Thesis for the Degree of Doctor of Philosophy

Human Facial Expression Recognition using Facial Movement Based Feature Extraction and Hidden Conditional Random Fields

Muhammad Hameed Siddiqi

Department of Computer Engineering
Graduate School
Kyung Hee University
South Korea

February 2016

Human Facial Expression Recognition using Facial Movement Based Feature Extraction and Hidden Conditional Random Fields

Muhammad Hameed Siddiqi

Department of Computer Engineering
Graduate School
Kyung Hee University
South Korea

February 2016

Human Facial Expression Recognition using Facial Movement Based Feature Extraction and Hidden Conditional Random Fields

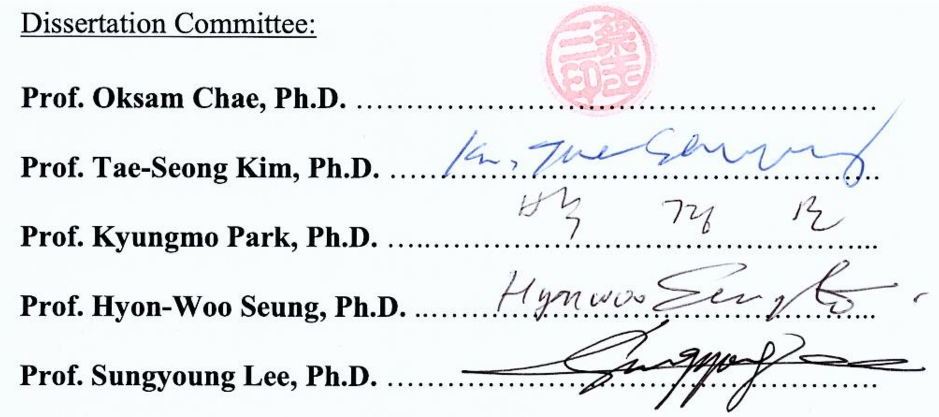
by

Muhammad Hameed Siddiqi

Supervised by

Prof. Sungyoung Lee, Ph.D.

Submitted to the Department of Computer Engineering
and the Faculty of the Graduate School of
Kyung Hee University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy



Dissertation Committee:

Prof. Oksam Chae, Ph.D. ....

Prof. Tae-Seong Kim, Ph.D. .... Kim, Tae Seong

Prof. Kyungmo Park, Ph.D. .... 박 경 모

Prof. Hyon-Woo Seung, Ph.D. .... Hyonwoo Seung

Prof. Sungyoung Lee, Ph.D. ....

Dedicated To

my beloved father and mother for their never ending support
and open-mindedness.

Abstract

Knowledge about people's emotions can serve as an important context for human behaviour analysis in context-aware systems. Hence, human facial expression recognition (FER) has emerged as an important research area over the last two decades. FER has a vital role in many research applications of image processing and pattern recognition such as human computer interaction, robot control and driver state surveillance, and human behavior studies in telemedicine and e-health environments. In healthcare, this technology could help identify the mood of a patient, especially in the case of physically disabled people. A paralyzed patient whose body is critically attacked by paralysis is completely unable to speak and the only way to understand him might be through facial expressions. Similarly, for stroke patients, it is necessary to identify their expressions and generate guidelines for their care and protection. This technology would be helpful to both the care-givers and medical experts to effectively use the resources and treat the patient. Similarly, for the people who are healthy but are unable to speak and understand, they could also be guided using FER systems, especially in times of their loneliness.

To accurately recognize expressions, FER systems require automatic face detection followed by the extraction of robust features from important facial parts. Furthermore, the process should be less susceptible to the presence of noise, such as different lighting conditions and variations in facial characteristics of subjects. Moreover, the expressions have high similarity among different expressions resulting in overlaps among feature values of different classes in the feature space. Though, several FER systems have been proposed in the past that showed promising results for a certain dataset, their performance was significantly reduced when tested with different datasets. Furthermore, these systems utilized existing publicly available standard datasets of facial expressions. However, these systems are far away from real life datasets collected in real world scenarios.

Accordingly, this thesis implements accurate and robust FER systems capable of providing high recognition accuracy even in the presence of aforementioned variations. The first system uses a hierarchical recognition scheme in order to solve the high-within class variance and low-between class variance problem. However, this system uses two level classification (two HMMs) which creates complexity issue; therefore, we proposed a second robust system. This system uses new methods for feature extraction and classification. In feature extracting, first noise reduction is achieved by means of wavelet decomposition, followed by the extraction of facial movement features using optical flow. These features reflect facial muscle movements which signify static, dynamic, geometric, and appearance characteristics of facial expressions. Finally, we have proposed an improved version of hidden conditional random fields (HCRF) model in order to classify the expressions which is capable of approximating a complex distribution using a mixture of Gaussian density functions.

The performance of the proposed FER system (including both the components: feature extraction and classification) was validated using publicly available standard datasets such as cohn-kanade and JAFFE datasets. These were used for the first system while, for the second system, the extended cohn-kanade (CK+), USTC-NVIE (both pose and spontaneous), MUG, MMI, Indian Movie Face Database (IMFDB), and Radboud Faces (RaFD) datasets were utilized.

For each dataset, a $10-$fold cross-validation scheme (based on subjects) was utilized. In other words, out of 10 subjects data from a single subject was used as the validation data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the validation data. Moreover, the system was trained on one dataset and tested on another dataset in order to show the robustness of the proposed system. For this evaluation, $n-$fold cross-validation scheme (based on dataset) was utilized. From the testing and validation, it is clear that the proposed FER system outperformed the existing state-of-the-art systems. Thus, the proposed FER system shows significant potential in its ability to accurately and robustly recognize human facial expressions using video data in naturalistic environments.

In most of the datasets, RGB cameras were utilized which may raise privacy concerns; therefore, in order to solve this concerns, depth camera will be utilized in the further study and then will check the accuracy and robustness of the proposed system.

Acknowledgement

In the name of Allah, the Beneficent, the Merciful.

”Read! In the Name of your Lord, Who has created (all that exists), He has created man from a clot (a piece of thick coagulated blood) Read! And your Lord is the Most Generous, Who has taught (the writing) by the pen. He has taught man that which he knew not.”

(Quran: Chapter 96: Surah Al-Alaq (The Clot), verses 1-5.)

First and foremost, I deliver my humble and earnest thanks to the **Almighty ALLAH** for showering His blessings in every possible form upon me. He gave me strength, courage, patience and introduced to me all those people who made my studies and stay in Korea a pleasant experience. Without His help I could not have made any step forward.

This dissertation signifies an enormous deal of time and effort not only on my part, but also on part of my advisor, Prof. Sungyoung Lee. I am grateful for the time and advice that you all provided me over the past four years. This thesis owes much of its contents to your ideas and guidance. You helped me in shaping up my research from day one, pushed me to get through the foreseeable research setbacks, and encouraged me to achieve the best of my aptitudes.

I am also thankful to my other thesis committee members for providing insightful and constructive comments to improve the quality of this dissertation. Their constructive criticisms on my work and insightful suggestions helped me in improving this dissertation a lot.

I am very grateful to my immediate advisor Prof. Adil Mehmood Khan for his invaluable help during formulation of my idea and papers preparation process. I would also like to thank all the current and former members of my lab for their kind support and for providing an amusing working environment. Every lab member is worthy to be praised and I appreciate them all.

I have no words to express my gratitude to my family for their endless support, and especially brothers Muhammad Raies, Muhammad Saeed Siddiqi, Dr. Muhammad Hanif Siddiqi, Dr. Muhammad Shafi Siddiqi, and Muhammad Zubair Siddiqi for their guidance, love and prayers. I would like to acknowledge the sacrifices made by my parents and sister for my better education and upbringing.

I would like to lengthen my thanks to all my friends and colleagues in KHU, Dr. Oresti Banos, Dr. Muhammad Shoaib Siddiqui, Dr. Asad Masood Khattak, Dr. Zeeshan PERVEZ, Dr. La The Vinh, Dr. Nguyen Duc Thang, Dr. Phan Tran Ho Truc, Dr. Jalal Ahmad, Dr. Muhammad Fahim, Dr. Wajahat Ali Khan, Dr. Bilal Amin, Le Ba Vui, Shujjat Husain, Kifayatullah Khan, Maqbool Hussain, Rahman Ali, Saeedullah, Muhammad Idris, Tae Ho Hur, Jaehun Bang, and others for their friendship and help to overcome the difficulties throughout my thesis research. Finally, I would like to thank my mother-in-law, my loving wife Gulzar Siddiqi for their love, prayers, support and encouragement that reinforced my spirits at some crucial junctures. Last but not the least, so much love to my sweet daughter ”Simran Hameed”.

Muhammad Hameed Siddiqi
Seoul, Korea
February, 2016

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Background

Expressions play a vital role in our daily communications, and recent years have witnessed a great amount of work being done to develop accurate and reliable facial expressions recognition (FER) systems. Such systems can be employed in many real life applications. Facial expression recognition (FER) is one of the most active areas of research in computer science, due to its importance in a large number of application domains. These systems are widely in demand for humanoid robotics, affective sensitive computing in human computer interaction (HCI), behavioral sciences, video games, and psychiatry [3]. A major breakthrough in the field of HCI is the ability of the computing systems to detect human faces and code them into various dimensions such as happy, neutral, sad, angry, happy, and disgust [4]. Human-robot communication is one practical example, where a robot could identify such facial expressions, which makes the robot to respond humans correctly, as a human would do [5]. FER technology is of huge importance for intelligent security systems, as it would real-time surveillance as a part of biometric authentication [6, 7], such as for banking, armed forces equipment, and high security installments. These systems could be used to secure personalized access to the data and many mobile applications have been developed [8].

FER systems can be categorized into two types: pose expression recognition systems [9–11] and spontaneous expression recognition systems [12–14]. Former case deals with recognizing artificial expressions: expressions produced by people when they are asked to do so [3]. On the other hand, the latter case deals with the expressions that people give out spontaneously, and these are the ones that can be observed on a day-to-day basis, such as during conservations, while watching movies [3], and etc. The focus of this study is both pose and spontaneous based FER.

In a typical FER system, there are two types of classifications: frame-based classification, and sequence-based classification. In frame-based classification methods, only the current frame is utilized with or without a reference image (neutral face image) in order to recognize the expressions; while, in sequence-based classification methods, the temporal information of the sequences are utilized in order to recognize the expressions of one or more frames [15]. In sequence-based methods, the geometrical displacement of facial feature points between the current frame and the initial frame is calculated [16]; while, frame-based methods do not have this property. The temporal information of expression in sequences of frames is important for facial expression analysis [17]; therefore, we employed sequence-based classification.

General FER system consists of four main components, preprocessing, feature extraction, feature selection, and recognition modules. In the preprocessing module, the noise is removed from the facial images, and the faces are detected and extracted, because it consists all of the important parts that convey the expressions [18]. Feature extraction deals with extraction the distinguishable features from each facial expression frame and quantizing them as discrete symbols. Feature selection module is used for selecting a subset of relevant features from a large number of features extracted from the input data. Finally, in the recognition module, a classifier is first trained using the training data and then used to generate labels for the expressions in the incoming video data.

1.1 Motivation

Cure provisioning to patients by remotely analyzing their facial expressions can be promising breakthrough development in telemedicine and healthcare domains. This will assist clinicians in their decision-making process by monitoring patients' facial expressions remotely, specifically in disease such as stroke. A concept of telestroke was introduced in [19] that explains a case study for monitoring acute stroke patient using telemedicine. Stroke patients' state can be defined using facial expression identification techniques and facial exercises can be recommended. Sad and happy facial expressions can be identified for stroke patients and therefore guidelines can be remotely provided for better care of patient. Another example is that psychiatrist can use telemedicine technology for treating a patient with post-traumatic stress disorder (PTSD) by monitoring his/her facial expressions remotely [20]. The study to monitor heart-failure patients using telemedicine [21] is a prime candidate for a facial expression analysis. These developments will facilitate clinicians and physicians for efficiently monitoring and management of patient disease. FER is an observable indication of person's sentimental state, mental activity, and behavior [22]. In Short, FER systems would be helpful to both the care-givers and medical experts to effectively use the resources and treat the patient. Similarly, for the people who are healthy but are unable to speak and understand, they could also be guided using FER systems, especially in times of their loneliness.

Most of the previous pose-based FER systems do not have the capability to accurately recognize the expressions using publicly available standard datasets of facial expressions. This lack of accuracy can be attributed to various causes, such as the failure to extract prominent features, and the high similarity among different facial expressions that results due to the presence of low between-class variance in the feature space.

Moreover, a great number of FER systems have been implemented, each surpassing the other in terms of classification accuracy. However, one major weakness found in the previous studies is that they have all used standard datasets for their evaluations and comparisons. Though this serves well, given the needs of a fair comparison with existing systems, it is argued that this does not go in hand with the fact that these systems are built with a hope of eventually being used in the real-world. It is because these datasets assume a predefined camera setup, consist of mostly posed expressions collected in a controlled setting, using fixed background and static ambient settings, and having low variations in the face size and camera angles, which is not the case in naturalistic environments.

In conclusion, challenges presented in Section 1.2 there is an opportunity to create more advanced FER system capable of handling posed and spontaneous FER issues as well as to incorporate more intelligent capabilities to transform experimental prototypes into actual usable applications. Thus the challenge facing this work, is to identify and characterize some of the most relevant limitations in the FER domain and develop solutions that may help overcome these complex problems.

Approaches to Human Emotion Recognition

The first step towards achieving the goal of recognizing the emotions of daily life conversation is to equip the emotion recognition systems with the recognizing capabilities. Three approaches have been mainly employed for this purpose: audio based, video based, and physiological based, as shown in Figure 1.1.

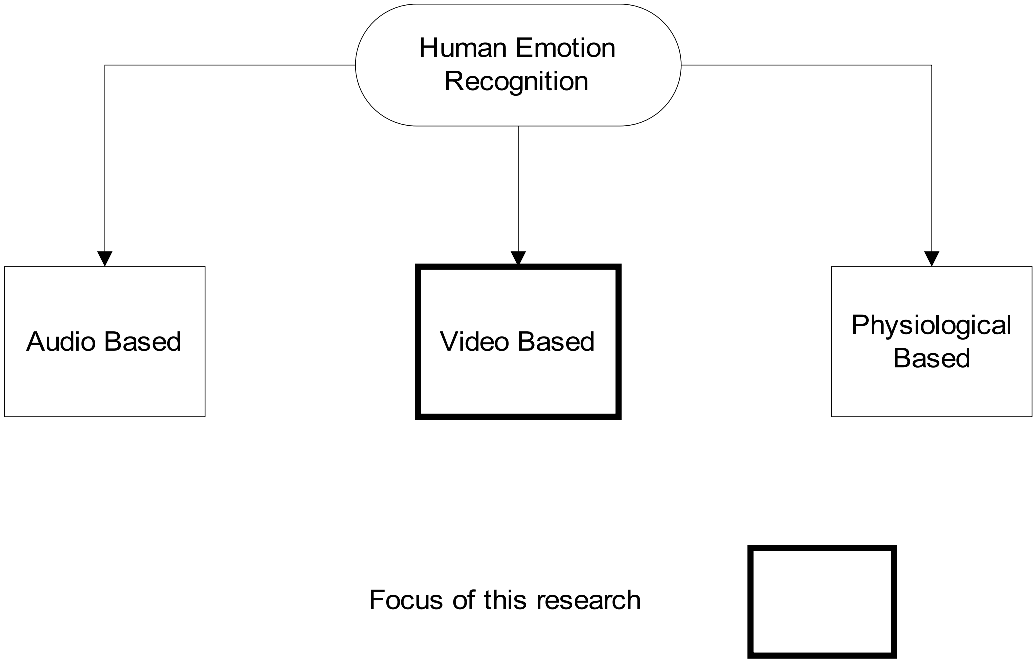


Figure 1.1: Three approaches employed for human emotion recognition.

Audio Based Systems

In audio-based emotion recognition systems, the audio sensors are used in order to recognize the human emotions [23]. Most of these systems work well in laboratory environments under predefined settings with fix positions of microphone in the frontal view of the face. All of these systems are subjects depended, which means that when the systems trained on one subject and test or validate using different subjects, the system fails to recognize the human emotions. The reason is that vocal tones based emotion recognition suffers due to culture, gender, and age factors. Also, the pitch of sound varies from subject to subject due to which the systems might not be able to recognize the emotions of a human correctly. Moreover, some environmental noise such as ambient noise may cause misclassification.

Physiological Based Systems

Physiological-based systems are developed to recognize the emotions of normal humans [24, 25]. These systems showed significant performance; however, the systems have some critical weaknesses such as the systems need special equipment due to which they are obtrusive to the users. This is one of the reasons that may prevent the widespread usage of such systems [26]. For instance, if we want to recognize the emotions of a human using these systems, the subjects need to use specialized bio-signal devices/sensors such as blood pressure sensor or heart rate sensor (electrocardiography (ECG) sensor) on their body [27]. However, these devices/sensors are not only invasive to subjects but also additionally much expensive [26].

Video Based Systems

Video-based systems designed to be born during normal expression to continually measure the mental state of humans in healthcare and daily life communication. A facial expression is a visible appearance of the emotional state, cognitive activity, intention, personality, and psychopathology of a person [28]; it plays a communicative role in interpersonal relations. Facial expressions express non-verbal communication cues in face-to-face interactions that might also complement speech by helping the listener to provoke the intended meaning of the spoken [29]. According to [30], facial expressions have a substantial effect on a listening interlocutor and it has an amazing contribution (55%) in our daily life communication than of the voice intonation (38%) and spoken words (7%). For instance, in healthcare domain, such systems could help to identify the mood of a patient, especially in the case of physically disabled people. A paralyzed patient whose body is critically attacked by paralysis is completely unable to speak and the only way to understand him might be through facial expressions. An FER system could identify the message that the patient is delivering through expressions. Such a system could be trained to identify correctly the moving parts of a patient face and identify whether the subject is happy, sad, angry, or displaying some other expressions. Similarly, for stroke patients, it is necessary to identify their expressions and generate guidelines for their care and protection.

1.2 Problem Statement

Different FER systems have been proposed; however, recognizing human facial expressions accurately is still a major concern for such systems. Several factors can reduce the accuracy of an FER system. For example, two such factors are the lack of robust features and high similarity among different facial expressions that can result in high within-class variance and low between-class variance in the feature space. Due to this property, most of the expression classes merge with each other, making it very hard for a classifier to distinguish among different expressions. Moreover, no one clarify exactly what parts of the databases were used, what the training and testing protocols were utilized. Most of the systems were provided performances on one dataset; however, they did not provide recognition rates across different spontaneous datasets, having expressions captured under different lighting conditions; using subjects of different gender, age (infants, adults, elderly), makeup, and race.

Challenges in Human Facial Expression Recognition

There are several issues that directly impact the success of any given FER system. Factors which contribute to the complexity of the recognition task can explained as given below [15].

- The key differences in facial appearance and features among individuals' faces play an important role and cause significant consequences in FER system is facial analysis. These differences may include explicit features such as facial shape, gender, person's ethnicity, variant hair styles etc. For example, female or children have smooth/less textured skin, less hairs on their body, long/short eyebrows, and significant differences in internals of eyes such as black and brown etc. make it difficult to track an eye correctly. The facial features can be obscured because of wearing goggles, jewels, persons having beards, and normal facial differences between ethnicities such as East Asians and Africans. Moreover, in spite of facial explicit features and looks differences, the use of expressions and its intensity, ways of responses to a certain condition, and the degree of plasticity resulting in weak or strong expressiveness can also affect the accuracy of FER systems. Therefore, it is very important to consider all these variant features and develop systems that can work on such variety datasets using sophisticated state-of-the-art algorithms.

- Facial expressions in the real world occur in a realistic manner and they are always triggered because of some action units. These expressions can occur in a serial way showing explicit dependence on the successor expression or change in response to a specific action unit. Similarly, the transitions from various states of expressions may not always involve static/normal state as presumed in the traditional FER systems [15]. Generally, in FER systems it is assumed that each expression is singular and is followed by a steady/normal state/position, which is completely against the reality. To efficiently analyze the expressions, analyzing a stream of expressional data without intervention and training the systems on test-bed dataset that include action units in it that is a basic requirement of an FER system. For a system to be more comprehensive, it is needed to process the data that include additive and non-additive action units, single action units, and co-articulation effects.

- The intensity of facial expressions have been less exploited in the research community. Much work exist in this regard; however, they focus on recognizing the intensities of a face in expression based on type of a facial action. Although, it may seem right for some cases; however, it is really important to thoroughly investigate the process of recognizing the intensity of facial expressions properly and define the lowest and largest possible values for it. The techniques used for intensive expressions may not work properly for low intense expressions; therefore, all these necessary things to be taken care of.

