Figure 3.6: Few frames from hand gesture dataset.

3.4.1 Gesture and speech Dataset description

A comprehensive effort has been made to create a rich dataset capturing both audio and 3D gesture data related to common Algerian Sign Language (LSA) words and gestures. The first dataset consists of Arabic audio clips, named ' Isolated Word Algerian Language (IWAL1.0),' while the second contains 3D co-verbal gestures, referred to as ' Algerian Hand Gesture Dynamic Dataset AHCGDD1.0.' Both datasets are focused on the Arabic language spoken in North Africa, particularly in the Algerian region. The two datasets were recorded at the Speech Communication and Signal Processing Laboratory at the University of Science and Technology Houari Boumediene (USTHB) in Algiers, Algeria.

After several visits to the school for deaf and hearing Impaired pupils in Rouiba and attending multiple sessions, we collected Algerian sign language words and expressions data to create a gesture database. This approach ensures the authenticity of the gestures and their relevance to the ASL community. Six common and useful words were selected – "yamin" (right), "yasar" (left), "sahih" (correct), "khataa" (error), "iatadl" (adjust/balance), and "wasat" (center) – as the basis for developing an ASL gesture recognition system. The recording conditions were carefully controlled, with a uniform background, standard lighting, and a fixed distance of 2 meters between the participants and the camera. Figure 3.6 presented a few frame of each gesture.

3.4.1.1 Isolated Word Algerian Language dataset

This dataset consists of 6 classes of Arabic words spoken performed by 18 different students, all speakers repeated the sequence twenty times, resulting in a total of 2160 audio clips. The recording frequency is 16 Khz with a recording time of 2s. The sequences used in this database are as follows: يمين - يسار - اعتدل - وسط - صحيح – خطأ

3.4.1.2 Algerian Hand Gesture Dynamic Dataset

This dataset consists of 6 classes of with 3D co-verbal gestures corresponding of last sequences, collected by 18 different students, repeated twenty times. Each sequence contains 55 frames resulting in a total of 118800 images, 3D co-verbal gestures.

3.4.2 Evaluation setup

In order to evaluate the performance of the proposed modality fusion of acoustic and visual features of a speech/gesture recognition system and to assess the performance of the proposed descriptors. We carried out a set of classification experiments using the collected dataset as described in section 3.4.1 . We used standard evaluation protocol for the proposed datasets where for each class, 50% is assigned for training and 50% for the test, The recognition rate as represented in equation3.5 is used as a metric to evaluate the performance of the sixth architectures.

$$Accuracy = \frac{\text{Gesture/speech correctly recognized}}{\text{Total Gestures/speechs}} \times 100 \tag{3.5}$$

We present different experiments carried out to evaluate the performance of the proposed recognition system. The flowchart of proposed model is depicted in Fig.3.7.

One against all SVM and MLP classifiers are applied for recognizing the combined acoustic and visual feature modalities, where the input training consists of a matrix of size $(M * N)$; where M is the length of our feature vector used for hand and speech characterization and N is the length of the total class vector.

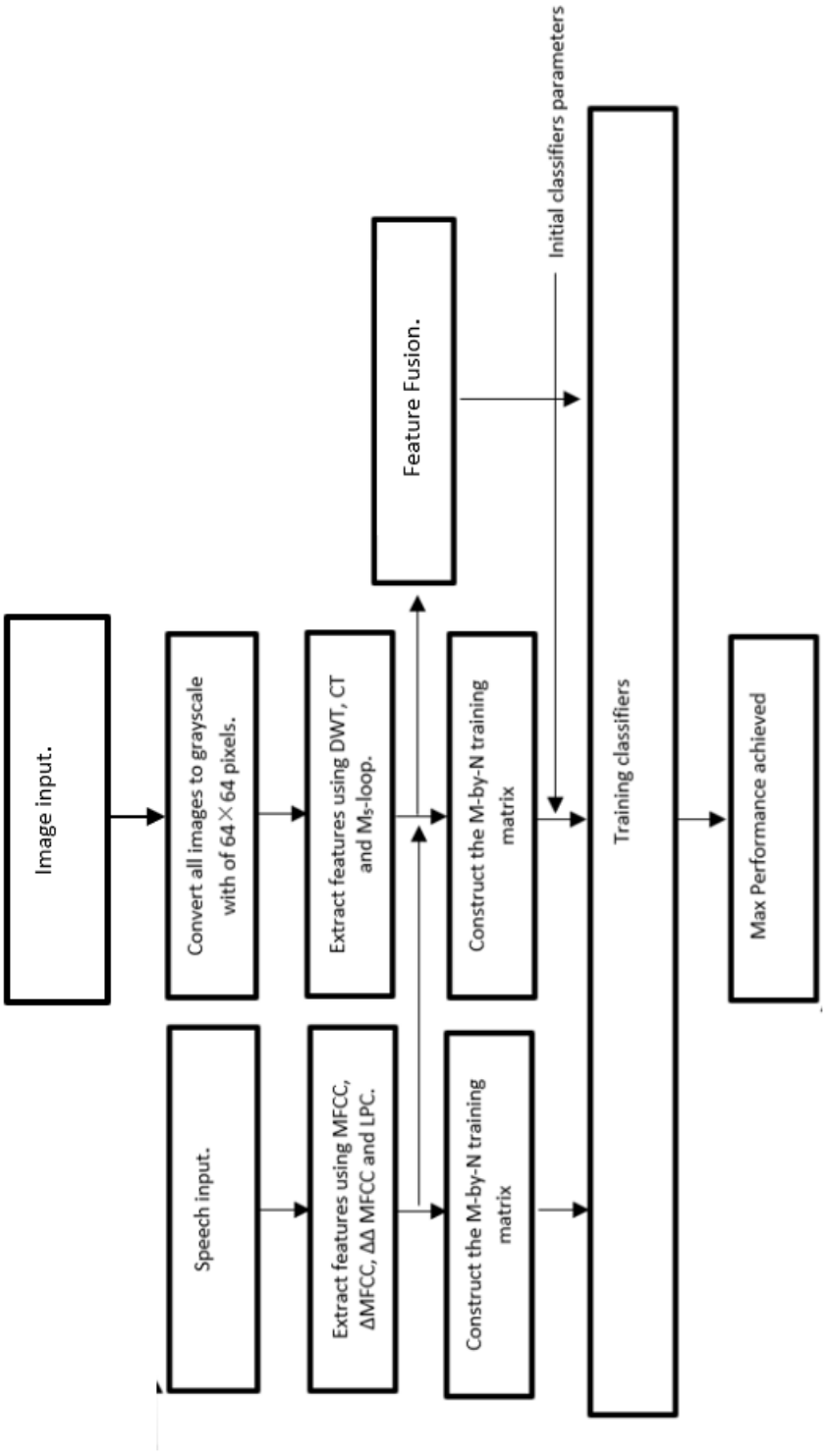


Figure 3.7: Flowchart of the proposed recognition system

3.4.2.1 Convolutional Neural Networks

Deep learning algorithms have demonstrated superior classification performance in different areas such as transfer learning, speaker and gesture recognition , etc [126] . These networks act as a black box, capable of automatically extracting essential information hidden within raw data. They function as a set of stacked layers, each analyzing details at an increasing level of complexity. Each layer deciphers the information. Ultimately, CNNs manage to synthesize these individual analyses into a global and complete representation, offering a fine understanding of audio and video data. This technology opens up an infinite field of possibilities for human-machine interaction and multimedia applications. Indeed, it enables the creation systems capable of reacting and adapting to their environment in a more natural and intuitive way. Whether it's identifying objects in a video, translating conversations in real time, or analyzing facial expressions, CNNs have the potential to revolutionize the way we interact with the digital world.

To assess performance, we examine the effectiveness of a deep learning architecture that integrates both 3D and 2D CNNs.

The principle of a CNN is to perform a convolution operation that produces filtered feature cards stacked on top of each other. A conventional neural network has the following characteristics [126]:

- A convolutional layer: This is the basic element of a CNN. The main purpose of this layer is to extract characteristics from the input features. The resulting output of the convolutional layer is given as follows:

$$C(x_{u,v}) = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j) x_{u-i,v-j} \tag{3.6}$$

  where f_k is the filter with a kernel size of $n \times m$ applied to the input x, and $n \times m$ a is the number of input connections to each CNN neuron (unit).

- A pooling layer: This layer reduces the number of features and makes the learned functions more robust by making them more invariant to changes in scale and orientation. Certain functions are used to reduce subregions, such as taking the average or maximum value. The max-pooling function given below is used by our CNN-based system [126].

$$M(x_i) = max \left\{ x_{i+k,i+l} \, |k| \leqslant \frac{m}{2}, |l| \leqslant \frac{m}{2} k, l \in N \right\} \tag{3.7}$$

where x is the input and m is the size of the filter.

- A Rectified Linear Unit (ReLU): The ReLU is an operation that replaces all negative values in the feature map with zero. The goal of ReLU is to introduce nonlinearity into our CNN-based system because most of the data we want our CNN to learn are nonlinear. Other nonlinear functions, such as tanh or sigmoid functions, can be used, but in most cases, ReLU is more efficient. Given the input x, ReLU uses the activation function $R(x) = max(0, x)$ to calculate its output [126].

- A fully connected layer: This layer takes all the neurons of the previous layer and connects them to each of its neurons. Adding a fully connected layer is a good method for learning nonlinear combinations of these features [126]. The output of this layer is given by:

$$F(x) = \sigma(W_{l \times k} * x) \tag{3.8}$$

where σ is the activation function, k is the size of the input x, and l is the number of neurons in the fully connected layer. This results in a matrix W.

- An output layer: The output layer is a one-hot vector representing the class of the given input vector. Therefore, its dimensionality is equal to the number of classes. In our work, we used 6 classes corresponding to 6 different gestures. The resulting class for output vector x is represented by:

$$C(x) = \{i | \exists i \forall j \neq i : x_j \leqslant x_j\} \tag{3.9}$$

- A softmax layer: The error is propagated back over a softmax layer. Let N be the dimension of the input vector [126]. Then, the softmax calculates a mapping such that:

$$S(x) : \mathrm{R}^{\mathrm{n}} \rightarrow [0, 1]\,; \tag{3.10}$$

For each component $1 \leq J \leq N$, the output is calculated as follows:

$$S(x) = \frac{e^{x_j}}{\sum_{i=1}^{N} e^{x_j}} \tag{3.11}$$

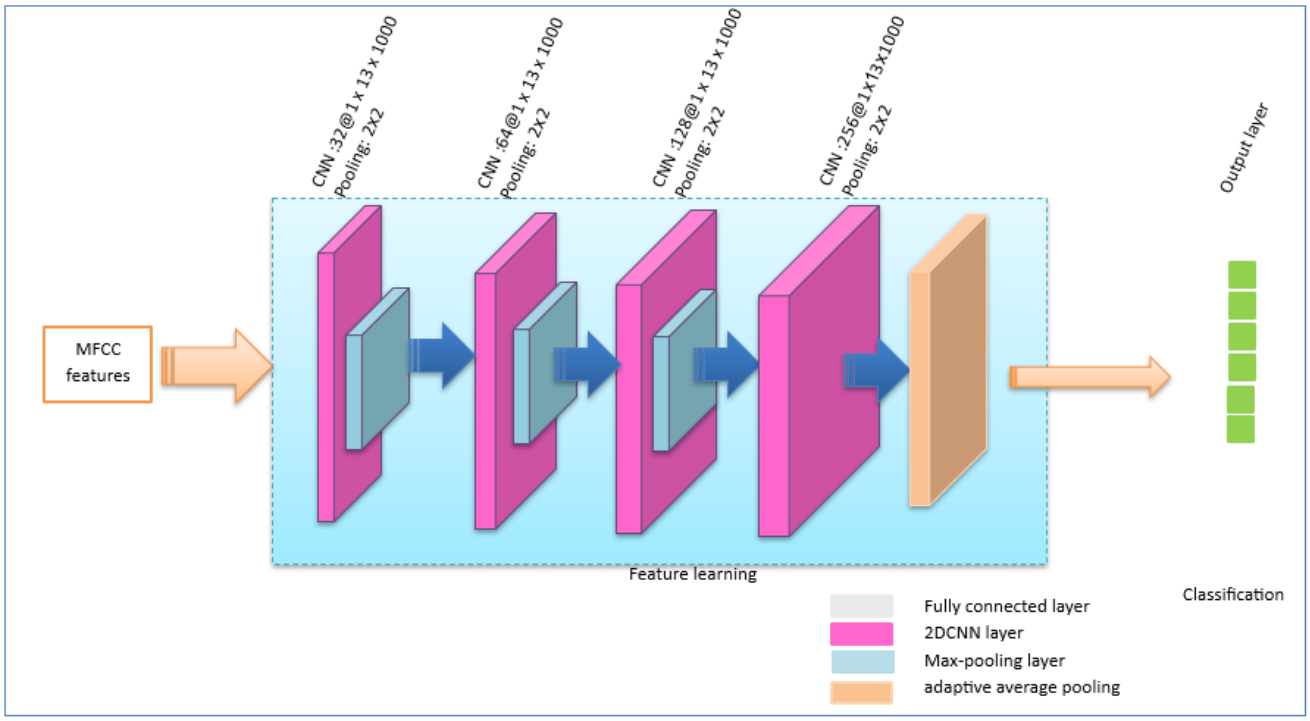


Figure 3.8: Single 2DCNN-based structure.