- In each research philosophy, it is recommended and most welcomed to perform experimentation on real world datasets that consist of the real world subject representations. Similarly, for an efficient and realistic FER system, it is more convenient and recommended to work on both pose and spontaneous facial expressions. The expression pattern identification models such as hidden Markov models and neural networks work on temporal data; therefore, they are likely to perform poor due to the pose and spontaneous facial expression differences [15]. The existing FER systems are trained and evaluated using the dataset generated from a directed set of expressions only differing in its time and place [31]. While pose and spontaneous expressions are communicated using different pathways such as pyramidal and extra pyramidal [32].

- For each system, the test-bed is the most important fundamental and basic step to evaluate and prove the accuracy of the system. The test-bed includes the environment of its real operation where the camera setting, the background environment, weather conditions, and crowdy place. In general case, the image data used for FER systems include images with static mono backgrounds, directed expressions, and no environment complexity. Therefore, the expression recognition, features extraction, accuracy of face detection can greatly be affected because of the environment complexity. In the complex test-bed environment, the system requires to be robust enough to recognize the emotions of people interacting with others, varying background, and non-directed expressions occurring spontaneously will definitely result in sophisticated and robust techniques being developed.
- The quality of image significantly affects the performance and accuracy of an FER system. The images used for the dataset can critically affect its performance such as images captured with variant resolutions and brightness effect of the operating/capturing place. Similarly, the distance between the subject and the camera, properties of the camera digitizer, image dimensions, and its orientation are the key factors that need to be tackled while developing an FER system. These factors when measured less or more will affect the accuracy of FER system. For example, difference or variation in the face position in a sequenced data relative to ambient light can affect the expression analysis apparently. The algorithms that are tested and have shown good performance for the datasets that are straightly oriented or without orientation may be poor when confronted with variant conditions. For compressed images and varying conditions, the existing techniques are surely expected to perform poor because of not knowing the boundary conditions, and it is very difficult task to evaluate them. Hence, we need systems that are tested and evaluated against a variant dataset that include all the described features.
- Most of the existing datasets relied on different kinds of video cameras which might not be the case in real world. These datasets did not take into account the color features (i.e., gender, race, and age), and were collected in controlled environments with constant lighting

conditions. Some of the previous datasets did not considered if the subject worn glasses or if the subject had beard. In FER domain, the size of the face has a vital role. However, in most of the previous datasets, due to the predefined setup of the cameras, the size of the faces were constant that might not be applicable in real world. All the previous datasets have been collected in static scenarios, either in indoor environment or outdoor environment. In some of the datasets, the subjects have just a slight variation of the angle which might not be the case in real world. Therefore, we need a comprehensive dataset that consider most of the shortcomings of the existing datasets.

- Some parts of the face such as lips, nose, eyes, forehead, etc, have much contribution in expressions making. In real life, the facial images vary from time to time due to the position of the corresponding camera such as frontal and non-frontal due to which some parts of the face may appear partially like nose and eyes. If some of these parts are occluded by other objects then the extraction of the informative features will be much complex. Therefore, most of the systems do not have the capability to recognize the spontaneous expression in naturalistic scenarios under these observations.
- The systems must be unobtrusive and robust to the subjects, gender, age (infants, adults, elderly), makeup, and race.

Limitations of Existing Systems

Majority of the FER systems included [13,16,17,33–37] had been developed in order to recognize the human expressions. Though, these approaches provided significant performance on existing publicly available standard datasets. However, the performance of these systems degrade across multiple datasets. Some systems have contribution only in feature extraction such as [1,38–46]; one problem that can be associated with the use of these methods is the fact that they are very sensitive to variations in pose, illumination, occlusion, aging, and face rotation [47,48]. Furthermore, these techniques are poor at handling data where classes do not follow the Gaussian distribution. Also, these techniques do not work well in case of a small sample size [48]. Furthermore, complexity-wise, most of these techniques are much expensive because of considering the entire face, as this requires more memory [49]. Lastly, these methods work well mostly in a controlled environment [50]. While, some systems such as [16, 51–53] have contributions in classification module. However, most of the aforementioned classifiers were employed for frame-based classification. On the other hand, the most commonly used sequence-based classification method is the Hidden Markov Models (HMMs) [54,55]. HMMs have their own advantage in handling sequential data when frame-level features are used, whereas vector-based classifiers, such as GMMs, ANNs, and SVMs, fail to learn the sequence of the feature vectors.

Nevertheless, conventional HMMs are based on Markovian property, which presumes that the current state depends only on the previous state. Because of this assumption, labels of two contiguous states must hypothetically occur consecutively in the observed sequence. Unfortunately, this presumption is not always true in reality. Some other limitations of HMMs include their generative nature and the independence assumption between states and observations [56].

Furthermore, another limitation seen among most of these methods is that they were evaluated under settings that are far from real-life scenarios. It means that they only utilized publicly available datasets and did not consider the real world challenges in their respective systems. Since the beginning of research in FER, the focus has been on designing new and improved methodologies, and evaluating them using publicly controlled-settings datasets for the sake of a fair comparison. Moreover, the existing publicly available FER datasets are mostly pose-based and assume a predefined setup. The facial expressions in these datasets are recorded using a fixed camera deployment with a constant background and static ambient settings. In a real-life scenario, FER systems are expected to deal with changing ambient conditions, dynamic background, varying camera angles, different face size, and other human-related variations. Little or no effort has been put into designing a new dataset that is closer to real-life situations, probably because creating such a dataset is a very difficult and time consuming task. And this is where the contribution of this work lies.

In Summary, there are lots of work have already been done for automatic FER, but a robust FER system is yet to be developed: a system capable of providing high recognition accuracy not just for one dataset but across different datasets, having expressions captured under different lighting conditions; using subjects of different gender, race, and age.

Study Goal and Methodology

The goal of this thesis is to investigate on the potential impact of some of the most prominent technological and practical issues in the use of FER systems for the pose and spontaneous based expression recognition, to demonstrate the limitations of state-of-the-art solutions and to provide alternatives to cope with these effects. In this way, this work seeks to contribute to a better understanding of the needs of realistic expression-aware applications and aims to help paving the path to a new generation of FER systems readily available for their use in real-world. In order to achieve the aforementioned goals, the system has the following main contributions.

The system uses a method for feature extraction in which first the noise reduction is achieved by means of wavelet decomposition, followed by the extraction of facial movement features using optical flow. These features reflect facial muscle movements which signify static, dynamic, geometric, and appearance characteristics of facial expressions. Post feature extraction, feature selection is performed using Stepwise Linear Discriminant Analysis, which is more robust in contrast to previously employed feature selection methods for FER. In order to classify the expressions, we have introduced a novel hidden conditional random fields (HCRF) model, which is able to approximate complex distributions using a mixture of full covariance Gaussian density functions. We assessed the performance of the proposed techniques using publicly available standard pose and spontaneous datasets of facial expressions.

1.3 Contribution

In this thesis, we have made the following contributions:

- We have proposed a hierarchical recognition system to overcome the problem of high similarity among different expressions. Expressions were divided into three categories based on different parts of the face. At the first level, LDA was used with an HMM to recognize the expression category. At the second level, the label for an expression within the recognized category is recognized using a separate set of LDA and HMM, trained just for that category. However, this system utilized two level classification (two HMMs for each layer) which is computational wise much expensive.

- Therefore, a single-level FER system has been proposed in order to solve the limitation of the proposed hierarchical recognition system. In this system, we proposed two methods for feature extraction and classification respectively. For the feature extraction, a new technique based on the facial movement features is proposed. The technique is using symlet wavelet transform coupled with optical flow to get the facial movement features. The reason for using the wavelet transform is to diminish the noise before extracting the facial movement features. Even though the proposed feature extraction method extracts good features, there might be some redundancy among these features.

  Therefore, previously proposed a robust feature selection method called Stepwise Linear Discriminant Analysis (SWLDA) is applied to the selected feature space. SWLDA selects the most informative features taking the advantage of the forward regression model and removes irrelevant features by taking the advantage of the backward regression model. To the best of our knowledge, it is the first time that SWLDA is being utilized as a feature selection technique for FER systems.

- Furthermore, for the system, we have developed the improved version of the hidden conditional random fields (HCRF). The proposed HCRF is capable of approximating a complex distribution using a mixture of Gaussian density functions. This model inherits the advantages of the existing HCRF model and completely tackles the limitations of the existing work, we propose the use of HCRF algorithm that is able to explicitly utilize mixture of full covariance Gaussian mixture hidden conditional random fields (FCGM-HCRF).

- In order to show the accuracy and robustness of the proposed FER system, large scale experimentations are performed using multiple pose and spontaneous datasets of facial expressions.

1.4 Structure of the Dissertation

The thesis has been organized into seven chapters, as given below.

- Chapter 1 has presented a brief introduction of the concepts of human FER. It discussed the importance of FER and the factors that make it challenging. Moreover, the problem associated with the existing FER systems were summarized in this chapter as well. Finally, an overview of my contribution was given.

- Chapter 2 provides different previous methods for a typical FER system and its limitations such as methods for feature extraction and recognition. Also, this chapter provides a background detail about the human FER systems and its applications.
- Chapter 3 provides an overview of the proposed FER systems which solve the limitations of the existing FER systems. It also explains different methods that proposed for feature extraction and recognition modules respectively.
- Chapter 4 provides an explanation about the existing pose and spontaneous datasets of facial expressions. Moreover, the chapter presents the experimental setup for the systems that were proposed in this dissertation.
- Chapter 5 provides the corresponding experimental results for each system under different settings.
- Chapter 6 concludes the thesis and also provides future directions in this research area. The main contribution of the thesis is also highlighted in this chapter.

Chapter 2

Related Work

2.1 Overview

This chapter first presents the state-of-the-art of the general FER research from the applications point of view. After that the existing methods relating to the implementation of an FER system are presented. Regarding the methodology, face detection and extraction, feature extraction, feature selection, and classification are the most important components in a typical FER system. Therefore, in this chapter, the existing works related to these components are analyzed. We have strong contributions in each of these components which are the main foci of this dissertation.

2.2 Facial Expression Recognition Methodologies

A typical FER system consists of four main important components, preprocessing, feature extraction, feature selection, and recognition modules. In the preprocessing module, the noise is removed from the facial images, and the faces are located. Locating faces is important because it consists all of the important parts (such as eyes, forehead, and lips) that convey expressions [18]. Feature extraction deals with extraction the distinguishable features from each facial expression frame and quantizing them as discrete symbols [57]. Feature selection module is used for selecting a subset of relevant features from a large number of features extracted from the input data. Finally, in the recognition module, a classifier is first trained using the training data and then used to generate labels for the expressions in the incoming video data. As we have contribution in feature extraction and classification modules respectively. Therefore, in the following sections, a summary of the related works for these components are presented.

2.2.1 Face Detection and Extraction

Essentially, face detection is the first-step for a typical FER system, with the purpose of localizing and extracting the face region from the background. Some factors like illumination, pose, occlusion, and size of the image make it difficult for FER systems to accurately detect faces from expression video [58].

A variety of methods have been proposed in literature for face detection, which are explained below.

2.2.1.1 Appearance and Feature Based Methods

These [59] methods have been proposed for face detection; however, both of them have their own limitations. The performance of appearance-based methods is excellent in static environment; however, their performance degrades with the environmental change [35]. Therefore, feature-based methods were proposed to overcome the limitations of appearance-based approaches. These methods are more robust in illumination, pose, and size of the image than the appearance-based methods. However, a prior knowledge is required for these methods, i.e., at the time of implementation for these techniques, it is compulsory to decide randomly which intensity information will be important [60]. Due to these characteristics, these methods are known as heuristic techniques. Moreover, these methods have trouble in automatic feature detection [61].

2.2.1.2 Geometric Based Methods

The geometric based methods [62] have some common limitations such as in these approaches, parts of a face like mouth, eyes, and nose are positioned with their attributes and their reciprocal relationships. Face is recognized by calculating the distance, angles, and areas between these parts of the face. These approaches are quite sufficient for small databases with stable lighting condition and stable viewpoint [63]. However, their performance degrades with the variation in lighting conditions and viewpoint [63].

2.2.1.3 Knowledge Based Methods

Similarly, knowledge-based approaches [64] are the rule-based techniques that convert human knowledge of what constitutes a typical face to capture the relationships between the facial features [65]. However, it is very hard for these approaches to build an appropriate set of rules. If the rules are too general then there could be several false positives, or there could be false negatives if the rules are in too detail [66]. Moreover, these approaches are incapable of finding faces in complex images [66].

2.2.1.4 Template Based Methods

Template-based approach [67] is a simple process that has been widely employed to locate the human face in the input image. However, this method is very sensitive to pixel misalignment in sub-image areas and depends on facial component detection [68].

2.2.2 Feature Extraction

According to the face descriptors, there are two types of features: global features and local features. For global features, the features are extracted from the entire face, whereas for local features, parts of the face, such as eyes, mouth, nose and forehead are used.

2.2.2.1 Global Features

Global feature extraction methods are known as holistic methods. These include Nearest Features Line-based Subspace Analysis [38], Eigenfaces and Eigenvector [39, 69] and [40], Fisherfaces [41], global features [42], Independent Component Analysis (ICA) [43, 70], Principal Component Analysis (PCA) [1, 44, 71], frequency-based methods [45], Gabor wavelet [46]. One problem that can be associated with the use of these methods is the fact that they are very sensitive to variations in pose, illumination, occlusion, aging, and rotation changes of the face [47, 48]. Furthermore, these techniques are poor at handling data where classes do not follow the Gaussian distribution. Also, these techniques do not work well in case of a small sample size [48]. Furthermore, complexity-wise, most of these techniques are much expensive because of considering the entire face, as this requires more memory [49]. Lastly, these methods work well mostly in controlled environments [50].

2.2.2.2 Local Features

On the other hand, local feature extraction methods compute local descriptors from parts of the face and then integrate this information into one descriptor. These include Local Feature Analysis (LFA) [72], Gabor features [73], Non-negative Matrix Factorization (NMF) and Local non-negative Matrix Factorization (LNMF) [74], and Local Binary Pattern (LBP) [16, 75]. Among these methods, LBP is the most commonly employed feature extraction technique. However, LBP does not provide the directional information of the facial frame [76].

Some recent studies have tried to solve the limitations of LBP. These methods include Local Transitional Pattern (LTP) [37], Local Directional Pattern (LDP) [33], Local Directional Pattern Variance (LDPv) [77]. Most of these methods exploited other information instead of employing intensity to overcome the problems due to noise and illumination change [35]. However, performance of these methods still degrade in non-monotonic illumination change, noise variation, change in pose, and expression conditions [35]. Another commonly used local feature extraction method for expression recognition is Local Fisher Discriminant Analysis (LFDA) [36]. But, LFDA fails to determine the essential assorted structure when face image space is highly nonlinear [78]. Furthermore, authors of [79] employed pixel and color segmentation for feature extraction to detect facial expressions. However, the performance of this approach also degrades with variation in illumination.

2.2.3 Feature Selection

Feature selection module helps reducing the dimensions of the feature space by selecting only the distinguishable features. Please see Figure 2.1 in order to appreciate the use of a feature selection technique in an FER system. Figure 2.1 shows 3D feature plots for six expression classes.

It can be seen from Figure 2.1 that the feature values for the six classes are highly merged, which can result in a high misclassification rate. It is because the use of inappropriate coefficients results in high within-class differences and low between-class differences.

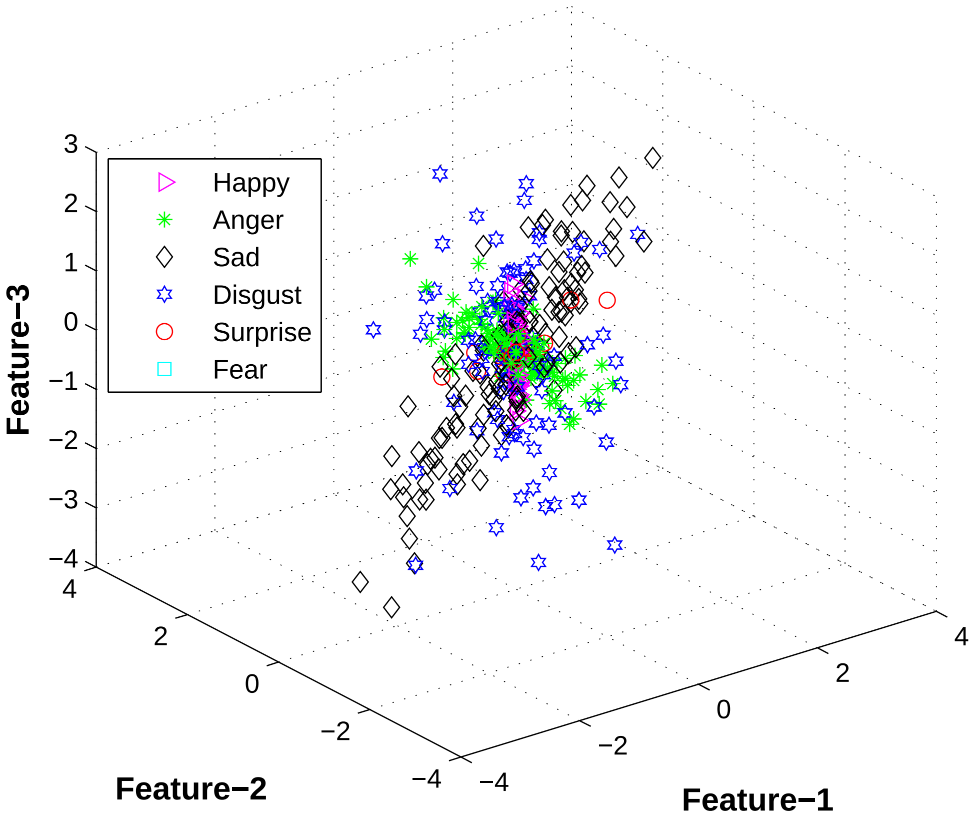


Figure 2.1: 3D-feature plot of Cohn-Kanade dataset for six different types of facial expressions, where each expression has twelve expression frames.

In order to solve the above problem, a method is required to address the aforementioned problem and to reduce the dimension space and to increase the class separability. This idea is employed by various feature selection methods, mainly Principal Component Analysis (PCA) [80], Linear Discriminant Analysis (LDA) [81], kernel discriminant analysis (KDA) [82], and Generalized Discriminant Analysis (GDA) [83].

PCA has poor discriminating power [84]. LDA-based methods suffer from limitations that their optimality criteria are not directly associated to the classification capability of the achieved feature representation [85]. Zia *et al.* [55] applied LDA to the extracted feature space to improve the class separation among different classes with the assumption that the variance is distributed uniformly among all the classes. However, this is not the case. For example, expressions like happy and sad are very similar to each other but can easily be distinguished from anger and fear (another pair with high similarity). Likewise, KDA does not have the capability to provide better performance in the case if the face images of the same subjects are scattered rather than dispersed as clusters [86]. Likewise, the solution of GDA might not be stable and perhaps is not optimal in terms of the discriminant ability if there is small sample sized training data [87].

2.2.4 Recognition

There are two approaches that utilized for expression recognition. One is frame-based recognition and the other is sequential-based recognition. Both are explained as.

2.2.4.1 Frame Based Recognition

As for the recognition module, a large number of methods have been employed for accurate expression classification. In [51], authors exploited artificial neural networks (ANNs) in order to classify different facial expressions and achieved a 73% recognition rate. However, ANN is a black box and has incomplete capability to explicitly categorize possible fundamental relationships [88]. Besides, ANNs may take long time to train and may trap in a bad local minima. Moreover, the authors of [89] and [16] employed support vector machines (SVMs) for their FER system. But, in SVMs, the observation probability is calculated using indirect techniques; in other words, there is no direct estimation of the probability [90]. Furthermore, SVMs simply disregard temporal dependencies among video frames, and thus each frame is expected to be statistically independent from the rest. Similarly, the authors of [52] and [53] utilized Gaussian mixture models (GMMs) to recognize different types of facial expressions. But facial features could be very sensitive to noise; therefore, fast variations in facial frames cannot be modeled by GMMs and might cause misclassification [15].

2.2.4.2 Sequential Based Recognition

Most of the aforementioned classifiers were employed for the frame-based classification. On the other hand, the most commonly used method is the Hidden Markov Models (HMMs) [54, 55] which is widely utilized for sequence-based classification. HMMs have their own advantage in handling sequential data when frame-level features are used, whereas vector-based classifiers, such as GMMs, ANNs, and SVMs, fail to learn the sequence of the feature vectors.

Nevertheless, conventional HMMs are based on Markovian property, which presumes that the current state depends only on the previous state. Because of this assumption, labels of two contiguous states must hypothetically occur consecutively in the observed sequence. Unfortunately, this presumption is not always true in reality. Some other limitations of HMMs include their generative nature and the independence assumption between states and observations [56]. A non-generative model such as maximum entropy Markov model (MEMM) was developed in order to resolve the limitations of HMM, and it produced better results compared to HMM [91]. However, MEMM has a commonly known drawback called the "*label bias problem*".

Conditional random fields (CRF) [56] and HCRF [92], the generalizations of MEMM, were then proposed in order to take the full advantage of MEMM and to solve the "*label bias problem*" [56]. HCRF extends the capability of CRF with hidden states making it able to learn hidden structure of the sequential data. Both of them use global normalization instead of per-state normalization. Thus, they allow weighted scores, making the parameter space larger than those of MEMM and HMM. The following discussion provides the underlying theory of HCRF, and analyzes the limitations in their existing implementations.

We consider a task of mapping from inputs X to labels $Y \in \Gamma$, for instance, $\Gamma = \{$happy, anger, sad, surprise, disgust, fear$\}$ in an FER problem. Each input X is a sequence of T frames, $X = x_1, x_2, ..., x_T$. The training set contains N pairs $(X_i, Y_i), i = 1, 2, ..., N$. In a Q-state HCRF, the conditional probability of a class label Y given input X and set of parameters of the model Λ is computed as

$$p\left(Y|X;\Lambda\right) = \frac{\sum\limits_{\overline{S}} \exp\left\{\Lambda \cdot f\left(Y, \overline{S}, X\right)\right\}}{z\left(X, \Lambda\right)}, \tag{2.1}$$

where

$$z\left(X, \Lambda\right) = \sum_{Y'\bar{S}} \exp\left\{\Lambda \cdot f\left(Y', \overline{S}, X\right)\right\}, \tag{2.2}$$

is the normalization factor to guarantee the sum-to-one rule of the conditional probability, where, Y' is the predicted label for the sequence, and $\overline{S} = \{s_1, s_2, \ldots, s_T\}$ is a sequence of hidden states. Each $s_i, i = 1, 2, ..., T$, can have an integer value from 1 to Q, the number of states, Λ is the parameter vector and $f\left(Y, \overline{S}, X\right)$ is known as the feature vector that consists of the following sufficient statistics used by the model.

$$\underset{y'}{\overset{Pr}{f}}\,(Y, \overline{S}, X) = \delta\,(y = y'), \;\; \forall y' \in Y, \tag{2.3}$$

$$\underset{ss'}{\overset{Tr}{f}}\,(Y,\, \overline{S},\, X) = \sum_{t=1}^{T} \delta(s_{t-1} = s)\,\delta(s_t = s'), \;\; \forall \{ss'\} \in \overline{S}, \tag{2.4}$$

$$\underset{s}{\overset{Occ}{f}}\,(Y, \overline{S}, X) = \sum_{t=1}^{T} \delta(s_t = s), \;\; \forall s \in \overline{S}, \tag{2.5}$$

$$f_s^{M_1}\,(Y, \overline{S}, X) = \sum_{t=1}^{T} \delta(s_t = s) x_t, \;\; \forall s \in \overline{S}, \tag{2.6}$$

$$f_s^{M_2}\,(Y, \overline{S}, X) = \sum_{t=1}^{T} \delta(s_t = s) x_t^2, \;\; \forall s \in \overline{S}, \tag{2.7}$$

where $\delta\,(s = s')$ is equal to one when $s = s'$, otherwise equal to zero. Thus, $\underset{y'}{\overset{Pr}{f}}\,(Y, \overline{S}, X)$ in (2.3) tracks the number of times the predicted labels are equal to the original labels. Similarly, $\underset{ss'}{\overset{Tr}{f}}\,(Y,\, \overline{S},\, X)$ in (2.4) determines the number of times the transition ss' occurs in $\overline{S}$, and this process is repeated for the entire state sequence. Likewise, $\underset{s}{\overset{Occ}{f}}\,(Y, \overline{S}, X)$ in (2.5) counts the occurrence of the state s. The first and second moments $f_s^{M_1}$ and $f_s^{M_2}$ in (2.6) and (2.7) respectively are the sum and sum of the squares of observations that align with the state s. It is to be noted that the term feature vector does not refer to the input features, but refers to the vector of sufficient statistics used by the model. Latter is referred to as the observation vector. The choice of the feature vector determines the dependencies of the HCRF model.

It can be seen from the above equations that with some specific set of parameters (Λ), HCRF's dependencies are similar to those of HMM. For example with above feature vector, the diagonal-covariance Gaussian distribution can be defined as

$$\underset{y'}{\overset{Pr}{\Lambda}} = \log(u_{y'}), \quad \forall y' \in Y, \tag{2.8}$$

$$\underset{ss'}{\overset{Tr}{\Lambda}} = \log(A_{ss'}), \quad \forall \{ss'\} \in \overline{S}, \tag{2.9}$$

$$\underset{s}{\overset{Occ}{\Lambda}} = -\frac{1}{2}\left(\log\left(2\pi\sigma_s^2\right) + \frac{\mu_s^2}{\sigma_s^2}\right), \tag{2.10}$$

$$\Lambda_s^{M_1} = \frac{\mu_s}{\sigma_s^2}, \tag{2.11}$$

$$\Lambda_s^{M_2} = -\frac{1}{2\sigma_s^2}, \tag{2.12}$$

where u in (2.8) is the prior distribution of Gaussian-HMM, and A in (2.9) is a transition matrix, then the numerator of the condition probability can be written as

$$\begin{gathered}\sum_{\overline{S}} \exp\left\{\Lambda \cdot f\left(Y, \overline{S}, X\right)\right\} = \\ \sum_{\overline{S}} u(s_1) \prod_{t=1}^{T} A\left(s_{t-1}, s_t\right) N\left(x_t^2, \mu_{S_t}, \sigma_{S_t}\right),\end{gathered} \tag{2.13}$$

where N denotes the Gaussian distribution. The conditional probability of X given Y is computed with a Gaussian-HMM by (2.13) that has a prior distribution u, and a transition matrix A.

A more generalized version of the HCRF model has been proposed by [2] in order to handle more complex distributions using a linear mixture of Gaussian density functions, and is given as

$$p\left(Y|X;\Lambda\right) = \frac{\sum\limits_{\overline{S}} \sum\limits_{m=1}^{M} \exp\left\{\Lambda \cdot f\left(Y, \overline{S}, m, X\right)\right\}}{z\left(X, \Lambda\right)}, \tag{2.14}$$

where M is the number of components in the Gaussian mixture.

Although, there are some existing works that employed the above HCRF model and showed good results [93, 94]. They did not address and overcome the limitations of the model. As we can see in the above equation, that the model can only utilize diagonal-covariance Gaussian distribution. In other words, the variables (columns of $x_i, i = 1, 2, ..., N$) are assumed to be pair-wise independent. Hereafter, we call this model *diagonal covariance Gaussian mixture hidden conditional* *random fields* (DCGM-HCRF). In addition, equations (2.10), (2.11), and (2.12) imply that with a particular set of values, the observation density at each state will converge to Gaussian form. Unfortunately, there is algorithm that could guarantee this convergence. Therefore, these assumptions may result in a decrease of accuracy.

In order to inherit the advantages of HCRF model and completely tackle the limitations of the existing work, we propose the use of HCRF algorithm that is able to explicitly utilize mixture of *full covariance Gaussian mixture hidden conditional random fields* (FCGM-HCRF).

2.3 Applications of Facial Expression Recognition

Expressions play a vital role in our daily communications, and recent years have witnessed a great amount of work being done to develop accurate and reliable FER systems. Such systems can be employed in many applications, which are described in subsequent subsections.

2.3.1 Human Computer Interaction

In human computer interaction (HCI), human faces are used to decode the well-known facial expressions such as happy, sad, anger, etc., face-to-face communication is considered as breakthrough towards perceptual primitives [4]. Expression recognition does not only help in identifying the affective state of the face, but also helps in identifying the cognitive state. Human interaction is distinguished to be identified by using two channels including the one, which transmits explicit messages about any or nothing and the one which transmits implicit messages about the action performer. Based on the detailed survey, the emotion recognition in HCI can efficiently be utilized through signal analysis for speech, faces, representation of emotions, and emotion-oriented representations [3, 5, 95].

2.3.2 Robotics

Emotion recognition currently has been widely adopted in the artificial intelligence to train the humanoid robots, affective sensitive computing in HCI [5], behavioral scenarios, and video games. The robotic technology can easily adopt the emotion recognition technology to detect the interacting human being facial expressions in real-time and respond to it. Using artificial intelligence techniques, robots can be trained to show the expressions, detect the expressions, and respond to the human expressions. In the field of healthcare, robotics have been introduced to perform surgeries, treatment, and hopefully provide patient-care and management in the near-future. This is one of the key area of facial recognition in robotics [96].

2.3.3 Biometric Authentication

Expression recognition technology is widely adopted in the field of security and also in the field of recognizing the human mood or behavior. To provide full-proof security to the world-wide banking systems, data management systems, security surveillance systems, nuclear management, and many other systems; biometric authentication are used to secure high security installments. These biometric authentication systems include facial recognition, voice recognition, and other mechanisms of security. Although, specific emotions may not be necessary to authenticate the user to the secure system such as in FER systems in healthcare etc., it is still necessary to identify the user face in any orientation and authenticate the right user of the system. Face detection is one of the key features in spite of the emotion recognition, and is quite tricky because of finding the exact match and differentiating between similar faces which is an obvious case [3, 7, 97, 98].

2.3.4 Healthcare Domain

Healthcare domain is one of the key domain in facial emotion recognition especially for the persons with a disability and abnormality. The current existing technologies recognize the patient emotions from physiological sensors, electrocardiogram signals, and other implicit signals which help to identify the user state and emotions [99, 100]. Those patients with no face expression can be understood by the help of these emotion recognition technologies. Emotions of human being can be identified using its implicit signals as described earlier to detect implicit message such as cognitive states and explicit signals of facial expressions to detect the facial emotions. Both can be used in healthcare to help the doctors in order to understand the patients. Hence, in such domain, emotion recognition helps to improve the quality of care.

2.3.5 Investigative Analysis

In law enforcement, advanced investigations with forensic operations and special treatment and security environments, there is important role of FER. Face recognition and identity resolution is one of the most important technology and breakthrough in the criminal investigation, identification, and tracking. In the investigative process, powerful 2D or 3D transformations of images and videos can be extracted and linked to reach to the final required target. Various images can be processed faster and analytical comparison can also be performed to extract precise expression recognition and criminal screening [101].

2.3.6 Gaming

Advanced feature extraction and image facial recognition give totally new dimension to the gaming field. As a primary example, Kinect of Microsoft has given advanced motion capabilities to the gaming based on Xbox360 with expression recognition and avoiding the hardware facilities. Similarly, a program launched by Viewdle deals on how to recognize whether a user is a vampire or a human being based on image recognition and face recognition [102]. Facial recognition in future is expected to be used for real-time game playing without hardware touch and will support head motions and facial expressions to react and act in the games. It will also be used in driving cars with the head movement and facial expressions.

2.3.7 Image Searching

One of the very recent and publicly available usage and applications of expression recognition released by Google includes Google image search. This option lets you search images in Google through upload image. It uses novel image recognition techniques to search images. Google image search enables the photographers to know where their images have been used and to know where they have been referred. In a recent image search report, it has been stated that Google image search is still patchy; however, it is expected to improve [103].

2.3.8 Solving Puzzles

Besides the image search, comparison, identification, facial expressions, and authentication; facial recognition is also recently used in the problem solving and game playing. Solving puzzles is one of the mostly played games and Googles Google Goggles is one of the recent achievement in solving the Suduko puzzle. This application lets a user solve the puzzle using state of the art image recognition techniques. It gives full solution of the solution once you provide in image of the puzzle [104].

Chapter 3

Proposed Methodology for Facial Expression Recognition

3.1 Overview

Human FER has emerged as an important research area over the last two decades. In order to accurately recognize expressions, FER systems require automatic face detection followed by the extraction of robust features from important facial parts. Furthermore, the process should be less susceptible to the presence of noise, such as different lighting conditions and variations in facial characteristics of subjects, and a classifier is required to accurately classify the expressions.

Accordingly, this chapter presents two FER systems. First one is the hierarchical FER system which has the capability to overcome the problem of high similarity among different expressions. Numerous techniques have been developed and validated for the purpose of feature extraction for FER systems. Among them, ICA is the mostly commonly used methods; therefore, in the proposed hierarchical system, we have decided to use ICA for feature extraction to extract the local features. Moreover, for dimension reduction, we utilized, linear discriminant analysis (LDA), and for recognizing the expressions, we used hidden Markov model (HMM). In the proposed hierarchical system, an expression is classified into one of these three categories (such as lips-based, lips-eyes-based, and lips-eyes-forehead-based) at the first level. At the second level, classifier (trained for the recognized category) is employed to give a label to the expression within that category. However, in this system, we used two level classification (two HMMs for each layer) which is computational wise much expensive.

Therefore, we have proposed second FER system that solved the limitation of the hierarchical system. In this system, we proposed an unsupervised technique based on active contour (AC) model for automatic face detection and extraction. In this model, a combination of two energy functions: Chan-Vese (CV) energy [105] and Bhattacharyya distance [106] functions are employed, which not only minimizes the dissimilarities within a face, but also maximizes the distance between the face and the background. For the feature extraction, we have proposed a robust feature extraction technique based on the facial movement features. The technique is based on symlet wavelet transform coupled with optical flow to get the facial movement features. The reason for using the wavelet transform is to diminish the noise before extracting the facial movement features. Furthermore, for the recognition, we proposed the improved version of the hidden conditional random fields (HCRF) that solves the limitations of the existing HCRF by utilizes full covariance Gaussian density function. Large scale experimentation is performed using multiple datasets to show the robustness of the proposed FER systems.

3.2 Hierarchical Recognition System

The architectural diagram for the hierarchical recognition system is given in Figure 3.1.

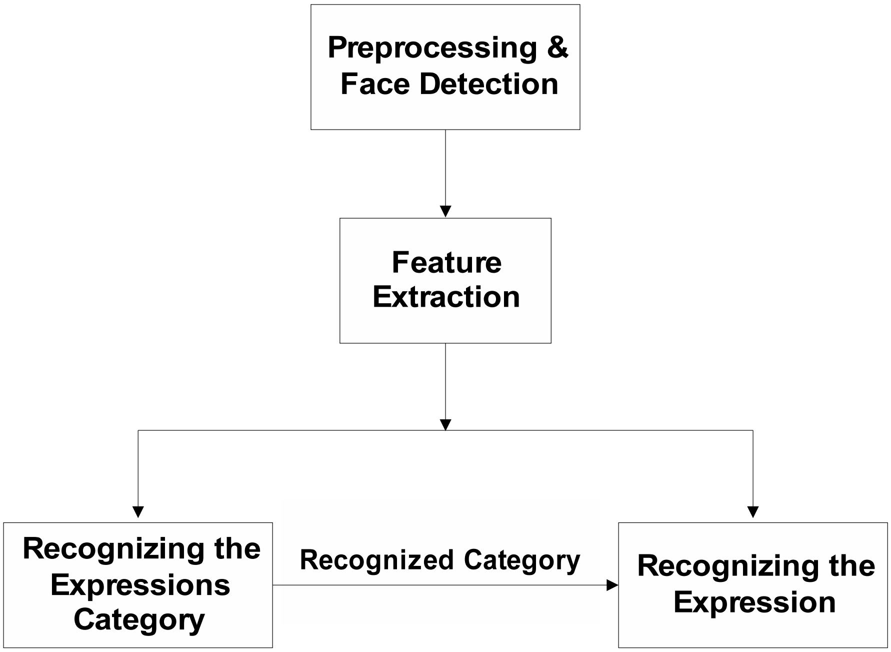

Preprocessing & Face Detection
Feature Extraction
Recognizing the Expressions Category
Recognized Category
Recognizing the Expression

Figure 3.1: Architectural diagram for the hierarchical recognition system.

3.2.1 Preprocessing

In preprocessing module, different environmental illuminations and lighting effects are diminished in order to increase the recognition accuracy, using several techniques like morphological filters, homomorphic filters, or median filters. This module is explained as below.

3.2.1.1 Diminishing Lighting Effects

Several techniques exist in literature to diminish such illumination effects, such as histogram equalization (HE) and local histogram equalization (LHE). However, HE produces unwanted artifacts and a washed-out look, so it is not recommended [107]. LHE, though better than HE, causes over-enhancement, and sometimes it produces checkerboards of the enhanced image [108]. Therefore, we used a new method called global histogram equalization (GHE) [109] for this purpose. GHE improves the image quality by increasing the dynamic range of the intensity using the histogram of the whole image. It obtains the scale factor from the normalized cumulative distribution of the brightness distribution of the original image and multiplies this scale factor by the original image to redistribute the intensity [109]. GHE finds the running sum of the histogram values and then normalizes it by dividing it by the total number of pixels. This value is then multiplied by the maximum gray-level value and then mapped onto the previous values in a one-to-one correspondence [109]. For more information, please refer to [109].

As described earlier, numerous techniques have been developed and validated for the purpose of feature extraction, dimension reduction, and classification for FER systems. Among them, independent component analysis (ICA), linear discriminant analysis (LDA) and hidden Markov model (HMM) are the most wisely used methods, and its performances has already been validated in [55]. Therefore, we decided to use ICA for feature extraction, LDA for dimension reduction, and HMM for classification. All are explained below.

3.2.2 ICA based Feature Extraction

Independent component analysis (ICA) is a technique used to seek independent components from multivariate statistical data. ICA assumes that the underlying sources are linearly mixed and statistically independent. General implementations of ICA can be found in the literature [110, 111].

If we assume that the sources are denoted by $S(t) = [S_1(t), S_2(t), ..., S_m(t)]^T$ and the multichannel observations are denoted by $X(t) = [X_1(t), X_2(t), ..., X_m(t)]^T$, then the linear mixture can be represented by

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \text{......} + \ a_{jn}s_n \quad for\ all\ j \tag{3.1}$$

or we can write this as

$$X = As \tag{3.2}$$

where the matrix A of size nxm represents linear memory-less mixing channels. The statistical model presented in Eq. 3.2 is called the independent component analysis or ICA model. The ICA model is a generative model, i.e., it describes how the observed data are generated by a process of mixing the components s_k. The independent components are latent variables, i.e., they cannot be directly observed. Also, the mixing matrix is assumed to be unknown. We observe the random vector x, and we must estimate both A and s by using it.

ICA starts with the very simple assumption that the components s_k are statistically independent. It will be seen below that we must also assume that the independent components have non-Gaussian distributions. However, in the basic model, we do not assume these distributions are known, but note that if they are known, the problem is considerably simplified. For simplicity, we assume that the unknown mixing matrix is square, but this assumption can sometimes be relaxed. Then, after estimating the matrix A, we can compute its inverse, W, and obtain the independent component simply by:

$$S = Wx \tag{3.3}$$

where $W = [w_1, w_2, ..., w_n]$ is the de-mixing matrix of size mxn. In general, ICA assumes that the number of channels are equal to the number of independent sources, i.e., $n = m$. Thus, n channels of data are decomposed into n ICs.

We notice that, in the ICA model, the time index t is dropped, as seen in Eqs. 3.2 and 3.3. We assume that each mixture x_j and each independent component s_k is a random variable, instead of a proper time signal.

3.2.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. LDA produces an optimal linear discriminant function that maps the input into the classification space on which the class identification of the samples is decided. LDA easily handles the case in which the within-class frequencies are unequal. The within S_W and between S_B class comparison is done by using the following equations.

$$S_B = \sum_{i=1}^{c} V_i \left(\overline{m}_i - \overline{\overline{m}}\right) \left(\overline{m}_i - \overline{\overline{m}}\right)^T \tag{3.4}$$

$$S_W = \sum_{i=1}^{c} \sum_{m_k \in C_i} \left(m_k - \overline{m}_i\right) \left(m_k - \overline{m}_i\right)^T \tag{3.5}$$

where V_i is the number of vectors in the i_{th} class C_i, and c is the number of classes, and in our case, c represents the number of facial expressions. Also, $\overline{\overline{m}}$ represents the mean of all the vectors, $\overline{m}$ is the mean of the class C_i, and m_k is the vector of a specific class. The optimal discrimination projection matrix D_{opt} is chosen from the maximization of the ratio of determinant of the between and within-class scatter matrices as

$$D_{opt} = \arg\max_{D} \frac{\left|D^T S_B D\right|}{\left|D^T S_W D\right|} = [d_1, d_2, \ldots, d_t]^T \tag{3.6}$$

where D_{opt} is the set of discriminate vectors of S_W and S_W corresponding to the $c-1$ largest generalized eigenvalues λ. The size of D_{opt} is $t \times r$, where $t \leq r$, and r is the number of elements in a vector. Then,

$$S_B d_i = \lambda_i S_W d_i, i = 1, 2, ..., c-1 \tag{3.7}$$

where the rank of S_B is $c-1$ or less, and hence, the upper bound value of t is $c-1$. Thus, LDA maximizes the total scattering of the data while minimizing the within scattering of the classes. For more details on LDA, please refer to [112].