The output of a layer is the input of the next layer, and most of the features learned from the convolution and clustering layers may be good. However, the combinations of these features could be even better. Conventionally, a CNN consists of several iterations of this succession of layers. One of the advantages of a CNN is its relatively rapid training.

3.4.2.2 2D-CNN Implementation for Speech Identification

The first proposed and implemented system in this study is the 2D-CNN system. The input to the 2D-CNN is represented by MFCC features, where each input map is in two dimensions with 13 coefficients per frame (see figure 3.8).

The speech model employs a 2D CNN architecture to process audio input represented as Mel-frequency cepstral coefficients (MFCCs). It consists of four convolutional layers, the first producing 32 output channels, followed by a ReLU activation and a 2×2 max-pooling layer. This pattern of convolution, ReLU, and max-pooling is repeated three more times with increasing numbers of output channels (64, 128, and 256, respectively). Each max-pooling layer uses a 2×2 kernel and stride. After the final convolutional layer, an adaptive average pooling layer reduces the feature maps to a $1x \times 1$ size, effectively performing global average pooling. The output is then flattened to a 256-dimensional feature vector. Dropout layers with a rate of 0.2 are applied after each max-pooling

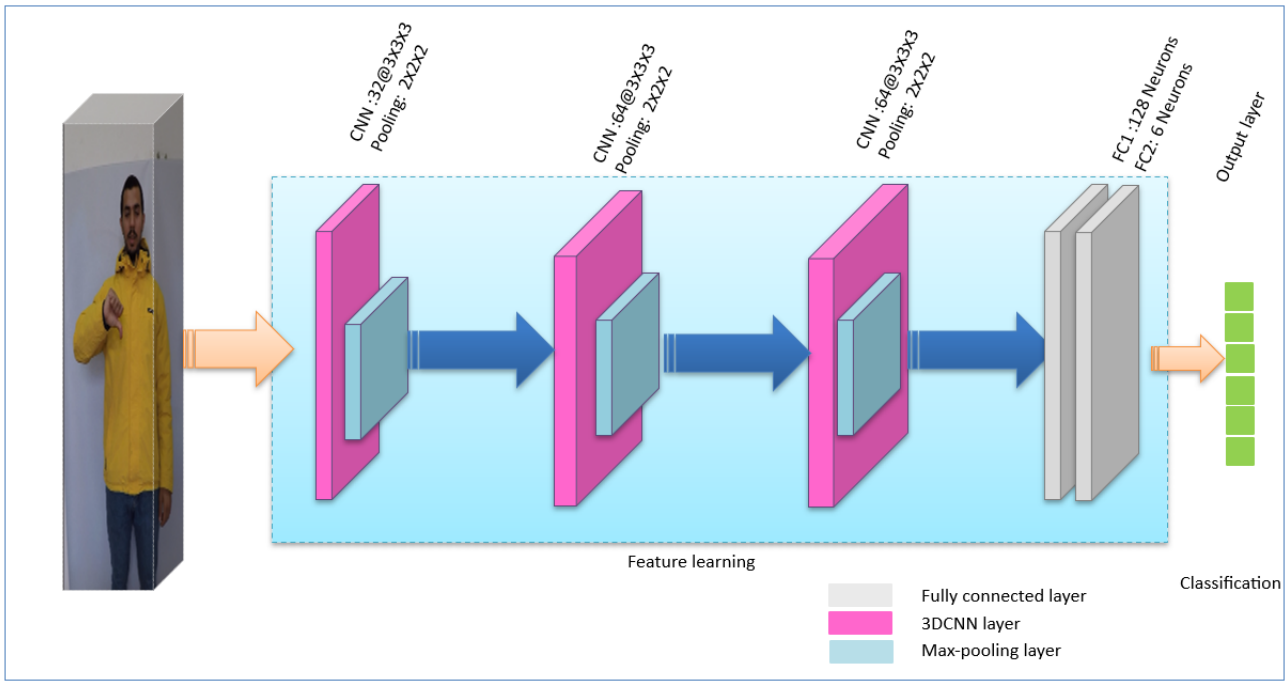


Figure 3.9: Hand gesture 3DCNN-based structure.

operation. Importantly, this model does not include any fully connected layers for classification; the 256-dimensional flattened output is passed directly to the feature fusion stage. The final classification into 6 word classes (using softmax) happens in a separate downstream component after feature fusion. Dropout is applied only after the max-pooling layers within the convolutional blocks.

3.4.2.3 3D-CNN Implementation for Dynamic gesture recognition

The second system introduced in this study leverages a 3D-CNN to tackle the problem of gesture recognition using sequential image data(see figure 3.9). In this approach, the input comprises sequences of images captured from gestures, where each image in the sequence represents a frame of the gesture. The system pre-processes these images by down-sampling and resizing them to a uniform dimension, ensuring consistency across the dataset.

The 3D-CNN model architecture is designed to capture spatial and temporal features from these image sequences, begins with three 3D convolutional layers, each employing 3x3x3 kernels. These convolutional layers are interleaved with 2x2x2 max-pooling layers to downsample the feature maps. Following the convolutional and pooling stages, the 3D feature maps are flattened into a 1D vector. This vector is then passed through a fully connected layer, reducing its dimensionality.

The convolutional layers are followed by a flattening layer that converts the 3D feature maps into

a 1D vector. This vector is then passed through a single fully connected layer.The final classification into gesture classes is performed by the last fully connected layer, which uses a softmax activation function.

The system performs feature-level fusion of 3D co-verbal gesture features and audio clips. The hand gesture model processes sequences of 3D images using multiple 3D convolutional and pooling layers to generate a 256-dimensional feature vector. Simultaneously, the audio model extracts a 256-dimensional feature vector from audio data using 2D convolutional layers, pooling operations, and adaptive average pooling. These two 256-dimensional vectors are concatenated to create a combined 512-dimensional feature vector. This combined vector is then passed through a fully connected layer to reduce its dimensionality to 256. Finally, this 256-dimensional vector serves as input to another fully connected layer for the final classification.

The final classification step of the fused model uses a fully connected layer to predict the gesture class. This layer takes the 256-dimensional fused feature vector (resulting from the concatenation and dimensionality reduction of the hand gesture and audio features) as input. It outputs a logits vector, where each element corresponds to a gesture class. A softmax function is then applied to these logits to produce a probability distribution over the possible gesture classes. The predicted gesture class is the one with the highest probability. The number of output neurons in this final fully connected layer corresponds to the number of gesture classes the model is trained to recognize.

The flowchart of the proposed recognition system is depicted in Fig 3.10.

3.5 Conclusion

In conclusion, this chapter provides an overview of the research conducted in this thesis. We have presented the architecture of the two modalities, gestural and acoustic, considered separately, as well as the fusion of the descriptor vectors for each modality.

Firstly, we described the acoustic recognition system, which is based on MFCC and LPC parameterization using MLP classifier. Secondly, we introduced the hand and sign gesture recognition system, which relies on several techniques using the SVM classifier.

Additionally, we detailed the architecture of the fusion stage, which aims to combine acoustic and gestural parameters. This fusion enables training using the MLP, SVM, and CNN classifiers.

The next chapter will focus on analyzing and discussing the results obtained for the two modal-

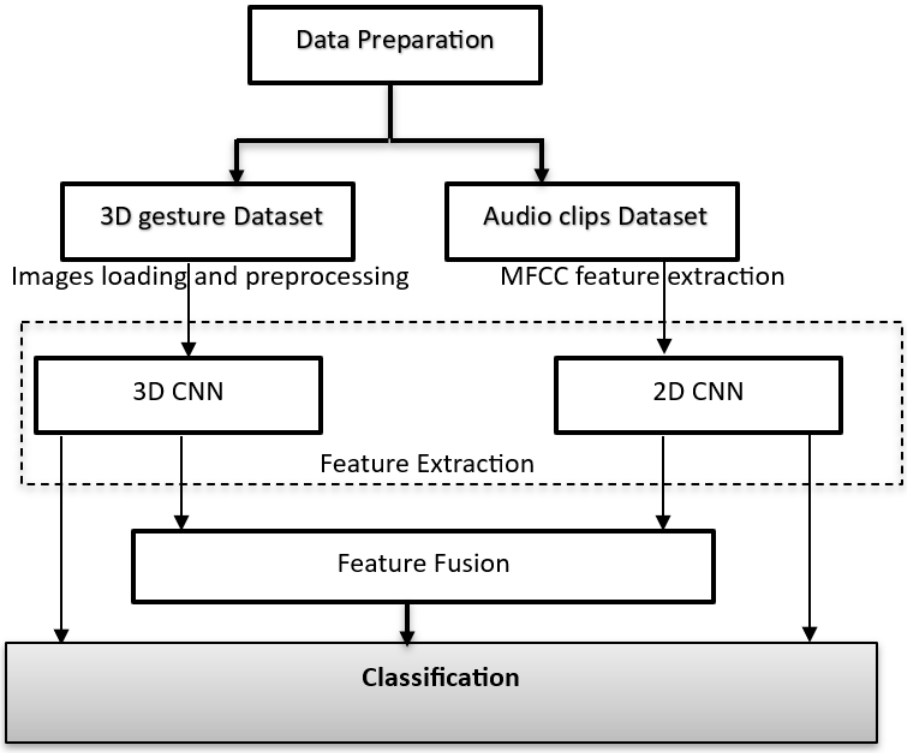


Figure 3.10: Flowchart of the proposed recognition system

ities, gestural and acoustic, both separately and in terms of the fusion of their descriptor vectors.

Chapter 4

Simulation Results and Evaluation

4.1 Introduction

To examine the effectiveness of the proposed systems, we conducted several experiments across different modalities. We analyzed the results for gestural and acoustic modalities separately, as well as for the fusion of their descriptor vectors.

Firstly, we present the results of the acoustic recognition system, which uses LPC and MFCC as will as first and second derivative descriptors parameterization with an MLP classifier. Comparisons between these parameterization techniques are also provided.

Additionally, we present and discuss the results of the hand and sign recognition system, which employs several techniques with an SVM classifier.Finally, we present the results of the fusion stage, combining acoustic and gestural parameters for training with MLP, SVM, and CNN classifiers.

4.2 Performance Analysis of the first proposed system

Evaluation and comparison experiments were conducted using the two datasets described in Chapter 3. We used MFCC, MFCC-PCA, LPC, and LPC-PCA as feature parameters, which were then classified using an MLP classifier.

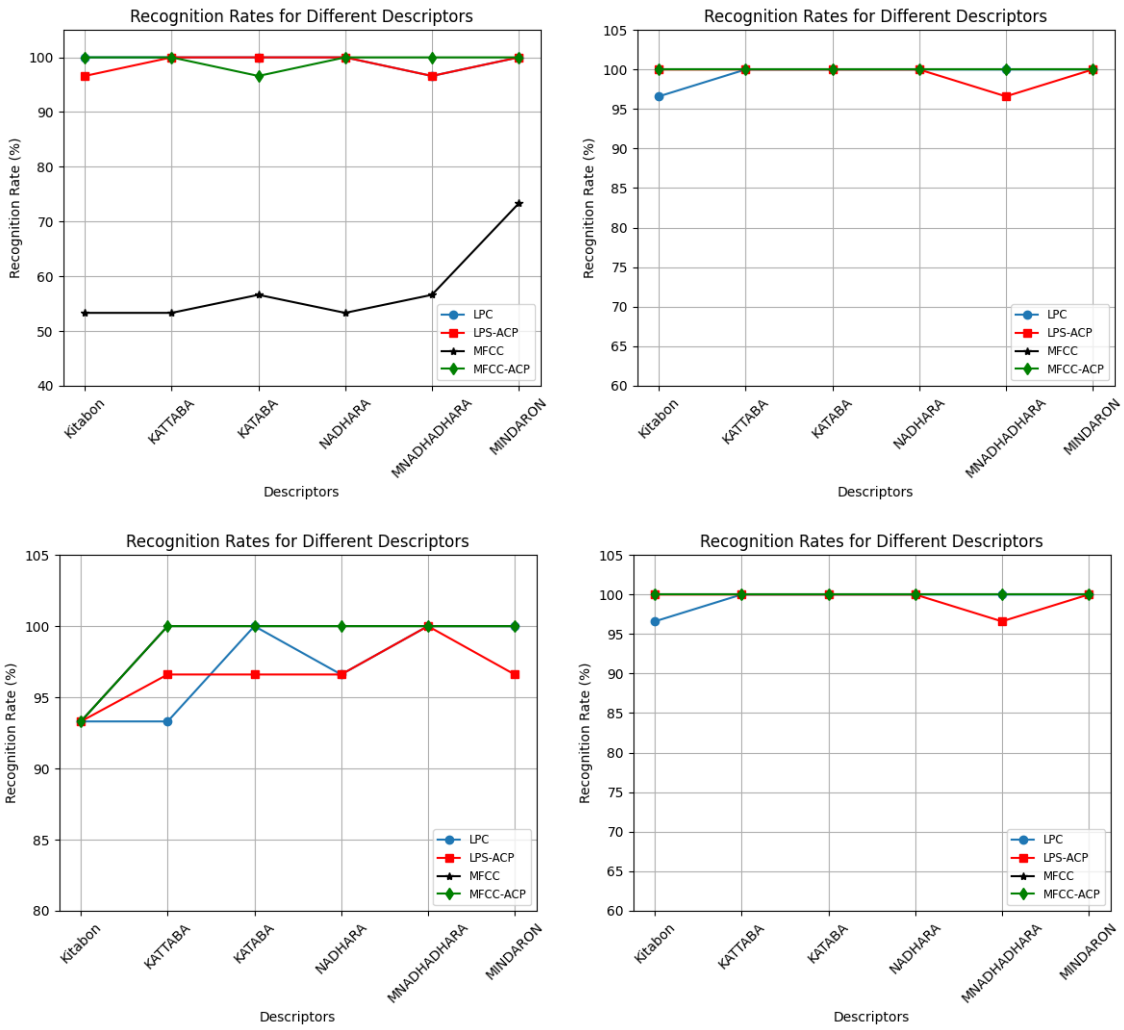


Figure 4.1: Global speaker recognition for Arabic data set: (a) using traingda function, (b) using trainrp function, (c) using trainoss function, (d) using trainscg function.