3.2.4 Hidden Markov Model

Hidden Markov model (HMM) is the most commonly used method for sequential data (facial expressions) classification, which provides a statistical model λ for a set of observation sequences. These observations are called frames in FER domain. A typical HMM has a sequence of observations of length T (i.e., $T = O_1, Q_2, ..., O_T$), a sequence of states S (i.e., $S = S_1, S_2, ..., S_N$, where N is the number of states in the model), and the time t for each state is denoted by Q (such that $Q = q_1, q_2, ..., q_N$). The states are connected by arcs, as shown in Figure 3.2, and each time, when a state j is entered, an observation is generated according to the multivariate Gaussian distribution $b_j(O_t)$ with the mean value μ_j and covariance matrix V_j correlated with that state. There is also transition probabilities correlated with them such that the probability a_{ij} is the resultant transition probability from state i to state j. The initial model probability for the state j is Π_j. An HMM can be defined by this set of parameters, such as $\lambda = A, B, \Pi$, where A indicates the probability of the state transition (such that $A = a_{ij}$, $a_{ij} = Prob(q_{t+1} = S_j | q_t = S_i)$, $1 \leq i$, $j \leq N$), where B represents the probability of observations (such that $B = b_j(O_t)$, $b_j = Prob(O_t | q_t = S_j) 1 \leq j \leq N$), and the initial state probability is indicated by Π (such that $\Pi = \Pi_j$, $\Pi_j = Prob(q_1 = S_1)$). All the equations are based on the work by [113] and make use of the initial state probability distribution.

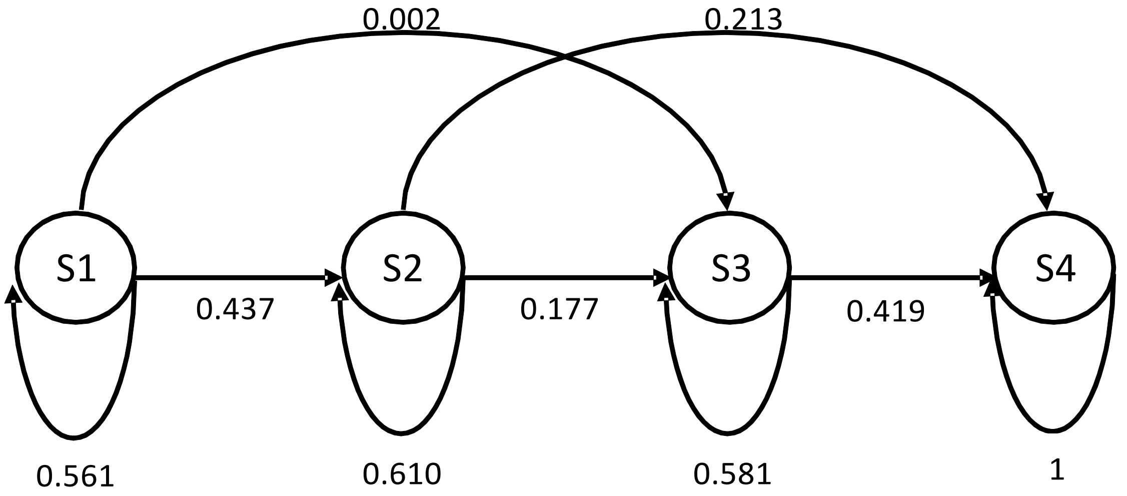


Figure 3.2: After training, the arrangement of HMM and transition probabilities for happy facial expression.

In the training step, for a given model λ, the multiplication of each transition probability by each output probability at each step t provides the joint likelihood of a state sequence Q and the corresponding observation O. This likelihood $P\left(O|\lambda\right)$ can be evaluated by summing over all possible state sequences:

$$P\left(O|\lambda\right) = \sum_{Q} P(O, Q|\lambda) \tag{3.8}$$

A simple procedure for finding the parameters λ that maximize the above equation in HMM, introduced in [114], depends on forward and backward algorithms $\alpha_t(j) = P(O_1...O_t, q_t = j|\lambda)$ and $\beta_t(j) = P(O_(t+1)...O_T|q_t = j, \lambda)$, respectively, such that these variables can be initiated inductively by the following processes:

$$\alpha_1\left(j\right) = \pi_j b_j\left(O_1\right), 1 \leq j \leq N \tag{3.9}$$

$$\beta_T\left(j\right) = 1, 1 \leq j \leq N \tag{3.10}$$

During testing, the appropriate HMM can then be determined by mean of likelihood estimation for the sequence observations O calculated based on the trained λ as

$$P\left(O|\lambda\right) = \sum_{i=1}^{N} \alpha_T\left(i\right) \tag{3.11}$$

The maximum likelihood for the observations provided by the trained HMM indicates the recognized label. The following formula has been utilized to model HMM (λ).

$$\lambda = (O,\, Q,\, \pi) \tag{3.12}$$

where O is the sequence of observations (i.e., $O_1, Q_2, ..., O_T$) and each state is denoted by Q (such as $Q = q_1, q_2, ..., q_N$), where N is the number of states in the model, and π is the initial state probabilities. The parameters that are used to model HMM (λ) for all experiments were 64, 4, and 4 respectively. These values have been selected by performing multiple experiments. For more details on HMM, please refer to [115].

3.2.5 Recognizing the Expression Category

This work is based on the theory that different expressions can be grouped into three categories based on the parts of the face that contribute most toward the expression [116–118]. This classification is shown Table 3.1

Table 3.1: The classified categories and facial expressions recognized in this system.

Category	**Facial Expressions**
Lips-Based	Happy
	Sad
Lips-Eyes-Based	Surprise
	Disgust
Lips-Eyes-Forehead-Based	Anger
	Fear

Lips-based expressions are those in which the lips make up the majority of the expressions. In lips-eyes-based expressions, both lips and eyes contribute in the expressions. In lips-eyes-forehead expressions, lips, eyes, and eyebrows or forehead have equal roles. In this hierarchical recognition FER system, an expression is classified into one of these three categories at the first level. At the second level, classifier (trained for the recognized category) is employed to give a label to this expression within this category.

At the first level, LDA was firstly applied to the extracted features from all the classes and an HMM was trained to recognize the three expression categories: lips-based, lips-eyes-based, or lips-eyes-forehead-based expressions. The LDA-features for these three categories are shown in Figure 3.3. A clear separation could be seen among the categories, and this is why the proposed hierarchical recognition scheme achieved 100% recognition accuracy at the first level.

3.2.6 Recognizing the Expressions

As mentioned earlier, once the category of the given expression has been determined, the label for the expression within the recognized category is recognized at the second level. For this purpose, LDA was applied separately to the feature space of each category and the result was used to train

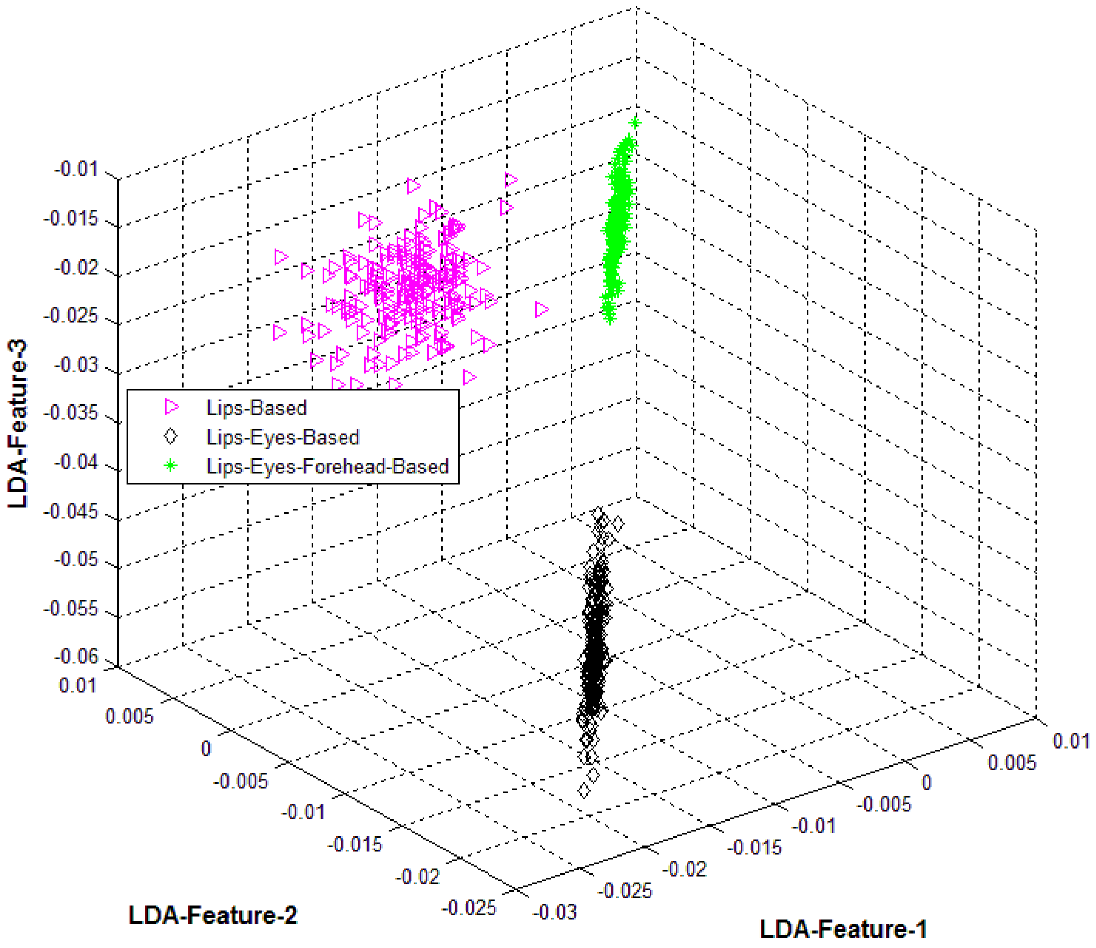


Figure 3.3: 3D feature plots of the proposed hierarchical recognition system after applying LDA at the first level for the three expression-categories such as lips-based, lips-eyes-based, or lips-eyes-forehead-based expressions. It can be seen that at the first level, the proposed hierarchical recognition system achieved 100% classification rate in expressions categories classification.

three HMMs, one HMM per category. Collectively, the overall classification for all the expression classes are shown in Figure 3.4.

These feature plots indicate that applying LDA to the features of three categories separately provided a much better separation as compared to single-LDA via single-HMM approach (see Figure 3.5). The single-LDA via single-HMM approach means that instead of applying LDA separately to each expression category and using separate HMMs for these categories, LDA is applied only once, to the features of all the classes. Figure 3.6(a) shows the basic steps of HMM training for a facial expression, where M indicates the the number of video clips for an expression, T presents the number of frames in each video, and O represents the observation symbol sequence for video. While, Figure 3.6(b) shows the testing process of a facial expression image sequence utilizing the likelihoods (i.e., L) of all trained facial expression HMMs (i.e., H).

3.3 Limitation of the Hierarchical Recognition System

Eventually, we envision that the proposed hierarchical system will be employed in smartphones. Even though our system showed high accuracy, it employs two-level-recognition with HMMs used at each level. This might become a complexity issue, especially when used in smartphones. One solution could be to use a lightweight classifier such as k-nearest neighbor (k-NN) at the first level; however, k-NN has its own limitations such as it is very sensitive to the presence of inappropriate parameters and sensitive to noise as well. Therefore, it can have poor performance in a real-time environment if the training set is large. In summary, new feature extraction, feature selection, and recognition methods should require in order to extract and select the most prominent features from the facial frames which maintain the same on a single layer classification.

3.4 A Robust FER System

3.4.1 Background

We developed an accurate and a robust FER system in order to solve the limitations of the hierarchical recognition system. In this system, we have proposed a new feature extraction technique based on wavelet transform coupled with optical flow. Moreover, we have proposed the improved

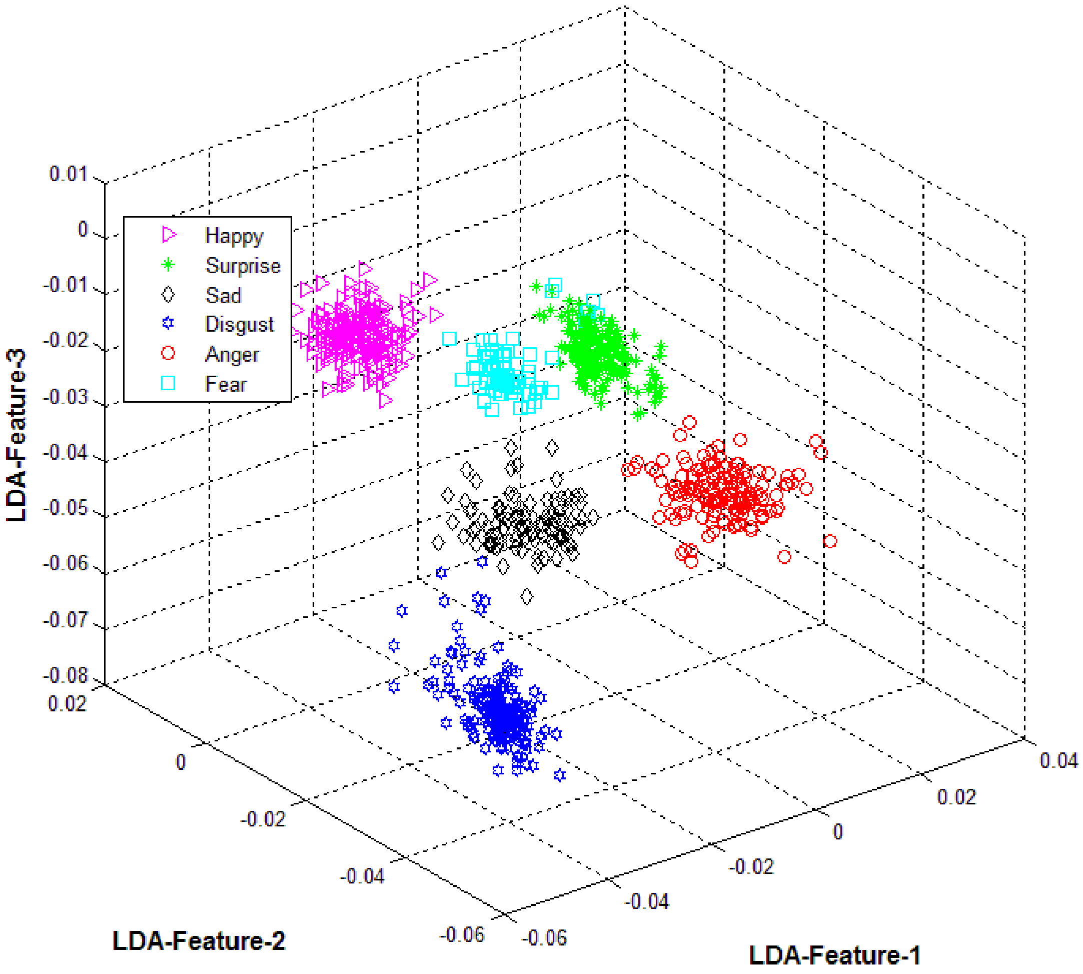


Figure 3.4: 3D feature plots of the proposed hierarchical system after applying LDA at the second level for recognizing the expressions in each category. It can be seen that at the second level, the proposed hierarchical system achieved much higher recognition rate as compared to a single-LDA via single-HMM shown in Figure 3.5.

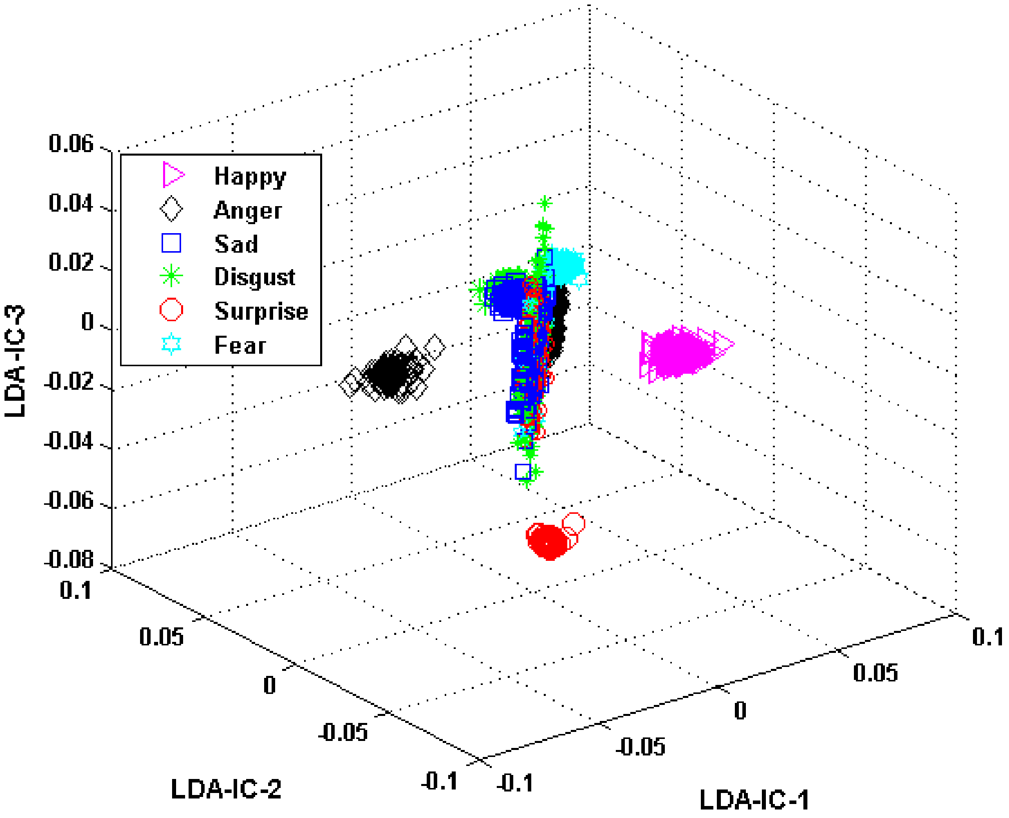


Figure 3.5: 3D-feature plot of single-LDA via single-HMM. It can be seen that using a single-LDA via single-HMM approach did not yield as good a separation among different classes as was achieved by the proposed hierarchical system (See Figure 3.4).

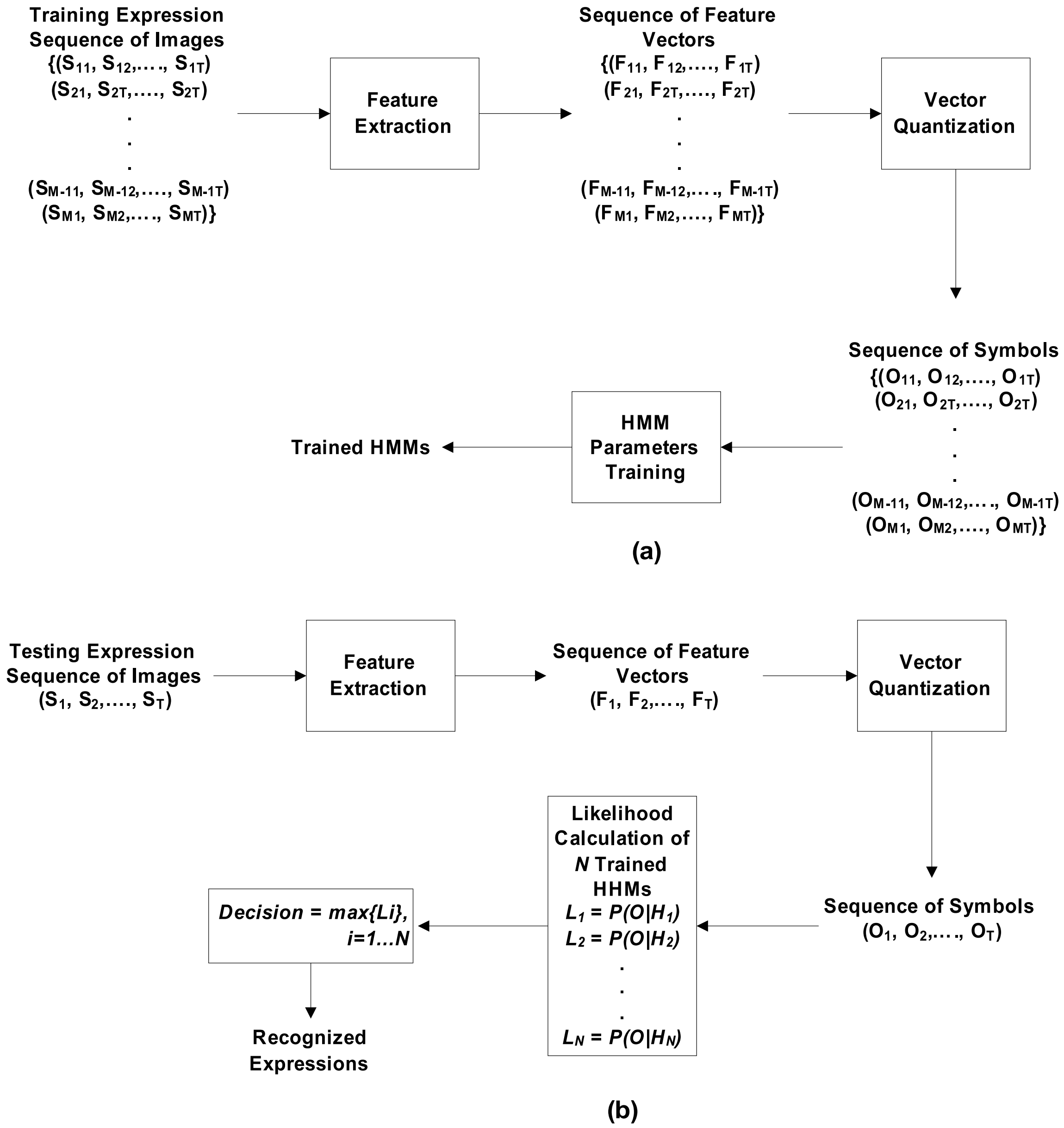

Training Expression Sequence of Images
$\{(S_{11}, S_{12},\ldots, S_{1T})$
$(S_{21}, S_{2T},\ldots, S_{2T})$
.
.
.
$(S_{M-11}, S_{M-12},\ldots, S_{M-1T})$
$(S_{M1}, S_{M2},\ldots, S_{MT})\}$
Feature Extraction
Sequence of Feature Vectors
$\{(F_{11}, F_{12},\ldots, F_{1T})$
$(F_{21}, F_{2T},\ldots, F_{2T})$
.
.
.
$(F_{M-11}, F_{M-12},\ldots, F_{M-1T})$
$(F_{M1}, F_{M2},\ldots, F_{MT})\}$
Vector Quantization
Sequence of Symbols
$\{(O_{11}, O_{12},\ldots, O_{1T})$
$(O_{21}, O_{2T},\ldots, O_{2T})$
.
.
.
$(O_{M-11}, O_{M-12},\ldots, O_{M-1T})$
$(O_{M1}, O_{M2},\ldots, O_{MT})\}$
HMM Parameters Training
Trained HMMs
(a)
Testing Expression Sequence of Images
$(S_1, S_2,\ldots, S_T)$
Feature Extraction
Sequence of Feature Vectors
$(F_1, F_2,\ldots, F_T)$
Vector Quantization
Sequence of Symbols
$(O_1, O_2,\ldots, O_T)$
Likelihood Calculation of *N* Trained HHMs
$L_1 = P(O|H_1)$
$L_2 = P(O|H_2)$
.
.
.
$L_N = P(O|H_N)$
Decision = max{Li}, i=1...N
Recognized Expressions
(b)

Figure 3.6: (a) HMM training for an expression, while (b) testing an expression in an image sequence.

version of the HCRF by employing the full covariance Gaussian distribution. The overall architectural diagram of the proposed FER system is given in Figure 3.7.

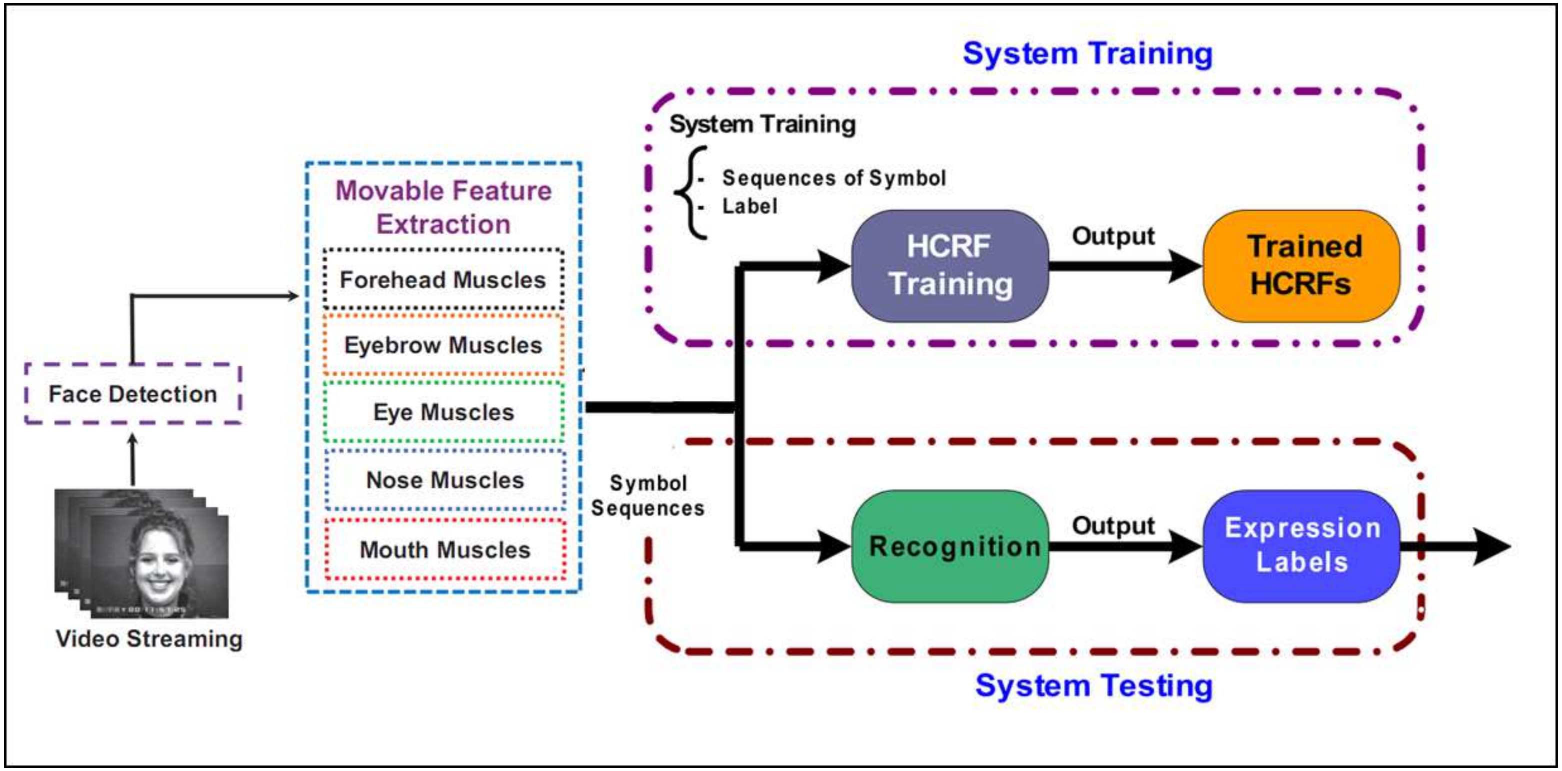


Figure 3.7: Architectural diagram for the proposed robust FER system.

As can be seen from the architectural diagram that there are four modules in a typical FER system. But we have contribution only in feature extraction and recognition modules only. The methods for each module are described as given below.

3.4.2 Face Detection and Extraction

The active contour (AC) model is a well-known technique in the field of image segmentation. It is a deformable spline influenced by constraint and image forces that pull it towards object contours. It tries to move into a position where its energy is minimized. The AC model tries to improve by imposing desirable properties such as continuity and smoothness to the contour of the object, which means that the AC approach adds a certain degree of prior knowledge for dealing with the problem of finding the object contour.

Recently, Chan-Vese (CV) proposed in [105] a novel form of AC for object segmentation based on level set framework. Its energy function is defined by the following equation.

$$\mathrm{F}\left(\mathrm{C}\right) = \int\limits_{\mathrm{inside(C)}} \left|\mathrm{I}\left(\mathrm{x}\right) - \mathrm{c_{in}}\right|^2 \mathrm{dx} + \int\limits_{\mathrm{outside(C)}} \left|\mathrm{I}\left(\mathrm{x}\right) - \mathrm{c_{out}}\right|^2 \mathrm{dx} \tag{3.13}$$

where c_{in} and c_{out} are respectively the average intensities inside and outside of the curve C. Compared to the other AC models, the CV AC model can detect the faces more exactly since it does not need to smooth the initial facial image (via the edge function $g\left|\nabla I_\sigma\right|^2$), even if it is very noisy. In other words, this model is more robust to noise. The CV AC model does not use the edge information but utilizes the difference between the regions inside and outside of the curve, making itself one of the most robust and thus widely used techniques for image segmentation, especially, in the area of face detection. However, the convergence of CV AC model depends on the homogeneity of the segmented faces. When the inhomogeneity becomes large, the CV AC model provides unsatisfactory results. Moreover, the global minimum of the above energy function does not always guarantee the desirable results. The unsatisfactory result of the CV AC model in this case is due to the fact that it tries to minimize the dissimilarity within each segment but does not take into account the distance between different segments.

The proposed methodology in this work is to incorporate an evolving term based on the Bhattacharyya distance [106] to the CV energy function that minimizes the dissimilarities within the object (face) and maximizes the distance between the two regions (face and background). The proposed energy function is given below:

$$E\left(C\right) = \beta F\left(C\right) + \left(1-\beta\right) B\left(C\right) \tag{3.14}$$

where $\beta \in [0,1]$. Note that the value of $B(C)$ is always within the interval $[0,1]$ whereas $F(C)$ is calculated based on the integral over the facial image plane.

The intuition behind the proposed model (in $E(C)$) is that we seek for a curve which is regular (the first two terms) and partitions the facial image into regions such that the differences within each region are minimized (the $F(C)$ term) like reducing environmental effects and the distance between the two regions (i.e., face and background) is maximized (the term $B(C)$).

For the level-set formulation, let is define ϕ as the level-set function, $I : \Omega \rightarrow \mathrm{Z} \subset R^n$ as a certain image feature such as intensity, color, texture, or a combination thereof, and $H(\bullet)$ and $\delta_0\left(\bullet\right)$ as the Heaviside and the Dirac function respectively. The energy function can then be rewritten as given in the following equations.

$$\begin{aligned} E\left(\phi\right) = {} & \gamma \int\limits_{\Omega} \left|\nabla H\left(\phi\left(x\right)\right)\right| dx + \eta \int\limits_{\Omega} H\left(-\phi\left(x\right)\right) \\ & + \beta \left[\int\limits_{\Omega} \left|I\left(x\right) - c_{in}\right|^2 H\left(-\phi\left(x\right)\right) + \int\limits_{\Omega} \left|I\left(x\right) - c_{out}\right|^2 H\left(\phi\left(x\right)\right) \right] \\ & + \left(1-\beta\right) \int\limits_{Z} \sqrt{p_{in}\left(z\right) p_{out}\left(z\right)} dz \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} p_{in}(z) &= \frac{\int\limits_{\Omega} \delta_0 (z - I(x)) H(-\phi(x)) dx}{\int\limits_{\Omega} H(-\phi(x)) dx} \\ p_{out}(z) &= \frac{\int\limits_{\Omega} \delta_0 (z - I(x)) H(\phi(x)) dx}{\int\limits_{\Omega} H(\phi(x)) dx} \end{aligned} \tag{3.16}$$

where γ and η are non-negative constants, and $p_{in}(z)$ and $p_{out}(z)$ are given in (3.16) are the local fitting functions [119] that depends on the level set function ϕ and needs to be updated in each contour evaluation. Differentiating w.r.t $\phi(x)$, then, the evaluation flow associated with minimizing the energy function in (3.15) is given as

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi} = \delta_0(\phi)\{M - (1-\beta)[N + O]\} \tag{3.17}$$

where

$$M = \gamma k + V_0 + \beta \left[(I - c_{in})^2 + (I - c_{out})^2 \right] \tag{3.18}$$

$$N = \frac{B(C)}{2} \left(\frac{1}{A_{in}} - \frac{1}{A_{out}} \right) \tag{3.19}$$

$$O = \frac{1}{2} \int_z \delta_0 (z - 1) \left(\frac{1}{A_{out}} \sqrt{\frac{p_{in}}{p_{out}}} - \frac{1}{A_{in}} \sqrt{\frac{p_{out}}{p_{in}}} \right) dz \tag{3.20}$$

where A_{in} and A_{out} are respectively the areas inside and outside the curve C. Thus, the proposed AC model overcome the limitation of conventional CV AC model in the area of face detection.

3.4.3 Feature Extraction

3.4.3.1 Noise Reduction via Wavelet Transform

In real-life scenarios, some environmental parameters (such as lighting effects) may produce some noise in the expression frames that could reduce the recognition rate. The proposed method employs symlet wavelet to reduce such noise. Facial frames are converted to gray scale prior to applying this step.

The wavelet decomposition could be interpreted as signal decomposition into a set of independent feature vector. Each vector consists of sub-vectors like

$$V_0^{2D} = V_0^{2D-1}, V_0^{2D-2}, V_0^{2D-3}, ..., V_0^{2D-n} \tag{3.21}$$

where V represents the 2D feature vector. If we have an expression frame X in the decomposition process, and it breaks up into the orthogonal sub images corresponding to different visualization. The following equation shows one level of decomposition.

$$X = A_1 + D_1 \tag{3.22}$$

where X indicates the decomposed image and A_1 and D_1 are called approximation and detail coefficient vectors respectively. If a facial frame is decomposed up to multiple levels, then (3.22) can be written as

$$X = A_j + D_j + D_{j-1} + D_{j-2} + \dots + D_2 + D_1 \tag{3.23}$$

where j represents the level of decomposition. The detail coefficients mostly consist of noise, so for feature extraction only the approximation coefficients are used. In the proposed algorithm, each facial frame is decomposed up to two levels, i.e., the value of $j = 2$, because by exceeding the value of $j = 2$, the facial frame looses significant information, due to which the informative coefficients cannot be detected properly, which may cause misclassification. The detail coefficients further consist of three sub-coefficients, so the (3.23) can be written as

$$\begin{aligned} X &= A_2 + D_2 + D_1 \\ &= A_2 + \left[(D_h)_2 + (D_v)_2 + (D_d)_2\right] + \left[(D_h)_1 + (D_v)_1 + (D_d)_1\right] \end{aligned} \tag{3.24}$$

where D_h, D_v and D_d are known as horizontal, vertical and diagonal coefficients respectively. Note that at each decomposition step, approximation and detail coefficient vectors are obtained by passing the signal through a low-pass filter and high-pass filter respectively.

In a specified time window and frequency bandwidth wavelet transform, the frequency is estimated. The signal (i.e., facial frame) is analyzed by using the wavelet transform [120].

$$C\left(a_i,\, b_j\right) = \frac{1}{\sqrt{a_i}} \int\limits_{-\infty}^{\infty} y\left(t\right) \Psi^{*}_{f.e} \left(\frac{t - b_j}{a_i}\right) dt \tag{3.25}$$

where a_i is the scale of the wavelet between lower and upper frequency bounds to get high decision for frequency estimation, and b_j is the position of the wavelet from the start to the end of the time window with the specified signal sampling period, t is the time, the wavelet function $\Psi_{f.e}$ is used for frequency estimation, and $C(a_i, b_i)$ are the wavelet coefficients with the specified scale and position parameters. Finally, the scale is converted to the mode frequency, f_m for each facial frame:

$$f_m = \frac{f_a\left(\Psi_{f.e}\right)}{a_m\left(\Psi_{f.e}\right) . \Delta} \tag{3.26}$$

where $f_a\left(\Psi_{f.e}\right)$ is the average frequency of the wavelet function, and Δ is the signal sampling period. Next, the facial movement feature have been extracted by exploiting the optical flow [121].

3.4.3.2 Feature Extraction via Optical Flow

The motion information of the facial pixels is found by employing the optical flow in order to generate the feature vectors for each expression frame. In order to find the optical flow of the two expression frames: first, a kernel is made for partial derivative of Gaussian (like g_x and g_y) with respect to x and y such as

$$g_x\left(i,j\right) = -\frac{j-k-1}{2\pi\,\Sigma^3}\exp\left(-\frac{\left(i-k-1\right)^2+\left(j-k-1\right)^2}{2\,\Sigma^2}\right) \tag{3.27}$$

$$g_y\left(i,j\right) = -\frac{j-k-1}{2\pi\,\Sigma^3}\exp\left(-\frac{\left(i-k-1\right)^2+\left(j-k-1\right)^2}{2\,\Sigma^2}\right) \tag{3.28}$$

where $k = \frac{N-1}{2}$ and N is the size of the kernel, x and y derivatives are computed for both frames. After that the images are smoothed by building a Gaussian kernel such as

$$kernel\left(i,j\right) = -\frac{1}{2\pi\,\sigma^2}\exp\left(-\frac{\left(i-k-1\right)^2+\left(j-k-1\right)^2}{2\,\sigma^2}\right) \tag{3.29}$$

$$A = \begin{bmatrix} \Sigma\, I_x, I_x & \Sigma\, I_x, I_y \\ \Sigma\, I_x, I_y & \Sigma\, I_y, I_y \end{bmatrix} B = \begin{bmatrix} \Sigma\, I_x, I_t \\ \Sigma\, I_y, I_t \end{bmatrix} \tag{3.30}$$

The resultant image is given as

$$R = A^{-1}\left(-B\right) \tag{3.31}$$

where R is the resultant image that has the pixels motion information. For instance, such pixel motion information for the two consecutive expression frames are shown in Figure 3.8.

The average feature vector is obtained by taking the average of the whole pixels motion information for all the facial frames in a video clip which is given below:

$$f_{ave} = \frac{R_1 + R_2 + R_3 + ... + R_K}{N} \tag{3.32}$$

where f_{ave} indicates the average feature vector of all the expressions frames that a single expression video have, R_1, R_2, R_3, ..., R_K are the motion information for each expression frame, K is the last frame of the expression video, and N represents the whole number of frames in each expression video.

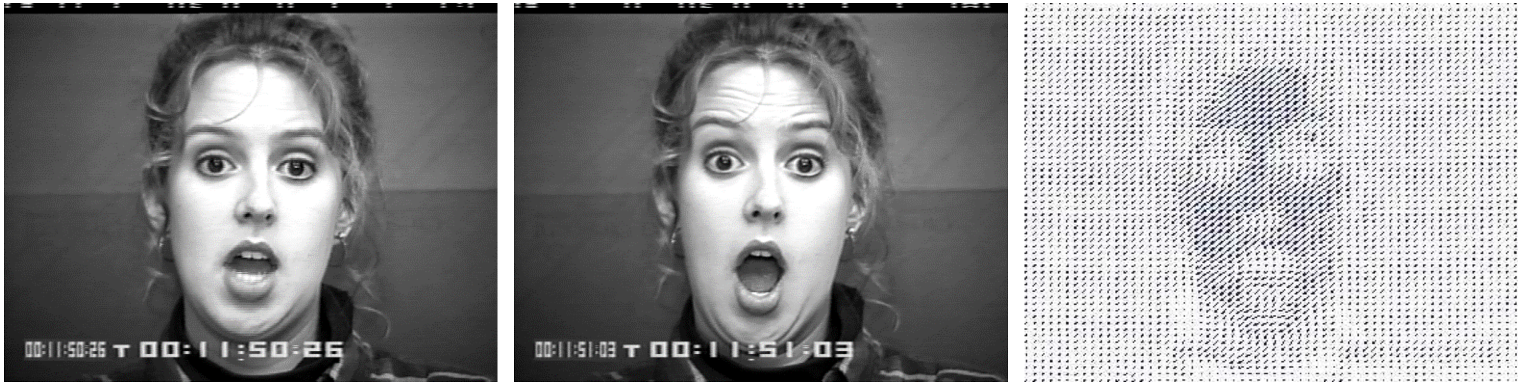


Figure 3.8: Two consecutive expression frames and its corresponding optical flow (pixel movement information).

The average feature vector is obtained by taking the average of the whole pixels motion information for all the facial frames in a video clip which is given below:

$$f_{ave} = \frac{R_1 + R_2 + R_3 + ... + R_K}{N} \tag{3.33}$$

where f_{ave} indicates the average feature vector of all the expressions frames that a single expression video have, R_1, R_2, R_3, ..., R_K are the motion information for each expression frame, K is the last frame of the expression video, and N represents the whole number of frames in each expression video.

3.4.4 Stepwise Linear Discriminant Analysis (SWLDA)

In this step, the most informative features are selected by using SWLDA, which maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. The forward and backward regression techniques enable SWLDA to effectively reduce the dimensions of the feature space.

In the forward regression step, the most correlated features are selected based on partial *F*-test values from the feature space. On the other hand, in the backward regression step, the least significant values are removed from the regression model i.e. lower *F*-test values. In both processes, the *F*-test values are calculated on the basis of defined class labels. The advantage of this method is that it is very efficient in seeking the localized features.