4.2.1 Results for speaker recognition rate depending on isolated words using different training functions

To identify the optimal MLP configuration, several processes were implemented. These involved determining the ideal number of hidden layers and selecting appropriate training algorithms, as detailed in Chapter 3, Section 1.

Figure 4.1 shows the overall accuracy obtained for the Arabic dataset using four descriptors (LPC, LPC-PCA, MFCC, and MFCC-PCA) with the training functions traingda, trainrp, trainoss, and trainscg.

It can be observed that the four architectures achieved good recognition rates for almost all training functions. However, for the MFCC architecture in the case of the trainoss and trainrp

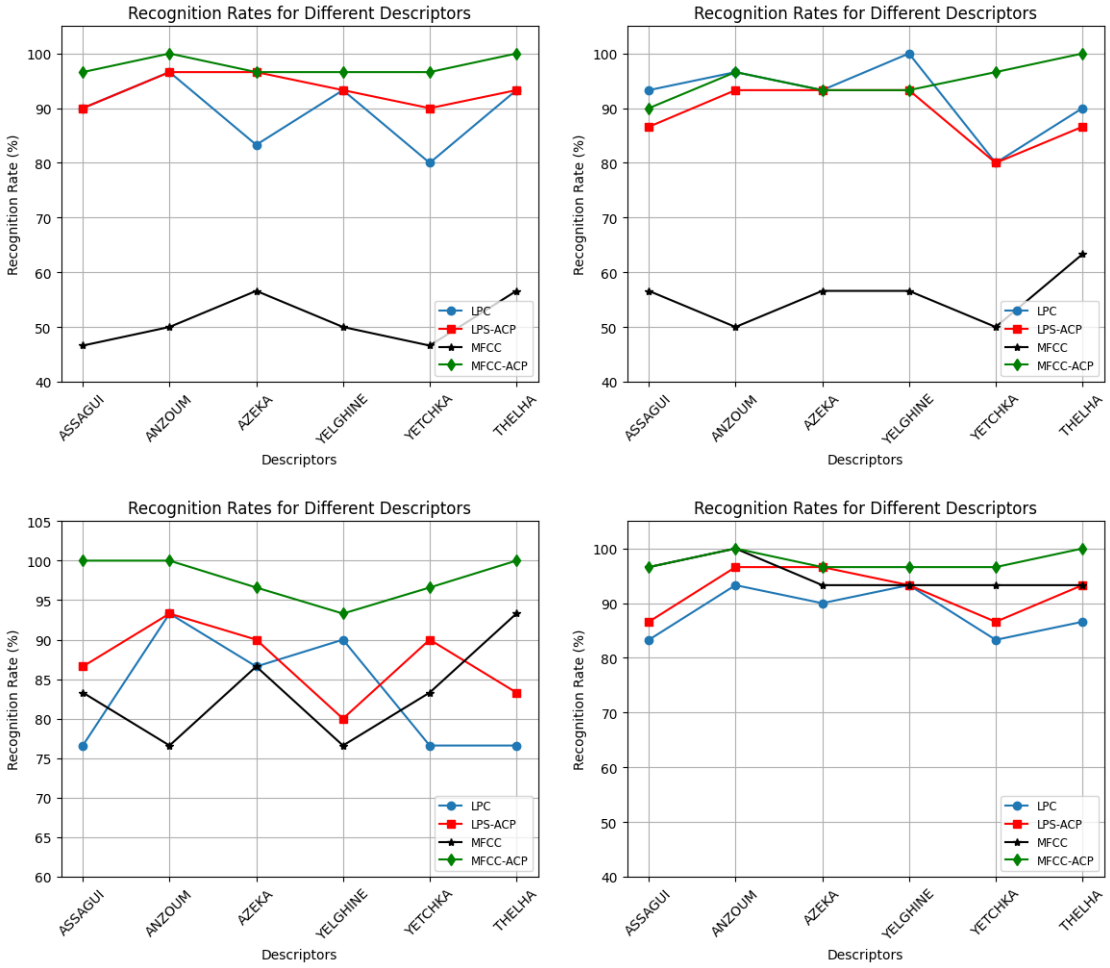


Figure 4.2: Global speaker recognition for Berber data set: (a) using traingda function, (b) using trainrp function, (c) using trainoss function, (d) using trainscg function.

functions, the results were not satisfactory. These results indicate that the chosen training algorithm during ANN training significantly influences the ANN predictions.

4.2.2 Results for speaker recognition rate depending on continuous speech using different training function

The system's efficacy in processing continuous speech was assessed using the trainrp, traingda, trainoss, and trainsgc training functions. In Figure 4.2 , a comparison of the classification results from the proposed MLP classifier is illustrated for the four descriptors (LPC, LPC-PCA, MFCC, and MFCC-PCA) employing these training functions.

The obtained results indicate a notable recognition rate for the extensively analyzed sentences, with the most favorable outcomes achieved through the trainsgc and traingda functions. Conversely,

Table 4.1: Recognition rates obtained over the two datasets

Descriptor	LPC-MLP	LPC-ACP-MLP	MFCC-MLP	MFCC-ACP-MLP
RR% using Trainsgc	**83 to 100**%	86 to 100%	**96 to 100**%	93 to 100%
RR% using Traingda	76 to 100%	83 to 100%	76 to 100%	93 to 100%
RR% using Trainrp	80 to 100%	80 to 100%	36 to 93%	90 to 100%
RR% using Trainoss	83 to 100%	**93 to 100**%	36 to 73%	**96 to 100**%

the use of trainrp and trainoss functions yielded unsatisfactory global recognition rates for the MFCC-MLP architecture. Importantly, the efficiency observed for continuous speech is comparable to that for isolated words.

4.2.3 Discussion

The system's performance is summarized in Table 4.1 , revealing that the incorporation of MFCC and LPC, followed by PCA, significantly enhances the identification system's performance for the extensively studied datasets.

The system's performance exhibits a considerable range, spanning from 36% to 100%.This variability is attributable to several factors, including speaker emotional and physical state, recording conditions, speech rate, and background noise. These diverse elements underscore the complexity of achieving consistent performance.

Moreover, the critical role of feature extraction and training functions in shaping system performance cannot be overstated. The trainrp function, acknowledged for its swiftness in handling pattern recognition problems, stands out for its efficiency and showcases relatively lower memory requirements compared to alternative algorithms.

Conjugate gradient algorithms, with a special mention to trainscg, demonstrate notable effectiveness across a spectrum of problems, especially in networks characterized by a multitude of weights. Their adaptability contributes to robust performance under varying conditions.

The SCG algorithm emerges as a noteworthy contender, nearly matching the speed of trainrp in pattern recognition tasks while maintaining relatively moderate memory requirements. This aspect is particularly significant in resource-conscious applications.

In summary, the system's performance is intricately tied to a range of factors, and the strategic selection of feature extraction and training functions significantly influences its robustness and efficiency across diverse scenarios.

4.3 Performance Analysis of the second proposed system

To showcase the efficacy of the proposed textural features, we performed experiments on five distinct datasets, comprising two ArSL (Arabic Sign Language) datasets named El-halawani and ArSL2008, two ASL (American Sign Language) datasets named Triesh and IEEE, and one dynamic dataset known as the Marcel dataset. The following paragraphs present the simulation results obtained for each dataset for the total descriptors.

4.3.1 Performance analysis on El Halawani ArSL dataset using SVM classifier

The Arabic sign gesture recognition datasets are constructed by choosing the total alphabets. Comparison with the traditional LBP, LDP and LOOP methods, presented in fig.4.3 is performed to demonstrate the performance of the proposed system. The results for the global recognition rate system varied from 80.7 to 90.7%. From this analysis, it is obvious that the proposed M_i-LOOP descriptors improves the discrimination capabilities of the traditional LBP, LDP, LOOP and LGBPHS, in almost all cases.

It is clearly observed that the M_i-LOOP achieves 90.7% accuracy when applied to the Halawani dataset. This dataset presents significant variability due to acquisition conditions, lighting variations, and other factors.

4.3.2 Performance analysis on ArSL2018 ArSL dataset using SVM classifier.

Figure 4.4 reports the accuracy obtained for ArSL2018 dataset. Our system achieves between 95.9–96.8% recognition rate in identifying the correct gesture while analyzing the total gestures. By comparing the most efficient descriptors it is clearly noticeable that the result obtained using M_2-LOOP are better compared to others, where 96.8% of accuracy is obtained.

4.3.3 Performance analysis on Jochen Triesch blue ASL dataset using SVM classifier.

Simulation results on American dataset with white and black background is provided in Fig.4.5. Experiments are conducted to assess the performance of the total descriptors. As we can see for white

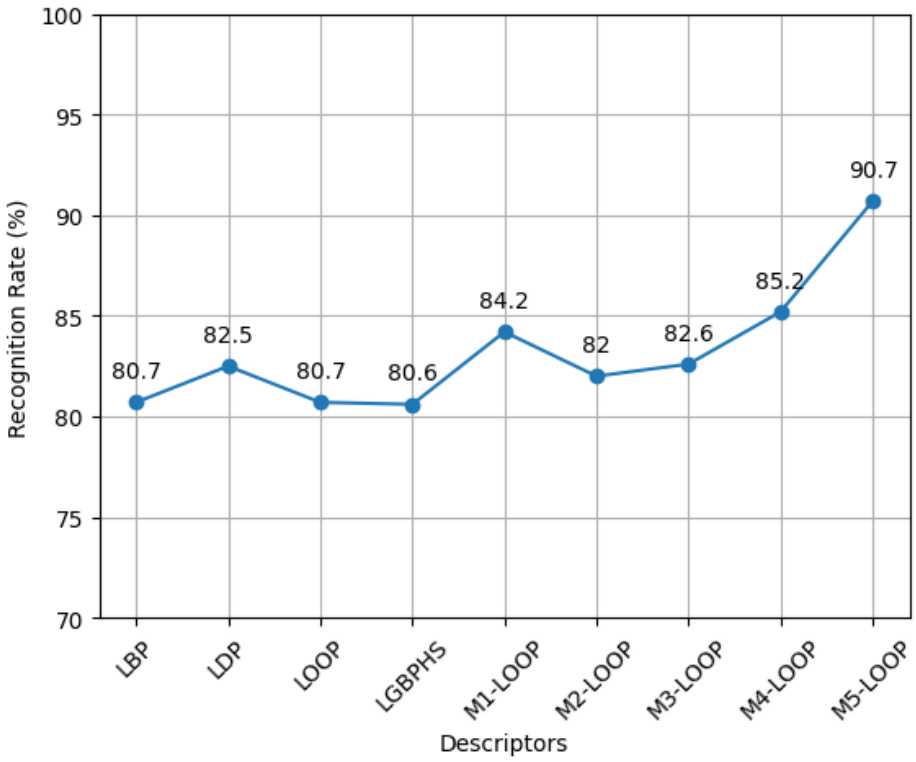


Figure 4.3: Global hand recognition for El Halawani ArSL dataset using SVM classifier

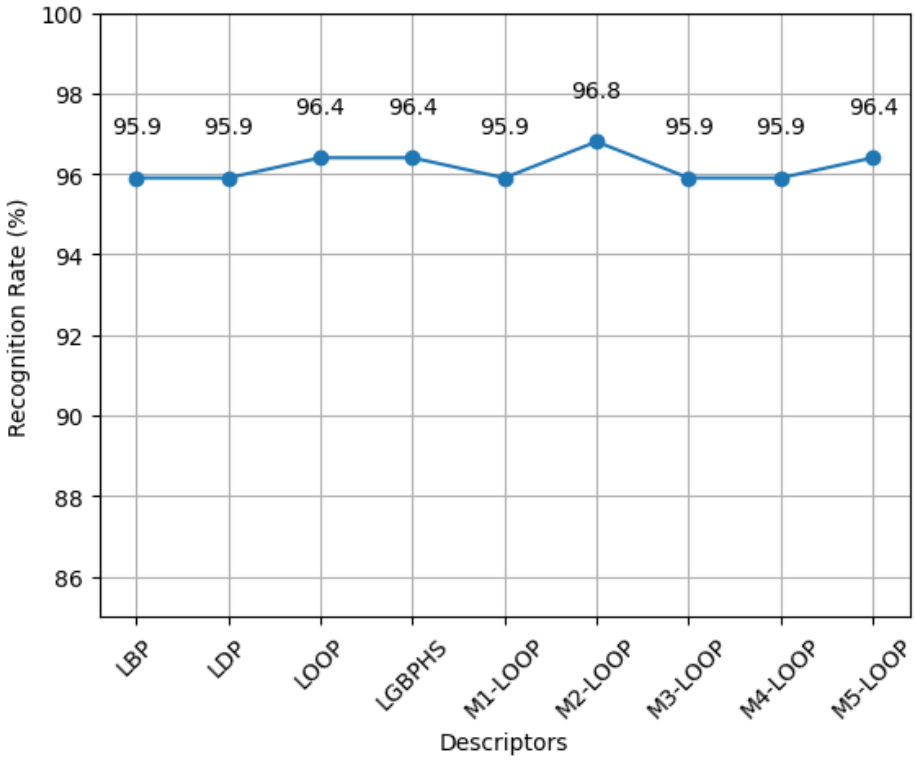


Figure 4.4: Global hand recognition for ArSL2018 ArSL dataset using SVM classifier.