In the beginning, there is no predictor in the model. Based on the significance test, i.e., partial *F-test* (the *t-test*), predictor is either entered or removed from the model in each iteration. Two predictors Alpha-to enter and Alpha-to remove are defined for significance level test. *Alpha-to enter* $a_e = 0.25$ and *Alpha-to-remove* $a_\gamma = 0.30$ are set as threshold parameters. These values are chosen based on various experiments. These threshold parameters show the significance level of the predictors which are entered or removed from the model, respectively. The algorithm stops when there are no more predictors to enter or remove from the stepwise model.

We present an example in which we have three independent predictors: x_1, x_2, and x_3, and an output (response) y. Each predictor fits into the model using a regression; that is, we regress y on x_1, x_2, $\ldots$, and x_{p-1}, where p is the total number of predictors ($p = 3$ in this case). The first predictor to enter into the stepwise model is the predictor that has the smallest *t-test p-value* (i.e., below a_e). This will continue until the stopping criterion is met (i.e., if there is no predictor with a *p-value* less than a_e).

Now suppose x_1 is the best predictor. Then fit each of the two predictor models that includes x_1 in the model, i.e., the model regresses y on (x_1, x_2), regress y on $(x_1, x_3)...y$ on (x_1, x_{p-1}). The second predictor enters into the stepwise model is the predictor that has the smallest *p-value*. If again there is no *p-value* less than a_e, the iteration stops.

Suppose this time x_2 is the best second predictor in the model. The analysis procedure then steps back and checks the *p-value* for $\beta_1 = 0$ (i.e., criterion for the removal of the predictor from the model). In this case, if the *p-value* is above a_γ for $\beta_1 = 0$, then the predictor is not significant compared to the new entry, and x_1 is removed from the stepwise model.

In contrast, suppose both x_1 and x_2 have made it into the two-predictor stepwise model. The analysis procedure then fits each of the three-predictor models with x_1 and x_2 in the model, i.e., it regresses y on (x_1, x_2, x_3), regresses y on $(x_1, x_2, x_4), ...,$ and regresses y on (x_1, x_2, x_{p-1}). The third predictor that enters the stepwise model is the predictor that has the smallest *p-value* less than a_e. The stopping criterion is met when there is no *p-value* less than a_e. In this case, the analysis checks the *p-value* $\beta_1 = 0$. If either *p-value* has not become significant (i.e., above a_γ), the predictor is removed from the stepwise model. This procedure will stop when adding an additional predictor does not yield a *p-value* below a_e. For more details on SWLDA, please refer to a previous study [122].

3.4.5 Vector Quantization

"Once the facial expressions are represented as features, then the optical-flow based features are symbolized by means of comparing with the codeword vectors of codebook. The purpose of this process in order to decode the sequential variations of the facial expression features, and then the discrete HCRFs have been utilized. As discrete HCRFs are commonly used trained and tested with symbols sequences, the feature vectors are symbolized by means of comparing with the codeword vectors of a codebook. To obtained the codebook, vector quantization algorithm is performed on the feature vectors from the training dataset. In the proposed FER system, we utilized Linde, Bunzo, and Gray (LBG) algorithm [123] has been utilized for codebook generation. The LBG approach selects the initial centroids and splits the centroids of the whole dataset. Then, it continues to split the dataset accordingly to the codeword size, and the optimization is performed to reduce the distortion.

Once the codebook is obtained, the index numbers of the codewords are regarded as symbols to be used with discrete HCRFs. As each expression image is converted to a symbol, an expression video clip of T consecutive images will result in T symbols after the vector quantization. Figure 3.9 presents the basic steps for the codebook generation from all the expression image vectors and symbol selection for a sample facial expression image vector" [1].

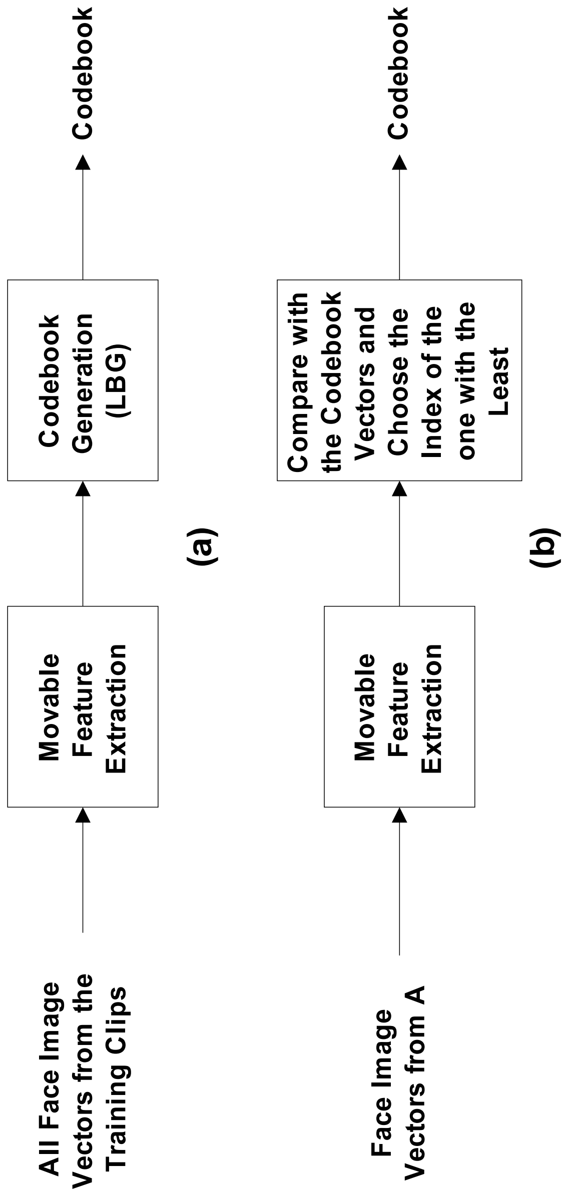

All Face Image Vectors from the Training Clips
Movable Feature Extraction
Codebook Generation (LBG)
Codebook
(a)
Face Image Vectors from A
Movable Feature Extraction
Compare with the Codebook Vectors and Choose the Index of the one with the Least
Codebook
(b)

Figure 3.9: (a) Steps for codebook generation, and (b) steps for symbol selection [1].

3.4.6 Classification using Hidden Conditional Random Fields (HCRF)

According to an exemplary aspect, there is provided a method including: dividing an input image measured from a variety of inputs and outputting a frame sequence; extracting a feature vector from the frame sequence; combining full covariance Gaussian distributions with a hidden conditional random fields model; receiving, by the hidden conditional random fields model, combinations of the feature vector and a label indicating a specific activity to obtain a parameter of the hidden conditional random fields model; receiving, by the hidden conditional random fields model to which the parameter has been applied, a feature vector extracted from a test input image measured for an actual activity to infer a label indicating the actual activity and indicate a sequence of a specific state; applying a gradient function-applied algorithm for analyzing the sequence of the specific state; and calculating probability of a state sequence.

As mentioned before that the existing HCRF utilizes diagonal covariance Gaussian distributions in the feature function and does not guarantee the convergence of its parameters to some specific values at which the conditional probability is modeled as a mixture of normal density functions. Because of this property, the existing HCRF losses a lot of information. This is one of the main disadvantages of the existing HCRF model.

In order to solve this limitation, we explicitly involve full covariance Gaussian density function in the feature functions at the observation level. Since there is no tool for a hidden conditional random fields model that can use the full-covariance Gaussian density function. For the prior and transition probabilities, we used the same equations of [2]. Mathematically, the contribution of the proposed model can be explained as

$$f_s^{Ob}\left(Y,\overline{S},X\right)=\sum_{t=1}^{T}\log\left(\sum_{m=1}^{M}\Gamma_{s,m}^{Obs}N\left(x_t^2,\mu_{s,m},\Sigma_{s,m}\right)\right)\left(\delta\left(s_t=s\right)\right), \tag{3.34}$$

The (3.34) presents the observation of the input at each state. Where M is the number of density functions, Γ is used in order to consider the contextual information of the whole observation, $\Gamma_{s,m}^{Obs}$ is the mixing weight of the m^{th} component with mean $\mu_{s,m}$ and covariance matrix $\sum_{s,m}$.

$N\left(x_t^2, \mu_{s,m}, \Sigma_{s,m}\right)$ in (3.34) can be computed as

$$N\left(x_t^2, \mu_{s,m}, \Sigma_{s,m}\right) = \frac{1}{(2\pi)^{D/2}|\Sigma_{s,m}|^{1/2}} \exp\left(-\tfrac{1}{2}\left(x_t^2 - \mu_{s,m}\right)'\Sigma_{s,m}^{-1}(x_t^2 - \mu_{s,m})\right), \tag{3.35}$$

where D is the dimension of the observation, and $\sum_{s,m}$ is the full covariance matrix.

As we can see in (3.34), by changing Γ, μ and $\sum$ we can create any mixture of the normal densities. So, the corresponding observation weight (Λ_s^{Obs}) is not necessary to be updated during the training phase. Therefore,

$$\Lambda_s^{Ob} = 1, \quad \forall s \in \overline{S}, \tag{3.36}$$

As a result, the conditional probability that is used to model the system can be rewritten as

$$p\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right) = \frac{\sum\limits_{\overline{S}} \exp\left(P\left(\overline{S}\right) + T\left(\overline{S}\right) + O\left(\overline{S}\right)\right)}{z\left(X,\Lambda,\Gamma,\mu,\Sigma\right)}, \tag{3.37}$$

where

$$P\left(\overline{S}\right) = \sum_{s\in\overline{S}} \Lambda_{y'}^{Pr} f_{y'}^{Pr}\left(Y,\overline{S},X\right), \tag{3.38}$$

$$T\left(\overline{S}\right) = \sum_{\{ss'\}\in\overline{S}} \Lambda_{ss'}^{Tr} f_{ss'}^{Tr}\left(Y,\overline{S},X\right), \tag{3.39}$$

$$O\left(\overline{S}\right) = \sum_{s\in\overline{S}} f_s^{Ob}\left(Y,\overline{S},X\right), \tag{3.40}$$

By putting the values of $P\left(\overline{S}\right)$, $T\left(\overline{S}\right)$, and $O\left(\overline{S}\right)$ from (3.38), (3.39), and (3.40) respectively in (3.37), the updated conditional probability can be rewritten in (3.4.6).

$$p\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right) = \frac{\sum\limits_{\overline{S}} \exp\left(\sum\limits_{s\in\overline{S}} \Lambda_{y'}^{Pr} f_{y'}^{Pr}\left(Y,\overline{S},X\right) + \sum\limits_{\{ss'\}\in\overline{S}} \Lambda_{ss'}^{Tr} f_{ss'}^{Tr}\left(Y,\overline{S},X\right) + \sum\limits_{s\in\overline{S}} f_s^{Ob}\left(Y,\overline{S},X\right)\right)}{z\left(X,\Lambda,\Gamma,\mu,\Sigma\right)}, \tag{3.41}$$

As mentioned before, our contribution is at the observation level; therefore, by putting the value of $f_s^{Ob}\left(Y,\overline{S},X\right)$ from (3.34), the updated conditional probability for the system can be rewritten in (3.4.6).

$$p\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)=\frac{\sum\limits_{\overline{S}=s_1,s_2,..,s_T}\exp\left(\Lambda_{y'}^{Pr}+\sum\limits_{t=1}^{T}\left(\Lambda_{s_{t-1},s_t}^{Tr}\right)+\log\left(\sum\limits_{m=1}^{M}\Gamma_{s_t,m}^{Obs}N\left(x_t^2,\mu_{s_t,m},\Sigma_{s_t,m}\right)\right)\right)}{z\left(X,\Lambda,\Gamma,\mu,\Sigma\right)},\tag{3.42}$$

The simple form of the conditional probability is defined in (3.43).

$$p\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)=\frac{Score\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{z\left(X;\Lambda,\Gamma,\mu,\Sigma\right)},\tag{3.43}$$

The procedure of the proposed HCRF follows exactly the procedure of the [2]. Based on equations (3.4.6) and (3.43), we can further update the conditional probability using the well-known forward and backward algorithms (as the algorithms used in HMM), which are defined in equations (3.44) and (3.45) respectively.

$$\alpha_\tau=\sum_{\overline{S}=s_1,s_2,..,\{s_\tau=s\}}\exp\left(\Lambda_{y'}^{Pr}+\sum_{t=1}^{\tau}\left(\Lambda_{s_{t-1},s_t}^{Tr}\right)+\log\left(\sum_{m=1}^{M}\Gamma_{s_t,m}^{Obs}N.\left(x_t^2,\mu_{s_t,m},\Sigma_{s_t,m}\right)\right)\right),$$

$$\alpha_\tau=\sum_{s'\in\overline{S}}\alpha_{\tau-1}\left(s'\right)\exp\left(\Lambda_{s's}^{Tr}+\log\left(\sum_{m=1}^{M}\Gamma_{s,m}^{Obs}N\left(x_\tau,\mu_{s,m},\Sigma_{s,m}\right)\right)\right),\tag{3.44}$$

$$\beta_\tau\left(s\right)=\sum_{\overline{S}=\{s_\tau=s\},s_{\tau+1},..,s_T}\exp\left(\Lambda_{y'}^{Pr}+\sum_{t=\tau}^{T}\left(\Lambda_{s_{t-1},s_t}^{Tr}\right)+\log\left(\sum_{m=1}^{M}\Gamma_{s_t,m}^{Obs}N\left(x_t^2,\mu_{s_t,m},\Sigma_{s_t,m}\right)\right)\right),$$

$$\beta_\tau\left(s\right)=\sum_{s'}\beta_{\tau+1}\left(s'\right)\exp\left(\Lambda_{ss'}^{Tr}+\log\left(\sum_{m=1}^{M}\Gamma_{s,m}^{Obs}N\left(x_\tau,\mu_{s,m},\Sigma_{s,m}\right)\right)\right),\tag{3.45}$$

Therefore, the $Score\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)$ of (3.44) is equal to the forward algorithm (α) and backward algorithm (β) as in (3.46).

$$Score\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)=\sum_{s\in\overline{S}}\alpha_T(s)=\sum_{s\in\overline{S}}\beta_1(s).\tag{3.46}$$

In the training phase, our goal was to find the parameters (Λ,Γ,μ, and $\sum$) to maximize the conditional probability of the training data. In the proposed HCRF model, we utilized (Limited-memory

Broyden-Fletcher-Goldfarb-Shanno) L-BGFS method to search the optimal point. However, instead of repeating the forward and backward algorithms to compute the gradients as others did [2], we run the forward and backward algorithms only when calculating the conditional probability, then we reuse the results to compute the gradients.

3.4.6.1 Analysis of Full Covariance Matrix

As described before, that we explicitly involve the full covariance matrix in the feature function at the observation level as shown in 3.34. For which the normal distribution N may be obtained through the equation shown in 3.35 that further has been explained in the following equations.

$$\begin{aligned} \frac{dScore\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{d\Lambda_s^{Pr}} &= \sum_{\overline{S}} \frac{dg\left(Y,\overline{S},X\right)}{d\Lambda_s^{Pr}} \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \sum_{\overline{S}} f_s^{Pr}\left(Y,\overline{S},X\right) \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \beta_1\left(s\right) \end{aligned} \tag{3.47}$$

The $dScore$ function is a gradient function for a variable of the prior probability vector.

$$\begin{aligned} \frac{dScore\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{d\Lambda_s^{Tr}} &= \sum_{\overline{S}} \frac{dg\left(Y,\overline{S},X\right)}{d\Lambda_{ss'}^{Tr}} \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \sum_{\overline{S}} f_{ss'}^{Tr}\left(Y,\overline{S},X\right) \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \sum_{t=1}^{T} \alpha\left(t,s\right)\beta\left(t+1,s'\right) \end{aligned} \tag{3.48}$$

The $dScore$ function is a gradient function for a variable of the transition probability vector.

$$\begin{aligned}\frac{dScore\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{d\Gamma^{Obs}_{s,m}} &= \sum_{\overline{S}} \frac{dg\left(Y,\overline{S},X\right)}{d\Gamma^{Obs}_{s,m}} \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \frac{\sum\limits_{\overline{S}} f^{Ob}_{s}\left(Y,\overline{S},X\right)}{d\Gamma^{Obs}_{s,m}} \exp\left(g\left(Y,\overline{S},X\right)\right) \\ &= \sum_{\overline{S}} \sum_{t=1}^{T} \frac{N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)}{\sum\limits_{m=1}^{M} \Gamma^{Obs}_{s,m} N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)} \\ = \delta\left(s_t = s\right) \exp\left(g\left(Y,\overline{S},X\right)\right) &\sum_{t=1}^{T} \frac{N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)}{\sum\limits_{m=1}^{M} \Gamma^{Obs}_{s,m} N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)} \alpha\left(t,s\right)\gamma\left(t+1\right)\end{aligned} \tag{3.49}$$

The $dScore$ function is a gradient function for a Gaussian mixture weight variable. Here, a function $Y\left(t\right)$ is calculated as

$$\gamma\left(t\right) = \sum_{s} \beta\left(t,s\right) \tag{3.50}$$

$$\frac{dScore\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{d\mu_{s,m}} = \sum_{t=1}^{T} \frac{\Gamma^{Obs}_{s,m} \frac{dN\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)}{d\mu_{s,m}}}{\sum\limits_{m=1}^{M} \Gamma^{Obs}_{s,m} N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)} \alpha\left(t,s\right)\gamma\left(t+1\right) \tag{3.51}$$

The $dScore$ function is a gradient function for the Gaussian distribution mean.

$$\frac{dScore\left(Y|X;\Lambda,\Gamma,\mu,\Sigma\right)}{d\Sigma_{s,m}} = \sum_{t=1}^{T} \frac{\Gamma^{Obs}_{s,m} \frac{dN\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)}{d\Sigma_{s,m}}}{\sum\limits_{m=1}^{M} \Gamma^{Obs}_{s,m} N\left(x_t,\mu_{s,m},\Sigma_{s,m}\right)} \alpha\left(t,s\right)\gamma\left(t+1\right) \tag{3.52}$$

The $dScore$ function is a gradient function for covariance of the Gaussian distributions.

Equations 3.47, 3.48, 3.49, and 3.50 represent an analysis method algorithm for calculating values of gradients for a feature function, the mean of Gaussian distributions, and the covariance of the prior probability vector, the transition probability vector, and the observation probability vector obtained from the existing HCRF.

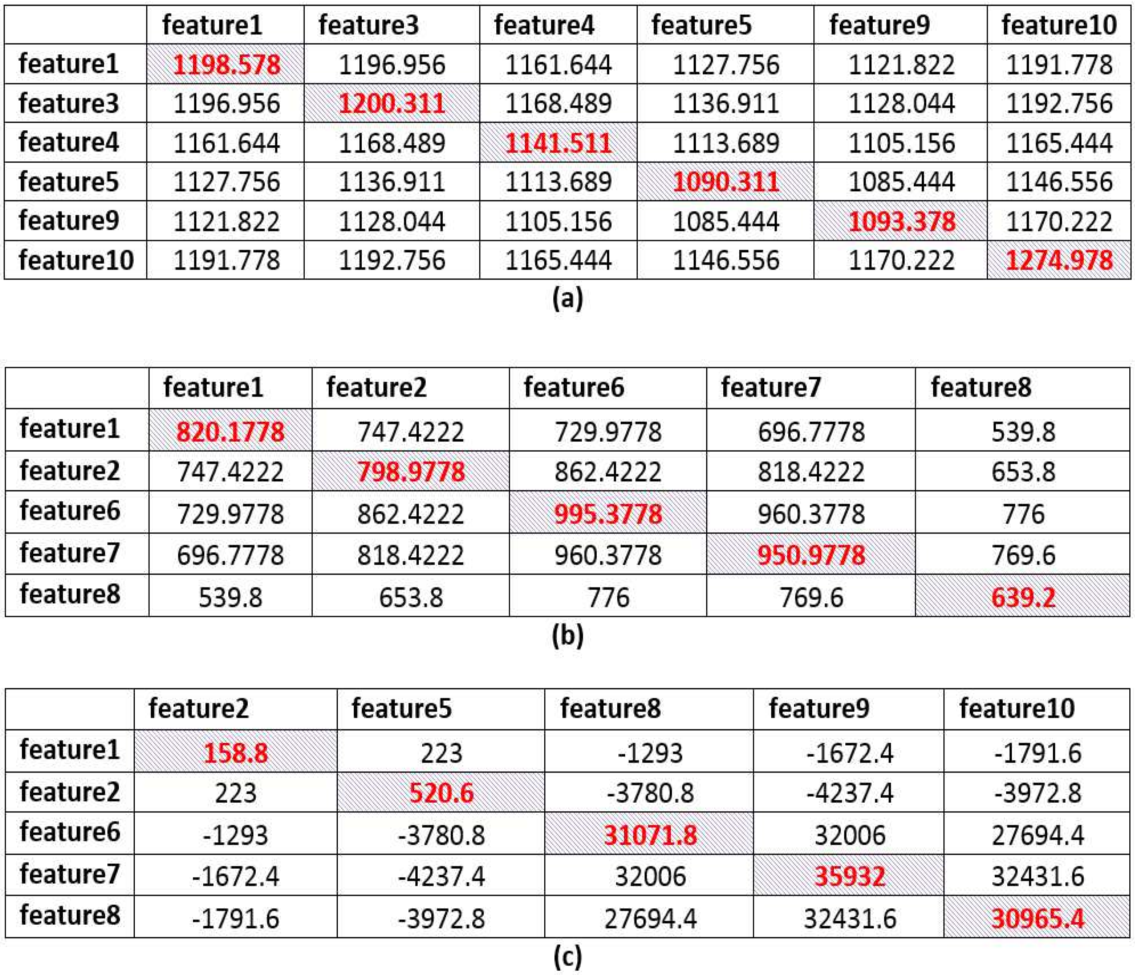
In the proposed model, the method is divided into a training step and an inference step in recognizing a variety of actual expressions. The training step refers to a step of inputting data whose labels are known, of a recognition target and training the hidden conditional random fields model. For example, in the case of emotion recognition based on video, video representing happy, sad, angry, fear, and emotions whose states are known in advance are input as the training data. In the inference step, the inputs to be actually measured are classified based on parameters calculated in the training step.