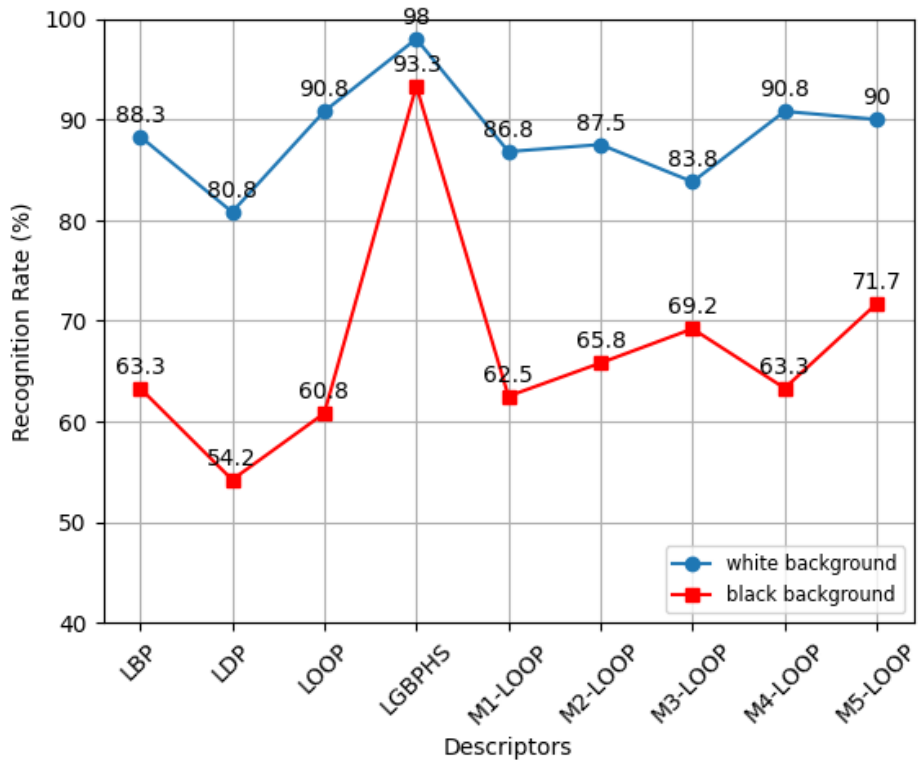


Figure 4.5: Global hand recognition for American data set for white and black background using SVM classifier

background, the nine descriptors allow obtaining good performances with an accuracy varying from 80.8% to 98%. The proposed M_i–LOOP descriptor presents a significant high performance over the LOOP descriptor. Better performances are obtained using LGBPHS, M_4–LOOP and LOOP. By analysis the result, it is clear that the use of the multi-scale and multi-orientation blue Gabor filter in LGBPHS descriptor permits a significant parameterization of the dataset where 98% was obtained.

Furthermore, we clearly see that LGBPHS achieved good performance over the rest descriptors on black background ASL. We note that the result obtained with LBP, LDP, LOOP and M_i-LOOP descriptors achieve less efficiency compared to results obtained for the white background. Another interesting observation is that the fifth architecture offers best results as compared with the first, second and the third one. Besides that, the LGBPH can be considered as the best feature describing the Triesh datatset with white and black background. An improvement of 7.3% was noticed for the white background using the LGBPHS descriptor.

Figure 4.6 summarizes results obtained using the static hand gesture Asl dataset (IEEE dataset). The results prove the effectiveness of our recognition system. The overall accuracy of identifying the total gestures varied from 88% to 96.4%. The best score of our system about 95.4%. The fifth architecture called Mi-LOOP/SVM offers best results, it achieved a recognition rate of 94.2%. We

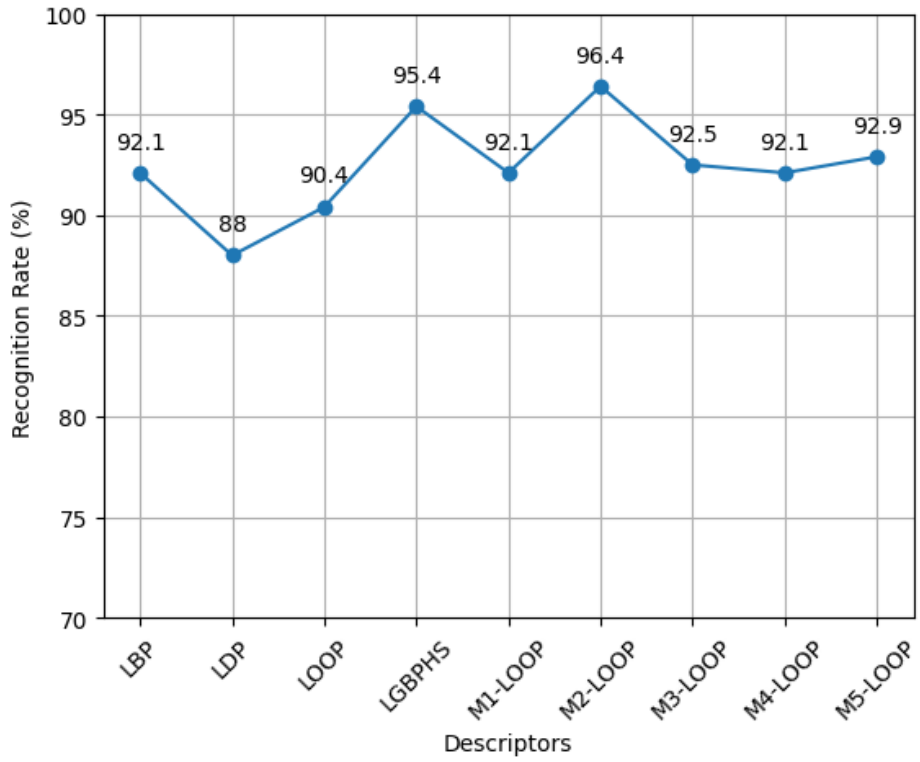


Figure 4.6: Global hand recognition for Static Hand Gesture Asl Dataset using SVM classifier

notice that the LGBPHS descriptor achieve more efficiency compared to LBP, LDP, LOOP and M_i-LOOP descriptors.

4.3.4 Performance analysis on Dynamic dataset (Sebastien Marcel) using SVM classifier

Figure 4.7 shows the best performance obtained for features generated from the dynamic database by using the nine descriptors. This database contains four different gestures (Clic, No, Rotate, Stop), each gesture is represented by 12 video sequences, and each sequence includes 55 frames performed by a different person. All images are converted from RGB to grayscale, a total of 55×12×4=2640 images was used to evaluate our system. The overall recognition rate increase in the range of [75 , 100], it can be clearly inferred that the proposed M_5–LOOP descriptor achieves the best performance comparing to others, it reaches 100%.

For better readability, the performance of the proposed descriptor M_i-LOOP, as well as the traditional methods LBP, LDP, LGBPHS and LOOP, on the five datasets are presented in Table 4.2. It can be drawn from the global analysis of this table that:

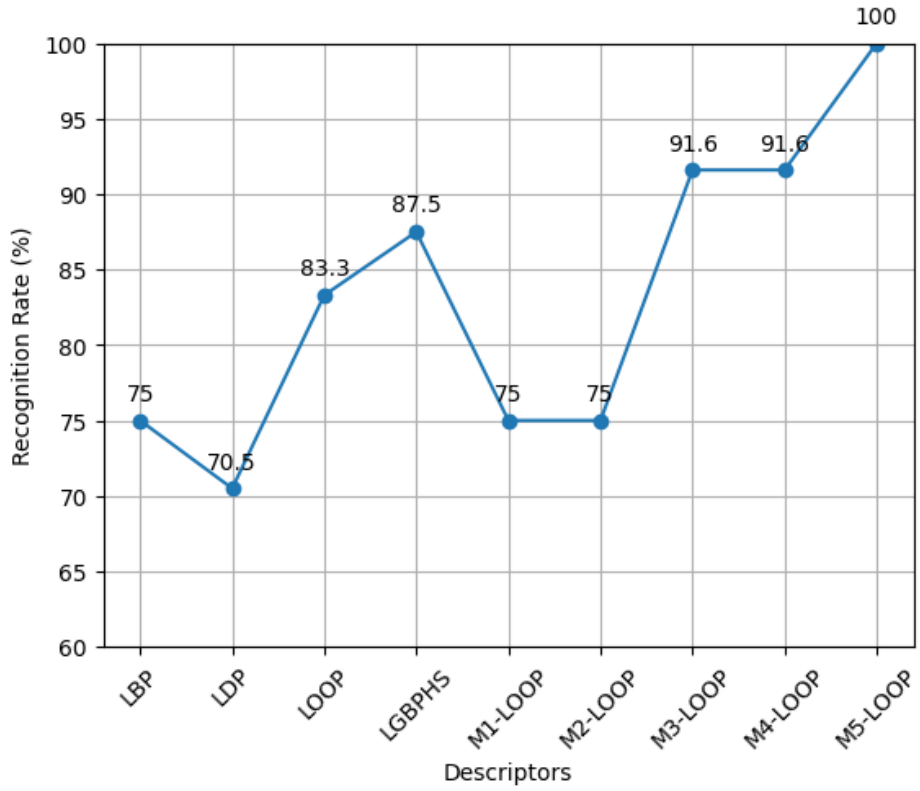


Figure 4.7: Global hand recognition for Dynamic dataset (Sebastien Marcel)

- All descriptors get promising results on five tested datasets namely, Triesch dataset with white background, El Halawani, Arsl 2018, Asl and Sebastien Marcel dataset. The best efficiency of this descriptors varies from 93,3% to 100%.

- It can be observed that there is a significant performance drop for the tested approaches on Triesh dataset black background. This result can be justified by the use of multi-scale and multi-orientation Gabor filter in LGBPHS descriptor, which permits a significant parametrization of the white and black dataset.

- It is obvious that the proposed M_i-LOOP descriptors improves the discrimination capabilities of the traditional LBP, LDP and LOOP, in all cases.

- Considering the ranking between the tested methods within each dataset, the best method in terms of accuracy is the proposed M_i-LOOP the second best method is LGBPHS which is in the top 2 descriptors on 3 datasets.

- Good performance was obtained for dynamic “Marcel Sebastien” dataset using Modified LOOP descriptor.

Table 4.2: Recognition rates obtained over the five datasets using SVM classifier

El Halawani		ArSL2018		Jochen Triesch white background		Jochen Triesch black background		Asl		Sebastien Marcel	
M5-LOOP	90.7%	**M2-LOOP**	96.8%	**LGBPHS**	98%	**LGBPHS**	93.3%	**M2-LOOP**	96.4%	**M5-LOOP**	100%
M4-LOOP	85.2%	M5-LOOP	96.4%	M4-LOOP	90.8%	M5-LOOP	71.7%	LGBPHS	95.4%	M4-LOOP	91.6%
M1-LOOP	84.2%	LGBPHS	96.4%	LOOP	90.8%	M3-LOOP	69.2%	M5-LOOP	92.9%	M3-LOOP	91.6%
M3-LOOP	82.6%	LOOP	96.4%	M5-LOOP	90 %	M2-LOOP	65.8%	M3-LOOP	92.5%	LGBPHS	87.5%
LDP	82.5%	M1-LOOP	95.9%	LBP	88.3%	M4-LOOP	63.3%	M1-LOOP	92.1%	LOOP	83.3%
M2-LOOP	82%	M3-LOOP	95.9%	M2-LOOP	87.5%	LBP	63.3%	M4-LOOP	92.1%	M2-LOOP	75%
LBP	80.7%	M2-LOOP	95.9%	M1-LOOP	86.6%	M1-LOOP	62.5%	LBP	92.1%	M1-LOOP	75%
LOOP	80.7%	LBP	95.9%	M3-LOOP	83.8%	LOOP	60.8%	LOOP	90.4%	LBP	75%
LGBPHS	80.6%	LDP	95.9%	LDP	80.8%	LDP	54.2%	LDP	88%	LDP	70.8%

4.3.5 Discussion and comparison with state-of-the-art methods

Our experimental results demonstrate that the proposed method can extract the gesture characteristics accurately. In summary, there are many advantages for the proposed Mi-LOOP hand parametrization method. The following findings can be drawn from the analysis of the performance evaluation.

- In terms of accuracy, it can be inferred that the proposed M_i-LOOP descriptor achieves performance that is significantly superior to LBP, LDP and LOOP. Furthermore, proposed-1 and proposed-2 are the best of the five masks on all the tested datasets.This result can be justified by the combination of 2nd derivative- Gaussian mask and the proposed M_i masks, which provides more detailed and discriminated information.

- In addition to its simplicity in terms of implementation and its effectiveness, the proposed M_i-LOOP can identify almost all hand gestures under different conditions.

- In terms of computational cost, the latency of recognizing gestures has been measured by the following expression:

$$T_T = T_{DES} + T_{SVMtest}, \tag{4.1}$$

  where T_{DES} is the latency of the descriptor computation and $T_{SVMtest}$ is the time taken to classify the total gestures.

  Table 4.3 shows the processing time taken by our system to extract and then classify the relevant features. s shown in table 4.3, we noticed that the nine descriptors, including the proposed method, require less processing time, except for the LGBPHS descriptor. Indeed, LGBPHS has high computational complexity due to many scales and orientations.

The M_i-LOOP descriptor seems to be the most appropriate in term of time processing. In light of these findings, the proposed M_i-LOOP descriptor as outstanding performance in accuracy and computational cost compared to the evaluated descriptors on all the tested datasets.

In order to better demonstrate the advantages of the proposed system, a comparative analysis is performed with state-of-the-art methods [45, 46, 47, 48, 49, 50, 51, 69, 127, 128, 129, 130, 131, 132, 133, 134, 135]. Table 4.4 reports comparison results blue for recognition rate (RR). From these results, it can be seen that the proposed descriptor outperforms the results obtained in almost

Table 4.3: Processing time of the nine descriptors

Descriptor	$\mathbf{T}_T$ (S)	$\mathbf{T}_{DES}$ (S)
LBP	41.08	**32.65**
LDP	63.80	57.35
LOOP	61.90	54.45
LGBPHS	1734.57	1730.85
M_1-LOOP	59.17	51.71
M_2-LOOP	57.54	50.92
M_3-LOOP	70.04	62.95
M_4-LOOP	63.56	55.47
M_5-LOOP	53.84	**50.73**

all of the methods [127, 128, 129, 130, 131, 132, 133, 134] and also remains competitive with [69]. Moreover, the major advantage of our system over the one designed in [69] is that it is able to achieve good results without using any hand segmentation algorithm. In addition, the approach used in [69] is computationally expensive compared to ours. Indeed, for real-time applications it is suitable to increase computational speed and reduce power consumption to improve efficiency.

It can be inferred that the proposed system achieves a performance that is competitive with or better than all the tested CNN-based methods. This can be justified by the fact that the CNN classifier requires a huge amount of data to perform well, which is lacking in the existing datasets in the field.

From our analysis, we can conclude that there is no universal parameterization technique in machine learning capable of delivering excellent results across all datasets.

4.4 Performance Analysis of the third proposed system

This section presents the results of our bimodal system. Our goal is to explore feature fusion techniques from the two recognition modules to improve recognition rates, rather than analyzing each modality independently. We present results for gestural and acoustic modalities separately, as well as for the fusion of their descriptor vectors.

Firstly, we describe the acoustic recognition system, which uses MFCC, ΔMFCC, $\Delta\Delta$MFCC, and LPC parameterization with MLP and SVM classifiers. A comparison of these parameterization techniques is provided. Further more, we describe the gesture recognition system, applying MLP and SVM classifiers to CT, DWT, and M5-LOOP coefficients. Results are presented for each stage individually, and for the fusion stage, which combines acoustic and gestural parameters using MLP

Table 4.4: Comparison with previous works

References	Dataset	Method	RR(%)
[127]	Triesch dataset	Krawtchouk features till 3rd order	84.9
[128]	Triesch dataset (black background)	CNN	88.5
[129]	Triesch dataset	DWT + Fratio	89.7
[45]	Triesch dataset	CNN-LSTM	97.8
[46]	Triesch dataset	DC-CNN	98
[47]	Triesch dataset	FiST CNN	94.9
[130]	Triesch dataset (white background)	Elastic graph matching	94.3
	Triesch dataset (black background)	Elastic graph matching	93.3
Our system	Triesch dataset (white background)	LGBPHS	**98**
	Triesch dataset (black background)	LGBPHS	**93.3**
[131]	ArSL	LBP	90.4
[69]	ArSL	ANFIS Network	93.5
Our system	ArSL	M_5-LOOP	**90.7**
[48]	Sebastien Marcel	CNN-2 model	96.4
[49]	Sebastien Marcel	CNN	95.9
[50]	Sebastien Marcel	CNN + MobileNet	94.4
Our system	Sebastien Marcel	LGBPHS	**96.4**
[132]	Sebastien Marcel	3DHOG + 3DHOOF	91.6
[133]	Sebastien Marcel	Contour-Based Similarity Images	85.4
[134]	Sebastien Marcel	EHD	95.8
[135]	Sebastien Marcel	mWLD	100
[51]	Sebastien Marcel	DPCNN	96.8
Our system	Sebastien Marcel	M_5-LOOP	**100**

and SVM classifiers.

As previously noted, the overall system comprises three main parts: (1) the creation of a well-structured, annotated dataset of 3D co-verbal gestures and corresponding Arabic audio clips from 18 participants; (2) the extraction of acoustic and visual features MFCC, ΔMFCC, ΔΔMFCC, and LPC from audio; CT, DWT, and M5-LOOP from 3D gestures) used as input for our recognition system; and (3) the implementation of various feature learning and classification architectures (MLP and SVM classifiers).

4.4.1 Performance analysis on the Arabic audio clips using MLP and SVM classifier

For the speech recognition system, we compared four types of features (see Chapter 2) as inputs to the SVM and MLP classifiers. Several experiments were conducted to perform better accuracy while varying the parameters of the MLP and SVM classifiers. Figure 4.8 compares the four types of features, namely MFCC, ΔMFCC, ΔΔMFCC, and LPC coefficients. According to the obtained results, the best recognition rate was 91.6%, achieved with the second derivatives ΔΔMFCC using the MLP classifier. It can be observed that there is a significant performance drop for the tested approaches when using ΔΔMFCCs. This result can be explained by the fact that ΔΔMFCCs provide a more detailed and dynamic description of the audio signal, capturing both the static spectral features and the temporal variations, which leads to a more complex parametrization of the dataset. Additionally,It is obvious that the MLP classifier achieves higher efficiency than the SVM in almost all cases.

4.4.2 Performance analysis on the 3D co-verbal gestures using MLP and SVM classifier.

For the hand recognition system, we compared three types of features (see Chapter 2) as inputs to the SVM and MLP classifiers. Figure 4.9 shows a comparison of these three types of features, namely DWT, CT, and M5LOOP coefficients, used as inputs to the MLP and SVM classifiers. According to the obtained results, all descriptors achieved promising recognition rates on the tested dataset. The best recognition rate was 96.6%, which was achieved using the DWT features with both MLP and SVM classifiers, as well as the M_5LOOP features with the MLP classifier. In addition, the CT