If in the training, the expression frame inserted as an input, then, in the preprocessing step, the different lighting effects are diminished and faces are detected and extracted from the expression frames. Then, the movable features are extracted from the different parts of the face to generate the feature vector. Then, the extracted feature vector is input to a full-covariance Gaussian-mixed hidden conditional random fields model of the proposed recognition model.

As described before, in the training step of the proposed HCRF model, a feature gradient is generally calculated by an LBFG method. However, in a current gradient calculation method, a forward and backward iterative execution algorithm is iteratively invoked, which requires a very great computational amount and accordingly degrades a computation speed. A new analysis method that decreases the invoking of the forward and backward iterative execution algorithms has been devised. Through the five gradient functions calculated using Equations 3.47, 3.48, 3.49, 3.50, and 3.51 real-time computation can be performed with a smaller computational amount and at a higher speed compared to an existing analysis method. The sample of full covariance matrices for the six basic expressions such as happy, anger, sad, surprise, fear, and disgust on Cohn-Kanade dataset are shown in Figures 3.10 and 3.11.

The training and testing for the proposed HCRF is given in Figure 3.12.

The overall work flow for the proposed FER system is presented in Figure 3.13



	feature1	feature3	feature4	feature5	feature9	feature10
feature1	**1198.578**	1196.956	1161.644	1127.756	1121.822	1191.778
feature3	1196.956	**1200.311**	1168.489	1136.911	1128.044	1192.756
feature4	1161.644	1168.489	**1141.511**	1113.689	1105.156	1165.444
feature5	1127.756	1136.911	1113.689	**1090.311**	1085.444	1146.556
feature9	1121.822	1128.044	1105.156	1085.444	**1093.378**	1170.222
feature10	1191.778	1192.756	1165.444	1146.556	1170.222	**1274.978**

(a)

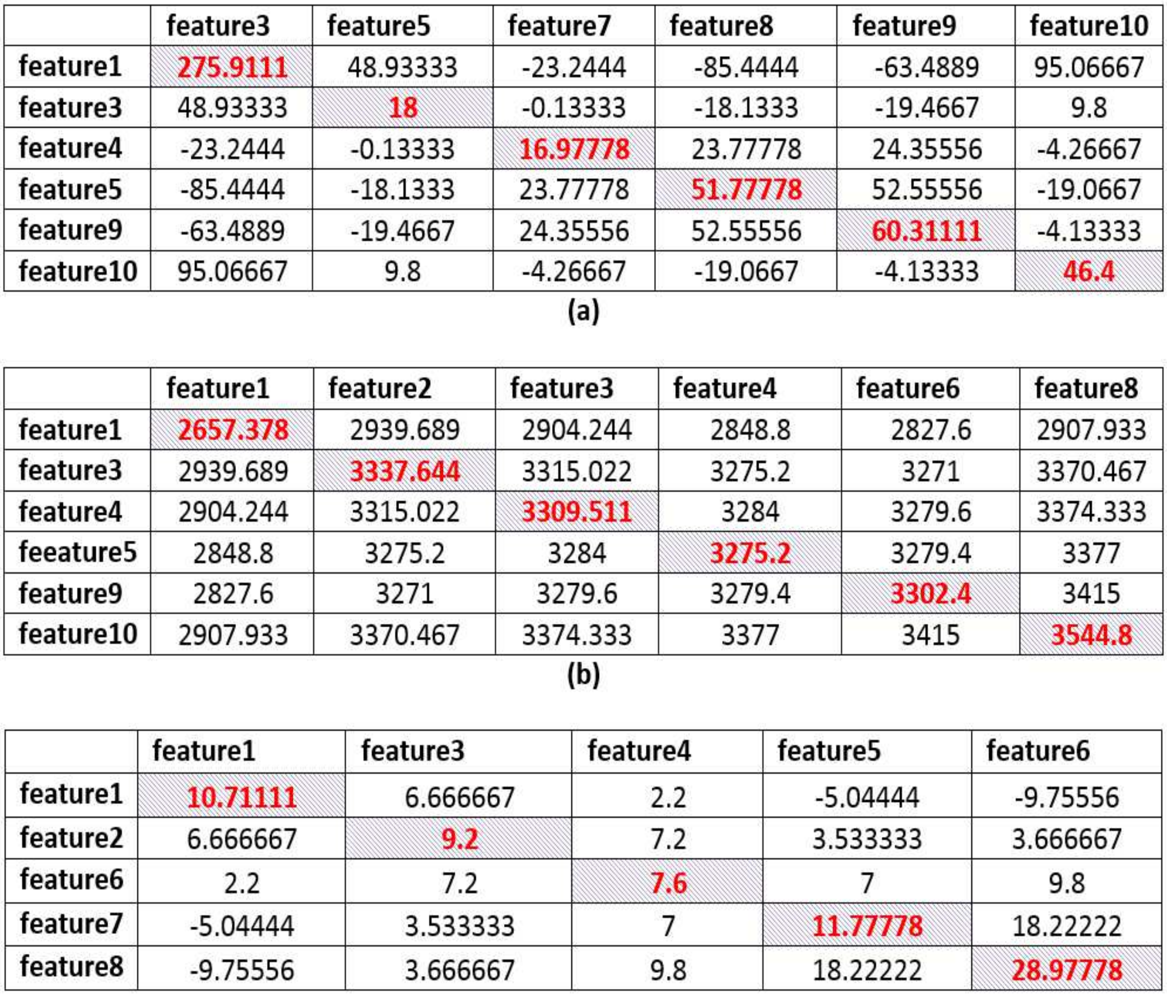
	feature1	feature2	feature6	feature7	feature8
feature1	**820.1778**	747.4222	729.9778	696.7778	539.8
feature2	747.4222	**798.9778**	862.4222	818.4222	653.8
feature6	729.9778	862.4222	**995.3778**	960.3778	776
feature7	696.7778	818.4222	960.3778	**950.9778**	769.6
feature8	539.8	653.8	776	769.6	**639.2**

(b)

	feature2	feature5	feature8	feature9	feature10
feature1	**158.8**	223	-1293	-1672.4	-1791.6
feature2	223	**520.6**	-3780.8	-4237.4	-3972.8
feature6	-1293	-3780.8	**31071.8**	32006	27694.4
feature7	-1672.4	-4237.4	32006	**35932**	32431.6
feature8	-1791.6	-3972.8	27694.4	32431.6	**30965.4**

(c)

Figure 3.10: Sample of full covariance matrices of the proposed recognition model for (a) happy expressions of different subjects, (b) anger expressions of different subjects, and (c) sad expressions of different subjects on Cohn-Kanade dataset.



	feature3	feature5	feature7	feature8	feature9	feature10
feature1	**275.9111**	48.93333	-23.2444	-85.4444	-63.4889	95.06667
feature3	48.93333	**18**	-0.13333	-18.1333	-19.4667	9.8
feature4	-23.2444	-0.13333	**16.97778**	23.77778	24.35556	-4.26667
feature5	-85.4444	-18.1333	23.77778	**51.77778**	52.55556	-19.0667
feature9	-63.4889	-19.4667	24.35556	52.55556	**60.31111**	-4.13333
feature10	95.06667	9.8	-4.26667	-19.0667	-4.13333	**46.4**

(a)

	feature1	feature2	feature3	feature4	feature6	feature8
feature1	**2657.378**	2939.689	2904.244	2848.8	2827.6	2907.933
feature3	2939.689	**3337.644**	3315.022	3275.2	3271	3370.467
feature4	2904.244	3315.022	**3309.511**	3284	3279.6	3374.333
feeature5	2848.8	3275.2	3284	**3275.2**	3279.4	3377
feature9	2827.6	3271	3279.6	3279.4	**3302.4**	3415
feature10	2907.933	3370.467	3374.333	3377	3415	**3544.8**

(b)

	feature1	feature3	feature4	feature5	feature6
feature1	**10.71111**	6.666667	2.2	-5.04444	-9.75556
feature2	6.666667	**9.2**	7.2	3.533333	3.666667
feature6	2.2	7.2	**7.6**	7	9.8
feature7	-5.04444	3.533333	7	**11.77778**	18.22222
feature8	-9.75556	3.666667	9.8	18.22222	**28.97778**

Figure 3.11: Sample of full covariance matrices of the proposed recognition model for (a) surprise expressions of different subjects, (b) fear expressions of different subjects, and (c) disgust expressions of different subjects on Cohn-Kanade dataset.

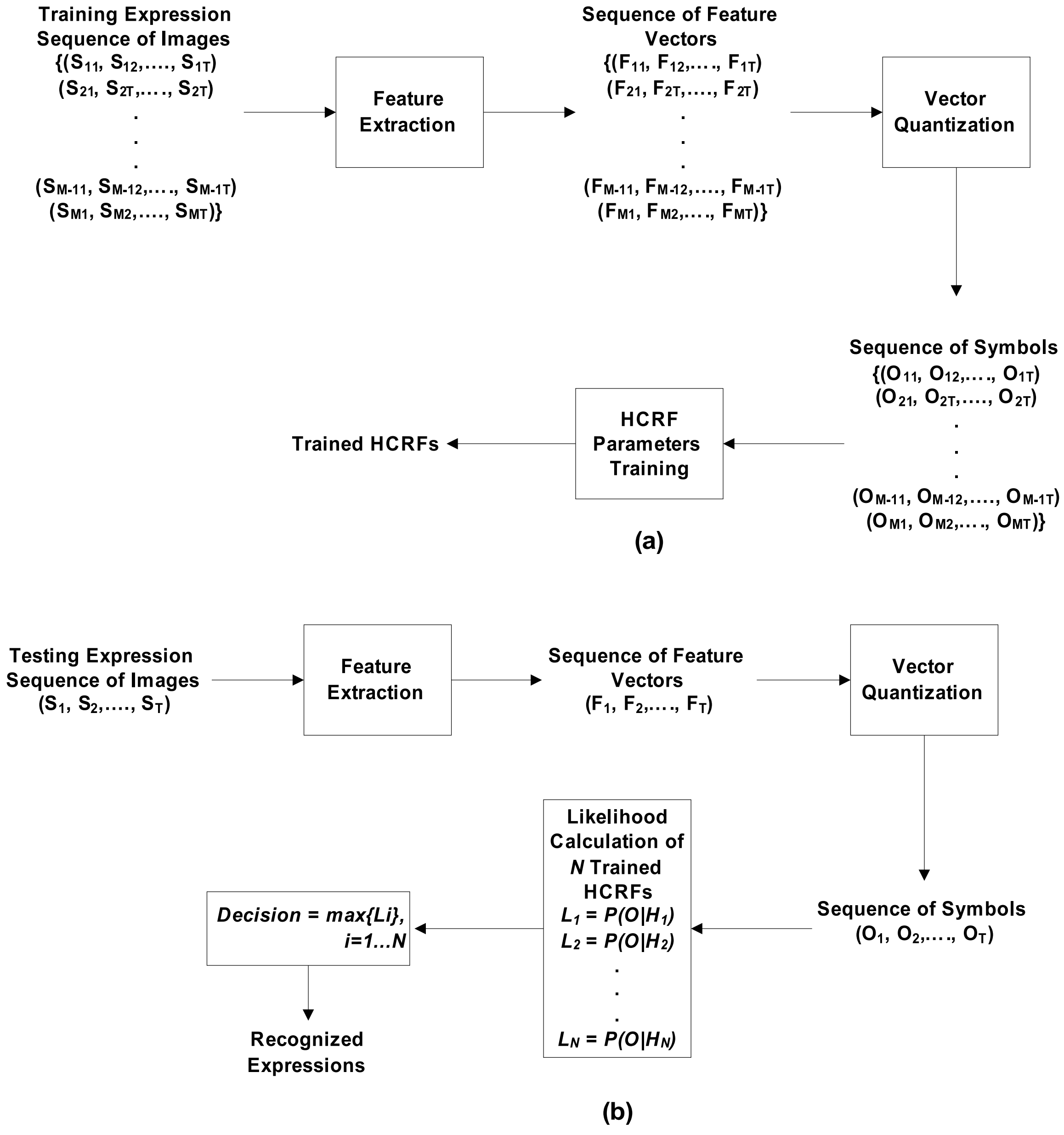

Training Expression Sequence of Images
{(S_{11}, S_{12},...., S_{1T})
(S_{21}, S_{2T},...., S_{2T})
.
.
.
(S_{M-11}, S_{M-12},...., S_{M-1T})
(S_{M1}, S_{M2},...., S_{MT})}
Feature Extraction
Sequence of Feature Vectors
{(F_{11}, F_{12},...., F_{1T})
(F_{21}, F_{2T},...., F_{2T})
.
.
.
(F_{M-11}, F_{M-12},...., F_{M-1T})
(F_{M1}, F_{M2},...., F_{MT})}
Vector Quantization
Sequence of Symbols
{(O_{11}, O_{12},...., O_{1T})
(O_{21}, O_{2T},...., O_{2T})
.
.
.
(O_{M-11}, O_{M-12},...., O_{M-1T})
(O_{M1}, O_{M2},...., O_{MT})}
HCRF Parameters Training
Trained HCRFs
(a)
Testing Expression Sequence of Images
(S_1, S_2,...., S_T)
Feature Extraction
Sequence of Feature Vectors
(F_1, F_2,...., F_T)
Vector Quantization
Sequence of Symbols
(O_1, O_2,...., O_T)
Likelihood Calculation of *N* Trained HCRFs
$L_1 = P(O|H_1)$
$L_2 = P(O|H_2)$
.
.
.
$L_N = P(O|H_N)$
Decision = max{Li}, i=1...N
Recognized Expressions
(b)

Figure 3.12: (a) HCRF training for an expression, while (b) testing an expression in an image sequence.

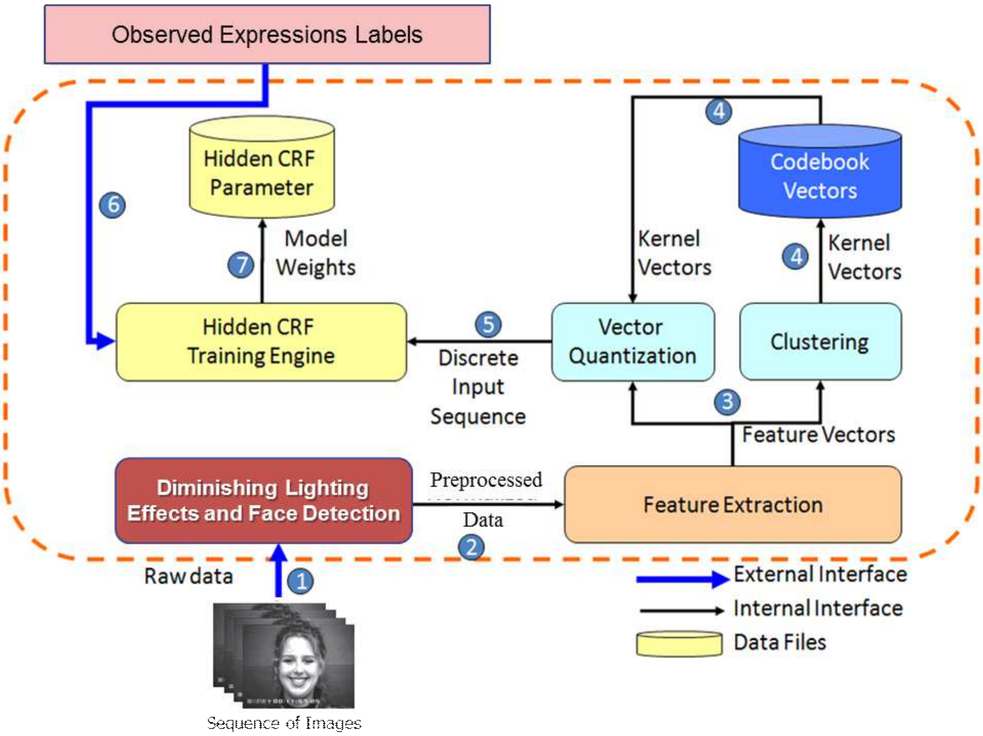


Figure 3.13: The overall flow diagram of the proposed FER methodology.

Chapter 4

Experimental Setup

4.1 Pose Based Datasets

The following pose-based datasets of facial expressions were utilized in this dissertation

4.1.1 Cohn-Kanade Dataset

In this facial expressions dataset, there were 100 subjects (university students) performed basic six expressions. The age range of the subjects were from 18 to 30 years and most of them were female. We employed those expression for which the camera was fixed in front of the subjects. By the given instructions, the subjects performed a series of 23 facial displays. Six expressions were based on descriptions of prototypic emotions such as happy, anger, sad, surprise, disgust, and fear. In order to utilize these six expressions from this dataset, we employed total 450 image sequences from 100 subjects, and each of them was considered as one of the six basic universal expressions. The size of each facial frame was 640×480 or 640×490 pixel with 8-bit precision for gray scale values. For recognition purpose, twelve expression frames were taken from each expression sequence, which resulted in a total of 5,400 expression images.

4.1.2 JAFFE Dataset

We also employed Japanese Female Facial Expressions (JAFFE) dataset in order to assess the performance of the proposed hierarchical recognition system. The expressions in this dataset were performed by 10 different (Japanese female) subjects. Each image has been rated on six expression adjectives by 60 Japanese subjects. Most of the expression frames were taken from the frontal view of the camera with tied hair in order to expose all the sensitive regions of the face. In the whole dataset, there were total 213 facial frames, which consists of seven expressions including neutral. Therefore, we selected 193 expression frames for six facial expressions performed by ten different Japanese female as subjects. The size of each facial frame was 256×256 pixels.

4.1.3 Extended Cohn-Kanade (CK+) Dataset

This facial expressions dataset contains 593 video sequences on seven facial expressions recorded from 123 subjects (university students). The age range of the subjects was from 18 to 30 years and most of them were female. Out of 593 video sequences, 309 were used in this work. The original size of each facial frame in some of the images is 640×480, and 640×490 pixel in others, with 8-bit precision for grayscale values.

4.1.4 USTC-NVIE Dataset

In this dataset, an infrared thermal and a visible camera was used in order to collect both spontaneous and posed expressions, but we utilized only posed-based expressions. There were 108 subjects, and their age range was from 17 to 31 years. Some of them worn glasses, whereas others were free of glasses. They were asked to perform a series of expressions with illumination from three different directions. The size of each facial frame was 640×480 or 704×490 pixels. In total, 1027 expression frames were utilized from this dataset.

4.1.5 MUG (Multimedia Understanding Group) Dataset

This is a pose-based dataset, in which 86 subjects were performed the six basic expressions with constant blue background with the frontal view of the camera. Two light sources of 300W each, mounted on stands at a height of 130cm approximately were used. A predefined setup with the help of umbrella was utilized in order to diffuse light and avoid shadow. The images were captured at a rate of 19 frames per second. The original size of each image was 896×896 pixels.

4.1.6 MMI Dataset

The MMI dataset of facial expressions is a fully web-searchable collection of visual and audio-visual recordings of subjects displaying a facial expression. This dataset contains a total of 238 video sequences performed by 28 subjects (male and female). The original size of each facial frame is 720×576 pixel.