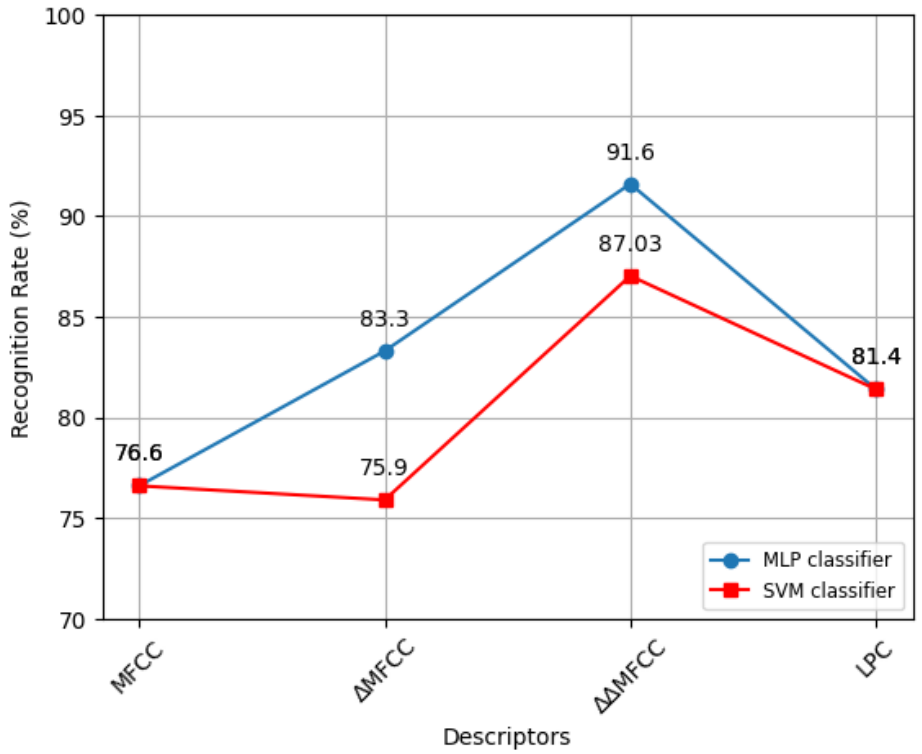


Figure 4.8: Global accuracy for speech using MLP and SVM classifier

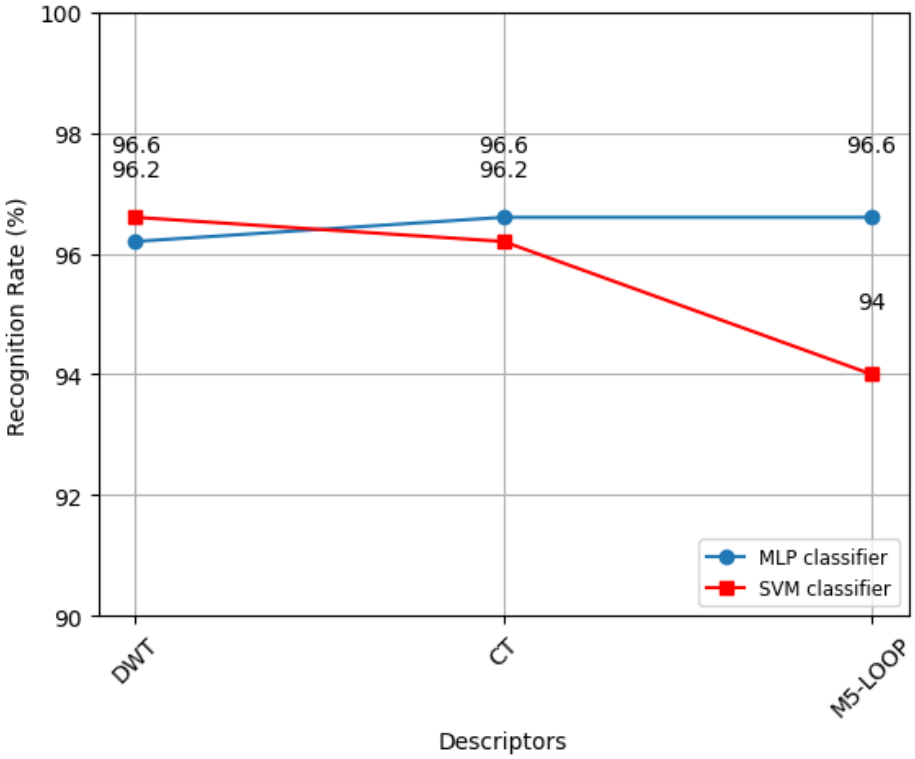


Figure 4.9: Global accuracy for gesture using MLP and SVM

features performs similar results, about 96% using SVM and MLP

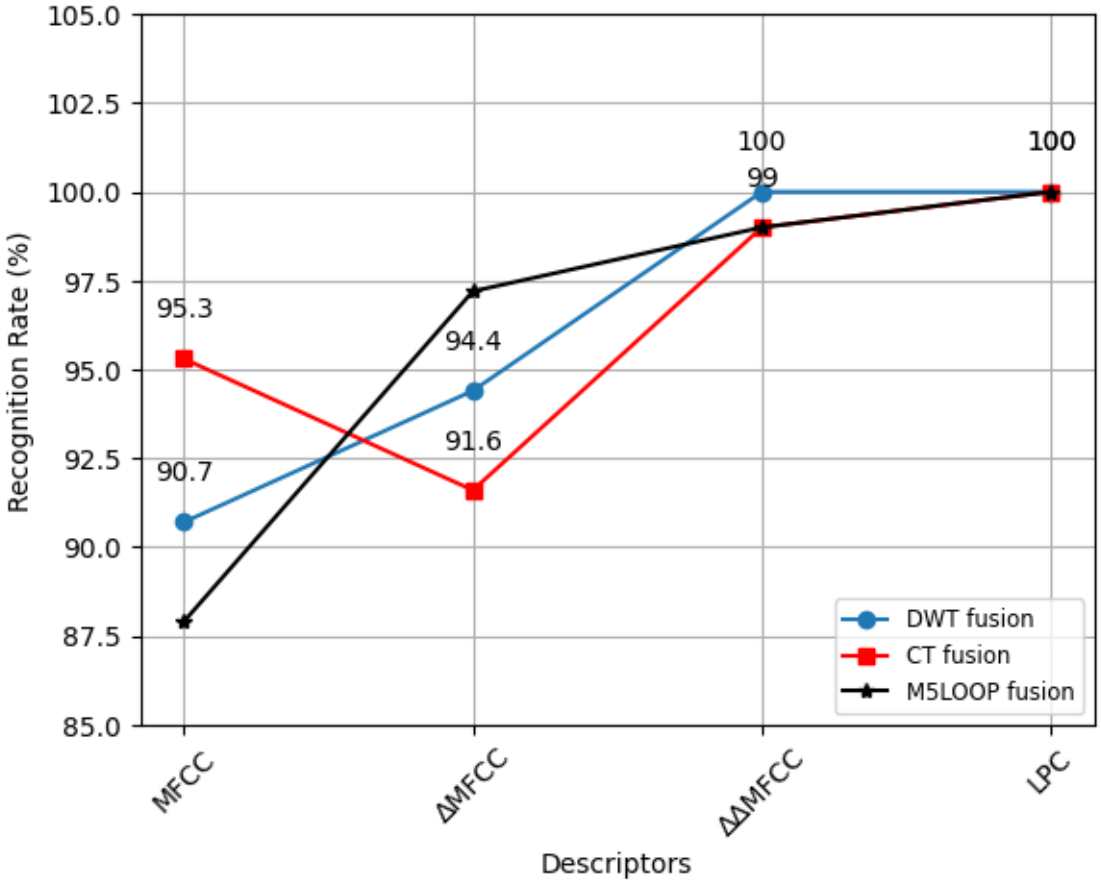


Figure 4.10: Global accuracy for seepch/gesture using MLP classifier

4.4.3 Performance analysis of the bimodal speech/gesture system

This section presents the results of our bimodal system, focusing on feature fusion techniques to improve recognition rates rather than analyzing individual modalities. This approach is based on the bimodal nature of human speech perception, where acoustic and gestural information are integrated, especially in noisy conditions. The following analysis presents the fusion results for each architecture.

After running the MLP classifier as well as the SVM on the corpus using twelfth architectures fusion where we combines the acoustic parameters with the gestural parameters . We obtained a global accuracy for each architecture,as shown in figure 4.10 and4.11 .

We demonstrate that hand gesture information, used to complement the acoustic speech signal, improves acoustic recognition scores. The fusion module combines the descriptor vectors from both modalities to make the final recognition decision.

We carried out experiments using different fusion methods in order to investigate their impact

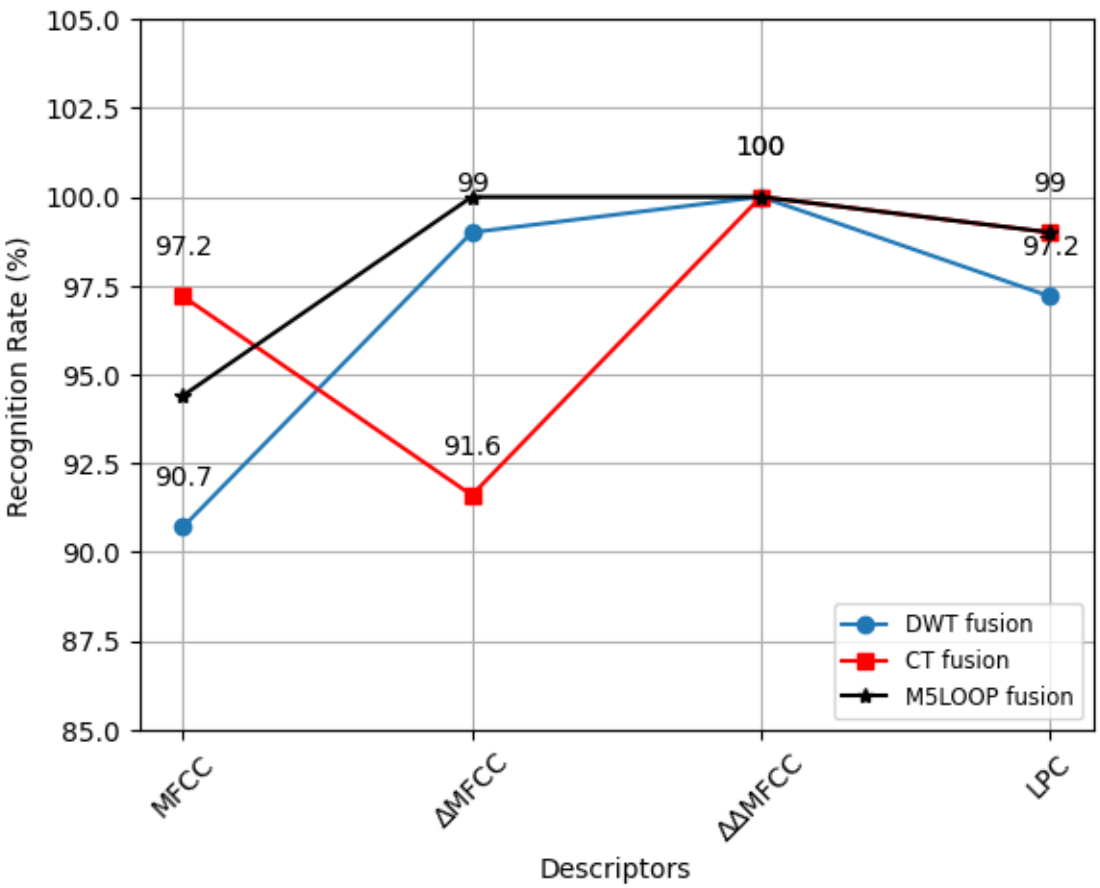


Figure 4.11: Global accuracy for seepch/gesture using SVM classifier

on the performance of the training procedure. From these results, the best recognition rate of 100 % was obtained with LPC fusion using the SVM classifier and ΔΔMFCC fusion using the MLP classifier. The lowest rates were obtained in the case of the MFCC fusion, were 87.5% of accuracy was obtained .

To better readability, a global analysis of the performance of the proposed bimodal system, as well as the mono-modal systems, is presented.

- The results clearly show that the ΔΔMFCC feature consistently outperformed MFCC, ΔMFCC, and LPC features across all audio clips.
- Acoustic modality results ranged from 75% to 91%, a variation attributable to factors such as speaker emotional and physical state, background noise, recording conditions, and speech rate.
- Comparing MLP and SVM classifiers, a slight performance difference was observed, while the SVM network converged faster than the MLP classifier.

- Although the fusion architecture using acoustic descriptors with DWT had a much shorter runtime (approximately 2 minutes) compared to CT (approximately 7 minutes), the SVM classifier yielded superior results to the MLP classifier.
- The best-performing architectures were $\Delta\Delta$MFCC and LPC, using the MLP classifier and DWT, CT, and M5-Loop descriptors.
- The results demonstrate that the gestural modality significantly improved speech recognition performance compared to the acoustic modality alone.

4.4.4 Discussion

The results clearly show that the bimodal system combining speech and gestures achieves the best recognition rate compared to the mono-modal systems.

We emphasize that the use of fusion information can significantly improve speech recognition system. The results obtained demonstrate not only the high performance of the proposed methodology, but also the fundamental possibility of recognizing speech and gestures.

In terms of accuracy, the fusion of speech and gesture information significantly improves overall recognition accuracy. The combination of speech and gesture data increases recognition accuracy because gestures provide spatial or temporal context that speech alone may not convey. This bimodal interaction feels more natural to users, enhancing the overall experience and potentially leading to more intuitive human-computer interactions.

In terms of robustness, the bimodal approach demonstrated enhanced resilience to errors that might occur in single-modality recognition.

To better demonstrate the advantages of the proposed system, a comparative analysis was performed using CNN architecture (see Table 4.5). The simulation results show that the mono-model system provides good results for hand gesture recognition, achieving a 99% accuracy rate, and for speech recognition, achieving a 96% accuracy rate. We conclude that the performance of both 2D-CNN and 3D-CNN models results in higher accuracy rates compared to classical systems. Notably, the fusion system achieved a 100% accuracy rate.

The CNN model was selected for comparison with our proposed machine learning architecture. Table 4.5 shows the overall accuracy achieved by the CNN, SVM, and MLP models, respectively, for the datasets used.

Table 4.5: Recognition rates obtained fo the different methods

Methods	**MLP**	**SVM**	**CNN**
Speech recognition system	87%	91%	99%
Hand recognition system	96%	96%	96%
Bimodal system	100%	100%	100%

Illustrating the overall accuracy, it was observed that: Compared our classical machine learning models (MLP and SVM) to CNN, the CNN have the advantage of not requiring a feature descriptor selection step. However, in most experiments, the CNN model outperforms or matches the performance of the other two models.

The CNN model improved the final classification rate of speech recognition systems, achieving 99% accuracy. This demonstrates the robustness of the hierarchical approach based on the use of Mel spectrograms, taking into account the entire set of the datasets.

4.5 Conclusion

In this chapter, we presented a comprehensive performance analysis of three proposed systems designed for speaker recognition and gesture recognition using various datasets and methodologies. The first proposed system focused on speaker recognition, examining the impact of different training functions on isolated words and continuous speech datasets.

The second proposed system was evaluated across multiple sign language datasets, including the El Halawani ArSL, ArSL2018, Jochen Triesch blue ASL, and a dynamic dataset from Sebastien Marcel. Each dataset posed unique challenges, and our system's ability to adapt and maintain high recognition rates was highlighted. Through a detailed discussion, we compared these results with existing methodologies, demonstrating the advantages and potential areas of improvement for our system.

The third proposed system explored multi-modal recognition, combining Arabic audio clips and 3D co-verbal gestures to enhance recognition accuracy. The performance analysis revealed that integrating multiple modalities significantly improves recognition rates compared to single-modality systems. The multi-modal speech/gesture system achieved superior results, reaching 100%

Overall, this chapter has shown that the proposed systems offer promising performance across various datasets and recognition tasks.

Conclusion

In conclusion, the recognition and interpretation of human body movements are pivotal areas of research, encompassing a wide range of actions from walking to communication. These movements provide significant insights into individuals' intentions, moods, and personalities. Despite the complexity and variability in gestures across different contexts and cultures, advancements in computer vision and machine learning offer promising solutions for accurately interpreting these movements.

This thesis focuses on developing a bimodal recognition system through the interpretation of gestural speech. By leveraging both acoustic and visual cues, this research aims to enhance the understanding of the interplay between verbal and non-verbal communication. The comprehensive approach includes creating a specialized audiovisual database, extracting relevant descriptors, and implementing advanced classification techniques.

We proposed a new bimodal dataset for a hand and speech recognition system. We developed three recognition systems using both classical and deep learning algorithms. The first contribution of this study is the design of a speaker-dependent recognition system using an MLP model with different activation functions. Various algorithms for extracting feature vector descriptors have been developed such as MFCC, LPC, and PCA for feature reduction. We presented a comparative performance analysis of speaker identification using different training function.The accuracy obtained is satisfactory for each technique, reaching up to 100%.

The second cotntribution consisted of presenting several textural features in the context of hand and sign gesture recognition systems. Five architectures, namely LBP/SVM, LDP/SVM, LOOP/SVM LGBHS/SVM, and Mi-LOOP/SVM, are implemented. This comprehensive investigation was carried out on five publicly available datasets: two for ArSL, two for ASL and one dynamic dataset (which describes four different gestures). Also new insights provided by proposing an improved LOOP descriptor, Mi-LOOP, which exploits the benefits of the LOOP, 2nd derivative Gaussian mask, and different proposed masks. Our experimental results show that the proposed

descriptor provides high accuracy with low computational cost. The effectiveness of the proposed system has been compared with several state-of-the-art approaches. The results show that our system outperforms most of the evaluated approaches, including deep learning-based techniques.

The third contribution provided new insights by investigating the ability of various modality fusion techniques and deep learning architectures for enhancing speech and gesture recognition systems. We explored the effectiveness of multiple feature extraction methods and fusion strategies, including MFCC/DWT, LPC/CT, and others. The performance of these fusion architectures was rigorously compared using both traditional classifiers (SVM and MLP) and advanced deep learning architectures combining 2D and 3D CNNs.

The research demonstrates significant advancements in multi-modal recognition systems, highlighting the potential of integrating multiple modalities and advanced machine learning techniques to interpret complex human movements more effectively. While this research has made significant advancements in the development of multi-modal recognition systems for speech and gesture interpretation, several areas warrant further investigation:

- Expansion of Multi-Modal Datasets: Future studies could focus on creating more diverse and comprehensive multi-modal datasets that include a wider range of gestures, speech patterns, and cultural contexts. This would improve the generalizability of recognition systems across different populations and applications.

- Improved Real-Time Processing: Enhancing the real-time processing capabilities of the proposed systems remains a critical challenge. Future research could explore more efficient algorithms and hardware acceleration techniques, such as leveraging GPU and TPU architectures, to achieve faster and more reliable performance in real-world applications.

- Integration of Additional Modalities: Beyond audio and visual cues, incorporating additional modalities, such as haptic feedback, physiological signals (e.g., heart rate, skin conductance), or even brain-computer interfaces, could further enrich the recognition systems and provide deeper insights into user intentions and emotional states.

- Advanced Deep Learning Architectures: Continued exploration of novel deep learning architectures, such as attention mechanisms, transformer models, and hybrid neural networks that combine convolutional and recurrent layers, could enhance the accuracy and robustness of multi-modal recognition systems.

- Personalization and Adaptability: Developing systems that can adapt to individual users over time by learning from their unique behavioral patterns and preferences would improve usability and user experience.

Bibliography

[1] Minh N Do and Martin Vetterli. Framing pyramids. *IEEE Transactions on Signal Processing*, 51(9):2329–2342, 2003.

[2] Roumiassa Ferhat and Fatma Zohra Chelali. Textural feature descriptors for a static and dynamic hand gesture recognition system. *Multimedia Tools and Applications*, 83(3):8165–8187, 2024.

[3] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419–429, 1999.

[4] Harriet Joseph Ottenheimer. *The anthropology of language: an introduction to linguistic anthropology*. Cengage Learning, 2005.

[5] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[6] Ilham Elouariachi, Rachid Benouini, Khalid Zenkouar, and Arsalane Zarghili. Robust hand gesture recognition system based on a new set of quaternion tchebichef moment invariants. *Pattern Analysis and Applications*, 23(3):1337–1353, 2020.

[7] Darpan Anand et al. An improved hand gesture recognition system based on optimized msvm and sift feature extraction algorithm. In *Proceedings of International Conference on Big Data, Machine Learning and their Applications*, pages 365–377. Springer, 2021.

[8] Nada B Ibrahim, Mazen M Selim, and Hala H Zayed. An automatic arabic sign language recognition system (arslrs). *Journal of King Saud University-Computer and Information Sciences*, 30(4):470–477, 2018.

[9] Manisha Kowdiki and Arti Khaparde. Automatic hand gesture recognition using hybrid meta-heuristic-based feature selection and classification with dynamic time warping. *Computer Science Review*, 39:100320, 2021.

[10] Igor LO Bastos, Michele F Angelo, and Angelo C Loula. Recognition of static gestures applied to brazilian sign language (libras). In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 305–312. IEEE, 2015.

[11] Houssem Lahiani and Mahmoud Neji. Hand gesture recognition method based on hog-lbp features for mobile devices. *Procedia Computer Science*, 126:254–263, 2018.

[12] Mahmood Jasim and Md Hasanuzzaman. Sign language interpretation using linear discriminant analysis and local binary patterns. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 1–5. IEEE, 2014.

[13] TG Subash Kumar and V Nagarajan. Local curve pattern for content-based image retrieval. *Pattern Analysis and Applications*, 22(3):1233–1242, 2019.

[14] Roumiassa Ferhat, Fatma Zohra Chelali, et al. Static and dynamic hand gesture recognition system using contourlet transform. In *The Proceedings of the Third International Conference on Smart City Applications*, pages 844–854. Springer, 2020.

[15] Priyanka Parvathy, Kamalraj Subramaniam, GKD Prasanna Venkatesan, P Karthikaikumar, Justin Varghese, and T Jayasankar. Development of hand gesture recognition system using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 12(6):6793–6800, 2021.

[16] Lianwei Li, Shiyin Qin, Zhi Lu, Dinghao Zhang, Kuanhong Xu, and Zhongying Hu. Real-time one-shot learning gesture recognition based on lightweight 3d inception-resnet with separable convolutions. *Pattern Analysis and Applications*, 24(3):1173–1192, 2021.

[17] Marco Fagiani, Emanuele Principi, Stefano Squartini, and Francesco Piazza. Signer independent isolated italian sign recognition based on hidden markov models. *Pattern Analysis and Applications*, 18(2):385–402, 2015.

[18] Anastassia Angelopoulou, Jose Garcia-Rodriguez, Sergio Orts-Escolano, Epaminondas Kapetanios, Xing Liang, Bencie Woll, and Alexandra Psarrou. Evaluation of different

chrominance models in the detection and reconstruction of faces and hands using the growing neural gas network. *Pattern Analysis and Applications*, 22(4):1667–1685, 2019.

[19] Tao Song, Honghua Zhao, Zhi Liu, Hao Liu, Yuanyuan Hu, and Dianmin Sun. Intelligent human hand gesture recognition by local–global fusing quality-aware features. *Future Generation Computer Systems*, 115:298–303, 2021.

[20] Hao Tang, Hong Liu, Wei Xiao, and Nicu Sebe. Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. *Neurocomputing*, 331:424–433, 2019.

[21] Haoming Liu and Zhenyu Liu. A multimodal dynamic hand gesture recognition based on radar–vision fusion. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023.

[22] Jimin Yu, Maowei Qin, and Shangbo Zhou. Dynamic gesture recognition based on 2d convolutional neural network and feature fusion. *Scientific Reports*, 12(1):4345, 2022.

[23] Liukai Xu, Keqin Zhang, Genke Yang, and Jian Chu. Gesture recognition using dual-stream cnn based on fusion of semg energy kernel phase portrait and imu amplitude image. *Biomedical Signal Processing and Control*, 73:103364, 2022.

[24] Abul Abbas Barbhuiya, Ram Kumar Karsh, and Rahul Jain. A convolutional neural network and classical moments-based feature fusion model for gesture recognition. *Multimedia Systems*, 28(5):1779–1792, 2022.

[25] Ekaterini Stergiopoulou and Nikos Papamarkos. Hand gesture recognition using a neural network shape fitting technique. *Engineering Applications of Artificial Intelligence*, 22(8):1141–1158, 2009.

[26] Jinqing Zheng, Zhiyong Feng, Chao Xu, Jing Hu, and Weimin Ge. Fusing shape and spatio-temporal features for depth-based dynamic hand gesture recognition. *Multimedia Tools and Applications*, 76(20):20525–20544, 2017.

[27] Mahmoud Elmezain, Ayoub Al-Hamadi, Jorg Appenrodt, and Bernd Michaelis. A hidden markov model-based continuous gesture recognition system for hand motion trajectory. In *2008 19th international conference on pattern recognition*, pages 1–4. IEEE, 2008.

[28] Mahmoud Elmezain, Ayoub Al-Hamadi, Gerald Krell, Sherif El-Etriby, and Bernd Michaelis. Gesture recognition for alphabets from hand motion trajectory using hidden markov models. In *2007 IEEE international symposium on signal processing and information technology*, pages 1192–1197. IEEE, 2007.

[29] Pavel A Popov and Robert Laganiere. Long hands gesture recognition system: 2 step gesture recognition with machine learning and geometric shape analysis. *Multimedia Tools and Applications*, pages 1–32, 2022.

[30] IA Adeyanju, OO Bello, and MA Azeez. Development of an american sign language recognition system using canny edge and histogram of oriented gradient. *Nigerian Journal of Technological Development*, 19(3):195–205, 2022.

[31] Shekhar Singh, Akshat Jain, and Deepak Kumar. Recognizing and interpreting sign language gesture for human robot interaction. *International Journal of Computer Applications*, 52(11), 2012.

[32] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi S Nayak. Hand gesture recognition for human computer interaction. *Procedia computer science*, 115:367–374, 2017.

[33] Hanwen Huang, Yanwen Chong, Congchong Nie, and Shaoming Pan. Hand gesture recognition with skin detection and deep learning method. In *Journal of Physics: Conference Series*, volume 1213, page 022001. IOP Publishing, 2019.

[34] Ming-Hsuan Yang, Narendra Ahuja, and Mark Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on pattern analysis and machine intelligence*, 24(8):1061–1074, 2002.

[35] Vaidyanath Areyur Shanthakumar, Chao Peng, Jeffrey Hansberger, Lizhou Cao, Sarah Meacham, and Victoria Blakely. Design and evaluation of a hand gesture recognition approach for real-time interactions. *Multimedia Tools and Applications*, 79(25):17707–17730, 2020.

[36] Linchu Yang, Ji'an Chen, and Weihang Zhu. Dynamic hand gesture recognition based on a leap motion controller and two-layer bidirectional recurrent neural network. *Sensors*, 20(7):2106, 2020.

[37] Zhou Ren, Junsong Yuan, and Zhengyou Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1093–1096, 2011.

[38] Jingren Tang, Hong Cheng, Yang Zhao, and Hongliang Guo. Structured dynamic time warping for continuous hand trajectory gesture recognition. *Pattern Recognition*, 80:21–31, 2018.

[39] Muaammar Hadi Kuzman Ali, M Asyraf Azman, Zool Hilmi Ismail, et al. Real-time hand gestures system for mobile robots control. *Procedia Engineering*, 41:798–804, 2012.

[40] Joongrock Kim, Sunjin Yu, Dongchul Kim, Kar-Ann Toh, and Sangyoun Lee. An adaptive local binary pattern for 3d hand tracking. *Pattern Recognition*, 61:139–152, 2017.

[41] Ana I Maqueda, Carlos R del Blanco, Fernando Jaureguizar, and Narciso García. Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns. *Computer Vision and Image Understanding*, 141:126–137, 2015.

[42] Houssem Lahiani, Monji Kherallah, and Mahmoud Neji. Hand gesture recognition system based on local binary pattern approach for mobile devices. In *International Conference on Intelligent Systems Design and Applications*, pages 180–190. Springer, 2017.

[43] CV Sheena and NK Narayanan. Static gesture classification and recognition using hog feature parameters and k-nn and svm-based machine learning algorithms. In *Machine Vision and Augmented Intelligence—Theory and Applications*, pages 157–166. Springer, 2021.

[44] Sakshi Sharma and Sukhwinder Singh. Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Systems with Applications*, 182:115657, 2021.

[45] Gajalakshmi Palanisamy and Sree Sharmila Thangaswamy. An efficient hand gesture recognition based on optimal deep embedded hybrid convolutional neural network-long short term memory network model. *Concurrency and Computation: Practice and Experience*, page e7109, 2022.

[46] Xiao Yan Wu. A hand gesture recognition algorithm based on dc-cnn. *Multimedia Tools and Applications*, 79(13):9193–9205, 2020.

[47] Akansha Tyagi and Sandhya Bansal. Hybrid fist_cnn approach for feature extraction for vision-based indian sign language recognition. *Int. Arab J. Inf. Technol.*, 19(3):403–411, 2022.

[48] Shroog Alshomrani, Lina Aljoudi, and Muhammad Arif. Arabic and american sign languages alphabet recognition by convolutional neural network. *Advances in Science and Technology. Research Journal*, 15(4), 2021.

[49] Ghazanfar Latif, Nazeeruddin Mohammad, Roaa AlKhalaf, Rawan AlKhalaf, Jaafar Alghazo, and Majid Khan. An automatic arabic sign language recognition system based on deep cnn: an assistive system for the deaf and hard of hearing. *International Journal of Computing and Digital Systems*, 9(4):715–724, 2020.

[50] Reem Aljuhani, Aseel Alfaidi, Bushra Alshehri, Hajer Alwadei, Eman Aldhahri, and Nahla Aljojo. Arabic sign language recognition using convolutional neural network and mobilenet. *Arabian Journal for Science and Engineering*, pages 1–8, 2022.

[51] Jiaqi Dong, Zeyang Xia, and Qunfei Zhao. Augmented reality assisted assembly training oriented dynamic gesture recognition and prediction. *Applied Sciences*, 11(21):9789, 2021.

[52] S Rubin Bose and V Sathiesh Kumar. In-situ recognition of hand gesture via enhanced xception based single-stage deep convolutional neural network. *Expert Systems with Applications*, 193:116427, 2022.

[53] Vimal Shanmuganathan, Harold Robinson Yesudhas, Mohammad S Khan, Manju Khari, and Amir H Gandomi. R-cnn and wavelet feature extraction for hand gesture recognition with emg signals. *Neural Computing and Applications*, 32(21):16723–16736, 2020.

[54] Mahin Moghbeli Damaneh, Farahnaz Mohanna, and Pouria Jafari. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using orb descriptor and gabor filter. *Expert Systems with Applications*, 211:118559, 2023.

[55] Soumen Das, Saroj Kr Biswas, and Biswajit Purkayastha. Automated indian sign language recognition system by fusing deep and handcrafted feature. *Multimedia Tools and Applications*, 82(11):16905–16927, 2023.

[56] Yu Zhang, Junlin Wang, Xin Wang, Haonan Jing, Zhanshuo Sun, and Yu Cai. Static hand gesture recognition method based on the vision transform-er. *Multimedia Tools and Applications*, 82(20):31309–31328, 2023.

[57] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020.

[58] Soheil Zabihi, Elahe Rahimian, Amir Asif, and Arash Mohammadi. Trahgr: Transformer for hand gesture recognition via electromyography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[59] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with center-symmetric local binary patterns. In *Computer vision, graphics and image processing*, pages 58–69. Springer, 2006.

[60] Taskeed Jabid, Md Hasanul Kabir, and Oksam Chae. Robust facial expression recognition based on local directional pattern. *ETRI journal*, 32(5):784–794, 2010.

[61] Taskeed Jabid, Md Hasanul Kabir, and Oksam Chae. Local directional pattern (ldp) for face recognition. In *2010 digest of technical papers international conference on consumer electronics (ICCE)*, pages 329–330. IEEE, 2010.

[62] Aminou Halidou, Xinge You, Mahamadou Hamidine, Roger Atsa Etoundi, Laye Hadji Diakite, et al. Fast pedestrian detection based on region of interest and multi-block local binary pattern descriptors. *Computers & Electrical Engineering*, 40(8):375–389, 2014.

[63] Shekhar Karanwal and Manoj Diwakar. Neighborhood and center difference-based-lbp for face recognition. *Pattern Analysis and Applications*, 24(2):741–761, 2021.

[64] Fujin Zhong and Jiashu Zhang. Face recognition with enhanced local directional patterns. *Neurocomputing*, 119:375–384, 2013.

[65] Tapabrata Chakraborti, Brendan McCane, Steven Mills, and Umapada Pal. Loop descriptor: local optimal-oriented pattern. *IEEE Signal Processing Letters*, 25(5):635–639, 2018.

[66] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-based recognition*, pages 227–243. Springer, 1997.

[67] Cao Dong, Ming C Leu, and Zhaozheng Yin. American sign language alphabet recognition using microsoft kinect. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 44–52, 2015.

[68] Gamal Tharwat, Abdelmoty M Ahmed, and Belgacem Bouallegue. Arabic sign language recognition system for alphabets using machine learning techniques. *Journal of Electrical and Computer Engineering*, 2021, 2021.

[69] Omar Al-Jarrah and Alaa Halawani. Recognition of gestures in arabic sign language using neuro-fuzzy systems. *Artificial Intelligence*, 133(1-2):117–138, 2001.

[70] Khaled Assaleh and M Al-Rousan. Recognition of arabic sign language alphabet using polynomial classifiers. *EURASIP Journal on Advances in Signal Processing*, 2005(13):1–10, 2005.

[71] Tamer Shanableh, Khaled Assaleh, and Mohammad Al-Rousan. Spatio-temporal feature-extraction techniques for isolated gesture recognition in arabic sign language. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(3):641–650, 2007.

[72] Omar Al-Jarrah, Ali Shatnawi, and Alaa H Halawani. Recognition of gestures in arabic sign language using neural networks. *Proceeding of Artificial Intelligence and Soft Computing*, pages 28–30.

[73] Richard Ohmann. Speech acts and the definition of literature. *Philosophy & Rhetoric*, pages 1–19, 1971.

[74] Soufiane Hourri and Jamal Kharroubi. A novel scoring method based on distance calculation for similarity measurement in text-independent speaker verification. *Procedia computer science*, 148:256–265, 2019.

[75] Mahdi Barhoush, Ahmed Hallawa, and Anke Schmeink. Speaker identification and localization using shuffled mfcc features and deep learning. *International Journal of Speech Technology*, 26(1):185–196, 2023.

[76] Mohamed Chetouani, Marcos Faundez-Zanuy, Bruno Gas, and Jean-Luc Zarader. A new nonlinear speaker parameterization algorithm for speaker identification. *arXiv preprint arXiv:2204.02609*, 2022.

[77] JiangKun Zhao, HouPan Zhou, HongLei Liu, and YongHai Du. Feature fusion method for speaker recognition based on embedding mechanism. In *International Conference on Signal Processing and Communication Security (ICSPCS 2022)*, volume 12455, pages 108–113. SPIE, 2022.

[78] VR Sreehari and Leena Mary. Automatic short utterance speaker recognition using stationary wavelet coefficients of pitch synchronised lp residual. *International Journal of Speech Technology*, pages 1–15, 2022.

[79] Mandar Nitin Kakade and Dharmraj B Salunke. Real-time speech recognition using convolutional neural network. In *Advanced Machine Intelligence and Signal Processing*, pages 467–477. Springer, 2022.

[80] Hong Lei, Yue Xiao, Yanchun Liang, Dalin Li, Heow Pueh Lee, et al. Dld: An optimized chinese speech recognition model based on deep learning. *Complexity*, 2022, 2022.

[81] Noshin Nirvana Prachi, Faisal Mahmud Nahiyan, Md Habibullah, and Riasat Khan. Deep learning based speaker recognition system with cnn and lstm techniques. In *2022 Interdisciplinary Research in Technology and Management (IRTM)*, pages 1–6. IEEE, 2022.

[82] Sunanda Mendiratta, Neelam Turk, and Dipali Bansal. Robust feature extraction and recognition model for automatic speech recognition system on news report dataset. In *Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces*, pages 589–601. Springer, 2022.

[83] Hiba Qasim Jaber and Huda Abdulaali Abdulbaqi. Real time arabic speech recognition based on convolution neural network. *Journal of Information and Optimization Sciences*, 42(7):1657–1663, 2021.

[84] Vaishali A Kherdekar and Sachin A Naik. Convolution neural network model for recognition of speech for words used in mathematical expression. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6):4034–4042, 2021.

[85] Mona A Azim, Wedad Hussein, and Nagwa L Badr. Spoken arabic digits recognition system using convolutional neural network. In *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021*, pages 164–172. Springer, 2021.

[86] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun. An enhanced speech emotion recognition using vision transformer. *Scientific Reports*, 14(1):13126, 2024.

[87] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A multi-modal transformer for visual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6658–6662. IEEE, 2021.

[88] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018.

[89] Denis Ivanko. *Audio-visual Russian speech recognition*. PhD thesis, Universität Ulm, 2022.

[90] Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3):141–151, 2000.

[91] Denis Ivanko, Alexey Karpov, Dmitrii Fedotov, Irina Kipyatkova, Dmitry Ryumin, Dmitriy Ivanko, Wolfgang Minker, and Milos Zelezny. Multimodal speech recognition: increasing accuracy using high speed video data. *Journal on Multimodal User Interfaces*, 12:319–328, 2018.

[92] Andrea Corradini and Philip R Cohen. Multimodal speech-gesture interface for handfree painting on a virtual paper using partial recurrent neural networks as gesture recognizer. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2293–2298. IEEE, 2002.

[93] Jian Ding, Jin Liu, Xiaolin Ning, and Zhiwei Kang. Dms-sk/blstm-ctc hybrid network for gesture/speech fusion and its application in lunar robot–astronauts interaction. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(07):2258005, 2022.

[94] Haodong Chen, Niloofar Zendehdel, Ming C Leu, and Zhaozheng Yin. Multi-modal fine-grained activity recognition and prediction in assembly. 2022.

[95] Xiaoyu Song, Hong Chen, and Qing Wang. A gesture recognition approach using multimodal neural network. In *Journal of Physics: Conference Series*, volume 1544, page 012127. IOP Publishing, 2020.

[96] Olena Vynokurova and Dmytro Peleshko. Hybrid multidimensional deep convolutional neural network for multimodal fusion. In *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pages 131–135. IEEE, 2020.

[97] Khaled Saleh. Hybrid seq2seq architecture for 3d co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 748–752, 2022.

[98] Haodong Chen, Ming C Leu, and Zhaozheng Yin. Real-time multi-modal human–robot collaboration using gestures and speech. *Journal of Manufacturing Science and Engineering*, 144(10):101007, 2022.

[99] Xiaoyu Qiu, Zhiquan Feng, Xiaohui Yang, and Jinglan Tian. Multimodal fusion of speech and gesture recognition based on deep learning. In *Journal of Physics: Conference Series*, volume 1453, page 012092. IOP Publishing, 2020.

[100] S Kay. Noise compensation for autoregressive spectral estimates. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(3):292–303, 1980.

[101] Douglas O'Shaughnessy. Linear predictive coding. *IEEE potentials*, 7(1):29–32, 1988.

[102] Yu Wu, Hua Mao, and Zhang Yi. Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems*, 161:90–100, 2018.

[103] Yu Su, Ke Zhang, Jingyu Wang, Daming Zhou, and Kurosh Madani. Performance analysis of multiple aggregated acoustic features for environment sound classification. *Applied Acoustics*, 158:107050, 2020.

[104] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[105] Cecile Fougeron and Nicolas Audibert. Variabilité intra-individuelle en parole lue. In *XXXIVe Journées d'Études sur la Parole–JEP 2022*, pages 145–153, 2022.

[106] Olivier Rioul and Pierre Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE transactions on information theory*, 38(2):569–586, 1992.

[107] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[108] Minh N Do and Martin Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on image processing*, 14(12):2091–2106, 2005.

[109] Peter J Burt. Edward, and eh adelson. the laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(532-540):340, 1983.

[110] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[111] Roberto H Bamberger and Mark JT Smith. A filter bank for the directional decomposition of images: Theory and design. *IEEE transactions on signal processing*, 40(4):882–893, 1992.

[112] Martin Vetterli. Multi-dimensional sub-band coding: Some theory and algorithms. *Signal processing*, 6(2):97–112, 1984.

[113] Li Wang and Dong-Chen He. Texture classification using texture spectrum. *Pattern recognition*, 23(8):905–910, 1990.

[114] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 786–791. IEEE, 2005.

[115] Vincent Torre and Tomaso A Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):147–163, 1986.

[116] ER Davies. Constraints on the design of template masks for edge detection. *Pattern recognition letters*, 4(2):111–120, 1986.

[117] Ferhat Roumiassa and Fatma-Zohra Chelali. Speaker identification and verification system for arabic and berber language. In *2020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP)*, pages 242–247. IEEE, 2020.

[118] Fatma Zohra Chelali and Amar Djeradi. Text dependant speaker recognition using mfcc, lpc and dwt. *International Journal of Speech Technology*, 20:725–740, 2017.

[119] Marcus Gallagher. *Multi-layer perceptron error surfaces: visualization, structure and modelling*. PhD thesis, Citeseer, 2000.

[120] Jayabalan Ramesh, Ponnusamy Thangapandian Vanathi, and Kandasamy Gunavathi. Fault classification in phase-locked loops using back propagation neural networks. *ETRI journal*, 30(4):546–554, 2008.

[121] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[122] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[123] Grégoire Mercier and Marc Lennon. Support vector machines for hyperspectral image classification with spectral-based kernels. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, volume 1, pages 288–290. IEEE, 2003.

[124] Y Liu and H Huang. Fuzzy support vector machines for pattern recognition and data mining. *International journal of fuzzy systems*, 4(3):826–835, 2002.

[125] Yea S. Huang and Ching Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE transactions on pattern analysis and machine intelligence*, 17(1):90–94, 1995.

[126] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, 2015.

[127] Bineet Kaur and Garima Joshi. Lower order krawtchouk moment-based feature-set for hand gesture recognition. *Advances in Human-Computer Interaction*, 2016, 2016.

[128] Aparna Mohanty, Sai Saketh Rambhatla, and Rajiv Ranjan Sahay. Deep gesture: static hand gesture recognition using cnn. In *Proceedings of International Conference on Computer Vision and Image Processing*, pages 449–461. Springer, 2017.

[129] Jaya Prakash Sahoo, Samit Ari, and Dipak Kumar Ghosh. Hand gesture recognition using dwt and f-ratio based feature descriptor. *IET Image Processing*, 12(10):1780–1787, 2018.

[130] Jochen Triesch and Christoph Von Der Malsburg. Robust classification of hand postures against complex backgrounds. In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 170–175. IEEE, 1996.

[131] Khadidja Sadeddine, Rachida Djeradi, Fatma Zohra Chelali, and Amar Djeradi. Recognition of static hand gesture. In *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, pages 1–6. IEEE, 2018.

[132] Salah eddine Agab and Fatma zohra Chelali. Hog and hoof spatio-temporal descriptors for gesture recognition. In *2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, pages 1–7. IEEE, 2018.

[133] Saeed Nasri, Alireza Behrad, and Farbod Razzazi. A novel approach for dynamic hand gesture recognition using contour-based similarity images. *International Journal of Computer Mathematics*, 92(4):662–685, 2015.

[134] Salah eddine Agab and Fatma zohra Chelali. Dynamic hand gesture recognition based on textural features. In *2019 International Conference on Advanced Electrical Engineering (ICAEE)*, pages 1–6. IEEE, 2019.

[135] Salah Eddine Agab, Roumaissa Ferhat, and Fatma Zohra Chelali. Robust hand gesture recognition based on mwld and ltep descriptors. In *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*, pages 1–5. IEEE, 2021.