4.2 Spontaneous Dataset

The following spontaneous datasets of facial expressions were utilized in this dissertation

4.2.1 USTC-NVIE spontaneous-based Dataset

In USTC-NVIE) [124] dataset, an infrared thermal and a visible camera was used in order to collect both spontaneous and posed expressions, but here, we only utilized the spontaneous-based expressions. There were 105 subjects under front illumination, 111 subjects under left illumination, and 112 subjects under right illumination, and their age range was from 17 to 31 years. Some of them worn glasses, whereas others were free of glasses. They performed a series of expressions with illumination from three different directions. The size of each facial frame was 640×480 or 704×490 pixels. In total, 910 expression frames were utilized from this dataset.

4.2.2 Indian Movie Face Database (IMFDB)

Indian Movie Face Database (IMFDB) [125] was collected from Indian movies of different languages. Most of the videos were collected from the last two decades which contain large diversity in age, illumination, and resolution. In IMFDB, the subjects were partially makeup and over-makeup. The images were from frontal, left, right, up, and down views of camera. The dataset has basic six expressions such happy, anger, sad, disgust, fear, and surprise with bad, medium, and high illumination. Images were taken from 67 male and 33 female actors of different age such as child (1–12 years), young (13–30 years), middle (31–50 years), and older (Above 50 years) with at least 200 images from each actor. Some subjects have glasses, beard, ornaments, hair, hand, or none. In order to maintain consistency among the images, a heuristic method for cropping was applied, and all the images were manually selected and cropped from the video frames resulting in a high degree of variability in terms of scale, pose, expression, illumination, age, resolution, occlusion, and makeup. The dataset consists of total 34512 images of 100 Indian actors collected from more than 100 videos. The size of each image which we used for our experiments was 140×180 pixels.

4.2.3 Radboud Faces Database (RaFD)

RaFD dataset [126] is a set of pictures of 67 models such as Caucasian males and females, Caucasian children, boys and girls, and Moroccan Dutch males, which presenting eight expressions. In this study, for a thorough validation, we have six basic expressions like happy, anger, sad, surprise, fear, and disgust. This dataset was created the Behavioral Science Institute of the Radboud University Nijmegen, which is located in Netherland, and can be used freely for non-commercial scientific research by researchers who work for an officially attributed university.

RaFD dataset is a high quality faces database, and accordingly to the Facial Action Coding System (FAUS), each model was trained to show all the expressions. Each emotion was shown with three different stare directions and all the pictures were taken from five camera angles simultaneously. The total images in RaFD dataset were 8040 for eight expressions. But in our study, we utilized total 5300 images for six expressions. In this study, the size of each expression image was 140×180 pixels.

4.3 Hierarchical Recognition System Validation using Pose Datasets

In order to assess the hierarchical recognition system, six universal expressions like: happy, sad, surprise, disgust, anger, and fear were used from two publicly available standard datasets. These datasets display the frontal view of the face, and each expression is composed of several sequences of expression frames. During each experiment, we reduced the size of each input image (expression frame) to 60×60, where the images were first converted to a zero-mean vector of size 1×3,600 for feature extraction. All the experiments were performed in Matlab using an Intel^R Pentium^R Dual-Core^{TM} (2.5 GHz) with a RAM capacity of 3 GB.

For a thorough validation, we performed the following setup for the corresponding experiments.

4.3.1 Recognition Rate on Individual Dataset

In the first experiment, the performance of the proposed hierarchical recognition system on each dataset (based on subjects) was analyzed. For each dataset, a $10-$fold cross-validation scheme (based on subjects) was used. In other words, out of 10 subjects data from a single subject was used as the validation data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the validation data.

4.3.2 Recognition Rate under the Absence of Hierarchical Scheme

In the second experiment, a set of experiments were performed in order to assess the effectiveness of the hierarchical recognition scheme in the proposed hierarchical recognition system. This experiment was performed on each dataset and the recognition performance was analyzed under the absence of the proposed hierarchical recognition process.

4.4 Experimental Setup for the Proposed Techniques on Pose Datasets

As described earlier, the motivation behind this work is to build an accurate and robust techniques whose accuracy is not affected by noise, race, and gender of subjects in a given image. Therefore, we have contributed in feature extraction and classification modules. For a thorough validation, five different experiments were performed.

4.4.1 First Experiment: Setup for the Face Detection and Extraction

In this experiment, the performance of the proposed face detection and extraction technique was analyzed.

4.4.2 Second Experiment: Setup Based on Subjects

In this experiment, performance of the proposed FER system (such as feature extraction, feature selection, and recognition) was validated using the four datasets. For each dataset, a $10-$fold cross-validation scheme (based on subjects) was used. In other words, out of 10 subjects data from a single subject was used as the validation data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the validation data.

4.4.3 Third Experiment: Setup Based on Dataset

In this experiment, $n-$fold cross-validation rule based on dataset was performed (in our case n =4). It means that from the four datasets, data from the three datasets were retained as the validation data for testing the system, and the data from the remaining dataset was used as the training data. This process was repeated four times, with data from each dataset used exactly once as the training data.

4.4.4 Fourth Experiment: Setup under the Absence of Each Module

In this experiment, a set of three sub-experiments were performed in order to show the effectiveness of sub-components of the proposed FER system, i.e., feature extraction (waveleet transform coupled with optical flow), and recognition model (HCRF). For this purpose, again 10-fold validation rule was used on each dataset. In the first case, ICA (a well-known local feature extraction technique) was utilized with HCRF instead of wavelet with optical. In the second case, the existing HCRF [2] model was used with the proposed feature extraction instead of using the proposed HCRF model.

4.4.5 Fifth Experiment: Comparison with the Existing Systems

Finally, in the fifth experiment, the performance of proposed FER system was compared against the state-of-the-art FER systems.

4.5 Experimental Setup for the proposed Techniques using Spontaneous Datasets

For a thorough validation, the following set of experiments were performed on spontaneous datasets.

4.5.1 First Experiment: Recognition Rate on Individual Dataset

In this experiment, the performance of the proposed techniques such as the proposed feature extraction, and recognition methods were tested and validated on existing spontaneous datasets. A $10-$fold cross-validation scheme (based on subjects) was utilized for each dataset, which means that out of 10 subjects data from a single subject was used as the validation data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the validation data.

4.5.2 Second Experiment: Based on Dataset

In this experiment, $n-$fold cross-validation rule based on dataset was performed (in our case n =3). It means that from the three datasets, data from the two datasets were retained as the validation data for testing the system, and the data from the remaining dataset was used as the training data. This process was repeated three times, with data from each dataset used exactly once as the training data.

4.5.3 Third Experiment: Experimental Setup under the Absence of Each Module

In this experiment, a set of experiments were performed in order to show the effectiveness of sub-components, i.e., wavelet transform with optical flow, and HCRF. For this purpose, four sub-experiments were performed on each dataset using the 10-fold validation rule. In the first two sub-experiments, ICA (a well-known feature extraction technique) was utilized with the proposed HCRF instead of using the proposed feature extraction method (i.e., wavelet transform with optical flow). In the next two sub-experiments, the existing HCRF [2] was used with wavelet transform coupled with optical flow instead of using the proposed HCRF model.

4.5.4 Fourth Experiment: Comparison with the Existing FER Systems

Lastly, in this experiment, the performance of the proposed techniques was compared with some well-known existing state-of-the-art FER systems. We borrowed the implementations for some of the methods, whereas the other methods were implemented by us for fair comparison.

Chapter 5

Experimental Results and Discussions

5.1 Experimental Results of the Proposed Hierarchical Recognition System

The two experiments that were performed for the proposed hierarchical system are given below.

5.1.1 Recognition Rate on Each Dataset

In this experiment, the proposed hierarchical system was validated on two different datasets. Each dataset possessed different facial features, such as the facial features of the Cohn-Kanade dataset are quite different from the facial features of JAFFE dataset. Which means that the eyes of the subjects in JAFFE dataset are totally different from the eyes of the subjects of Cohn-Kanade dataset. The proposed system was evaluated for each dataset separately that means for each dataset, 10-fold cross-validation rule (based on subjects) was applied. It means that out of ten subjects, data from a single subject was retained as the validation data for testing the proposed system, whereas the data for the remaining nine subjects were used as the training data. This process was repeated ten times, with data from each subject used exactly once as the validation data. The detailed results of this experiment for the two datasets are shown in Tables 5.1 and 5.2, respectively.

It can be seen that the proposed hierarchical recognition system consistently achieved a high recognition rate when applied on these datasets separately, i.e., 98.87% on Cohn-Kanade, and 98.83% on JAFFE datasets respectively. This means that, unlike Zia *et al.* [55], the proposed hierarchical recognition system is robust i.e., the system not only achieves high recognition rate on one dataset but shows the same performance on other datasets as well.

Table 5.1: Confusion matrix of the proposed hierarchical recognition sysetm using Cohn-Kanade dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	98.4	1.6	0	0	0	0
Sad	1.1	98.9	0	0	0	0
Anger	0	0	99	0	0	1
Disgust	0	0	0	98.4	1.6	0
Surprise	0	0	0	0	100	0
Fear	0	0	1.5	0	0	98.5
Average			98.87			

Table 5.2: Confusion matrix of the proposed hierarchical recognition system using Japanese (JAFFE) dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	99	1	0	0	0	0
Sad	2	98	0	0	0	0
Anger	0	0	100	0	0	0
Disgust	0	0	0	99	1	0
Surprise	0	0	0	3	97	0
Fear	0	0	0	0	0	100
Average			98.83			

5.1.2 Recognition Rate under the Absence of the Proposed Hierarchical Scheme

In this experiment, a set of experiments was performed to assess the effectiveness of the hierarchical scheme in the proposed hierarchical recognition FER system. This experiment was repeated two times and the recognition performance was analyzed on the two datasets. In each experiment, a single LDA and HMM were used to recognize all the expressions instead of using the hierarchical scheme. The results for the Cohn-Kanade and JAFFE datasets are shown in Tables 5.3 and 5.4, respectively.

It can be seen from Tables 5.3 and 5.4 that the hierarchical scheme is important and mainly responsible for the high recognition accuracy of the the proposed hierarchical recognition system.

Table 5.3: Confusion matrix of ICA+LDA and HMM on Cohn-Kanade dataset, while removing hierarchical recognition step (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	89	2	0	0	4	5
Sad	0	92	4	4	0	0
Anger	0	5	90	5	0	0
Disgust	0	0	11	89	0	0
Surprise	4	0	0	6	90	0
Fear	0	2	9	0	0	89
Average			89.8			

Table 5.4: Confusion matrix of ICA+LDA and HMM on Japanese (JAFFE) dataset, while removing hierarchical recognition step (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	82	5	3	2	5	3
Sad	0	90	3	2	3	2
Anger	1	3	93	2	0	1
Disgust	4	0	6	87	0	3
Surprise	2	4	0	0	88	0
Fear	2	0	6	7	0	85
Average			87.5			

When we removed the hierarchical recognition module, the recognition rate decreased significantly. These results support the theory that the problem of high similarity among the features of different expressions is a local problem. In other words, the features exist in the form of groups in overall feature space. The expressions within one group are very similar, whereas they are easily distinguishable from those in the other groups; therefore, to overcome this problem in an effective manner, these groups (or expression categories) should be separated first and then techniques like LDA should be applied to each category separately.

5.2 Experimental Results of the Proposed Techniques using Pose Datasets

The recognition results for the five different experiments are presented below.

5.2.1 First Experiment: Results of Face Detection and Extraction

It should be noted that in the proposed FER system, active contour evolution in a certain frame is performed independently of the other frames. It means that the face detection is performed on frame-by-frame bases. In any given frame, the only information utilized from the previous frame is the final contour obtained in the previous frame. This information is used to determine the initial position of the active contour in the current frame. First, an ellipse with major axis along y-axis of length 20 and minor axis along x-axis of length 20 is selected as the initial contour. In this experiment, the initial shape was the same for all frames, and only the center location varied. In each video sequence, the first frame is segmented using manual initialization such that the initial contour is closer to the face.

Then from the second frame, the position of the initial contour's center in the current frame is the mean value of the points along the final contour in the previous frame. For example, along the final contour of frame $n(n \geq 1)$, there are M points $\left(\mathrm{x}_\mathrm{i}^{(\mathrm{n})}, \mathrm{y}_\mathrm{i}^{(\mathrm{n})}\right), \mathrm{i} = 1..\mathrm{M}$. Then, the center $(\mathrm{c}_\mathrm{x}^{(\mathrm{n+1})}$, $\mathrm{c}_\mathrm{y}^{(\mathrm{n+1})})$ of the initial contour in the frame $(n+1)$is calculated as

$$\begin{pmatrix} (n+1) \\ c \\ x \end{pmatrix} = \frac{1}{M}\sum_{i=1}^{M} \overset{(n)}{\underset{i}{x}}; \begin{pmatrix} (n+1) \\ c \\ y \end{pmatrix} = \frac{1}{M}\sum_{i=1}^{M} \overset{(n)}{\underset{i}{y}} \tag{5.1}$$

Here we have explained how the contours are calculated, and the results are shown in Figure 5.1 (a), which compares the performance of the proposed AC model with that of the Chan-Vese active contour (CV AC) model. It is obvious that the proposed AC model showed better performance than the CV AC model as shown in Figure 5.1(b).

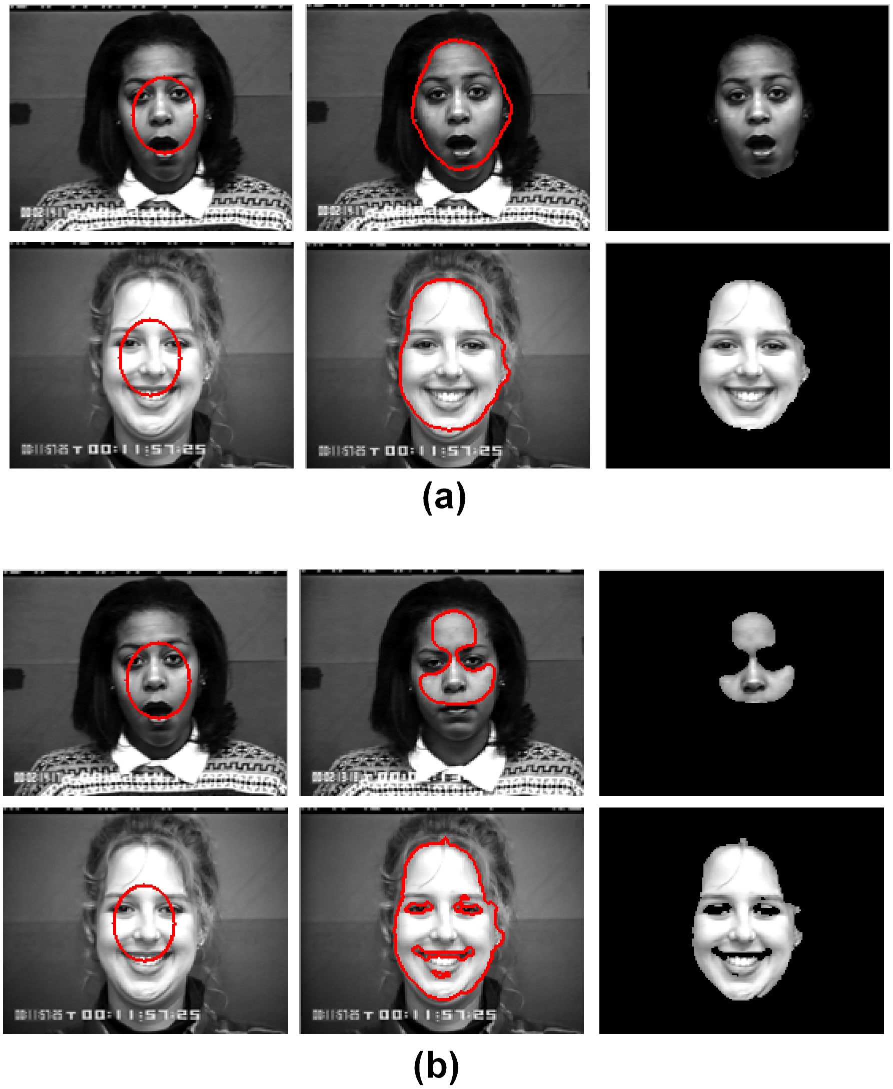


Figure 5.1: (a) Sample results of the proposed AC model for which $\gamma = 0.5$ and $\beta = 0.2$, while (b) indicates the sample results of CV AC model using $\gamma = 0.5$ and $\beta = 1.0$. Left image shows the initial contour, middle image represents the final contour, and last image presents the extracted face.

5.2.2 Second Experiment: Classification Results on Individual Dataset

The purpose of this experiment was to analyze the performance of the proposed techniques for all the four datasets. The recognition results for this experiment are shown in Table 5.5–Figure 5.2 (using Extended Cohn-Kanade (CK+) dataset), Table 5.6–Figure 5.3 (using USTC-NVIE dataset), Table 5.7–Figure 5.4 (using MUG dataset), and Table 5.8–Figure 5.5 (using MMI dataset).

Table 5.5: Confusion matrix of the proposed techniques using Extended Cohn-Kanade (CK+) dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	97	3	0	0	0	0
Sad	2	98	0	0	0	0
Anger	0	0	97	0	0	3
Disgust	0	0	0	98	2	0
Surprise	0	0	0	4	96	0
Fear	0	0	4	0	0	96
Average			97.0			

Table 5.6: Confusion matrix of the proposed techniques using USTC-NVIE dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	98	2	0	0	0	0
Sad	3	97	0	0	0	0
Anger	0	0	98	0	0	2
Disgust	0	0	0	97	3	0
Surprise	0	0	0	1	99	0
Fear	0	0	6	0	0	94
Average			97.16			

It is clear from Tables 5.5, 5.6, 5.7, and 5.8 that the proposed techniques consistently achieved a high recognition rate when applied to these datasets separately: 97.0% for the Extended Cohn-Kanade (CK+) dataset, 97.16% for USTC-NVIE dataset, 97.0% for MUG dataset, and 97.83% for MMI dataset. This indicates that the results are quite satisfactory for all the datasets.

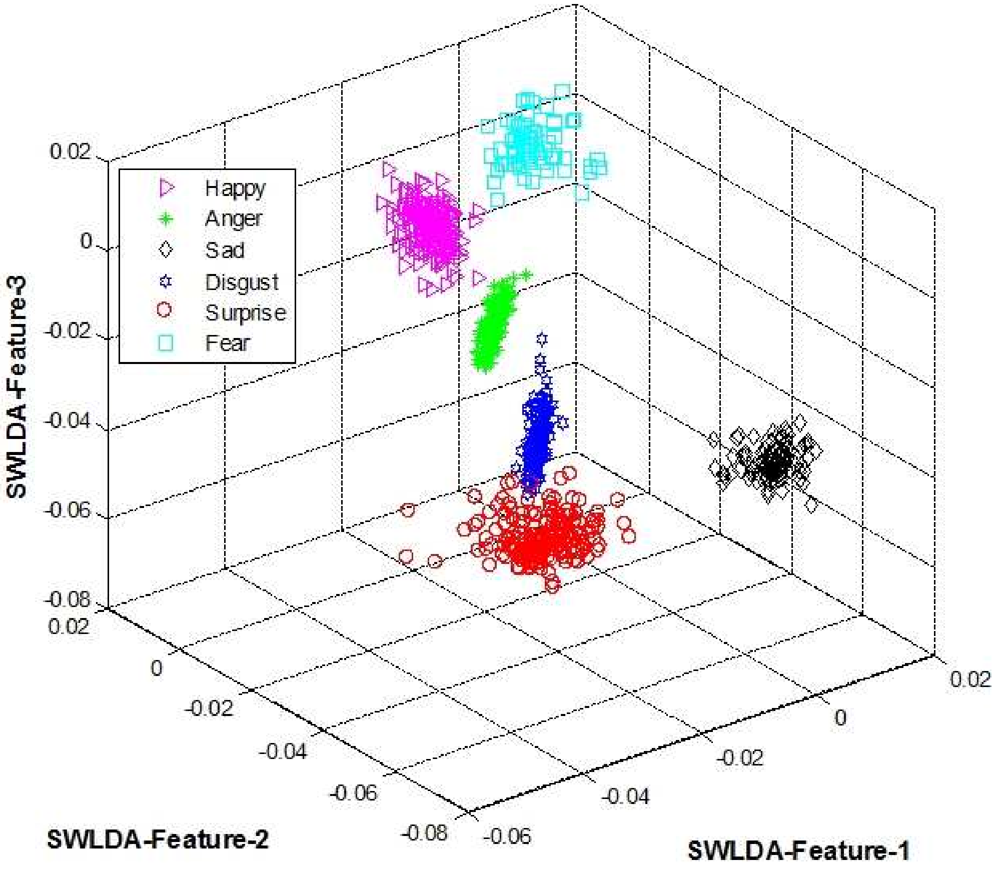


Figure 5.2: 3D-feature plot of the proposed techniques for six facial expressions. It is indicated that the proposed techniques provided best classification rate on Extended Cohn-Kanade (CK+) dataset.

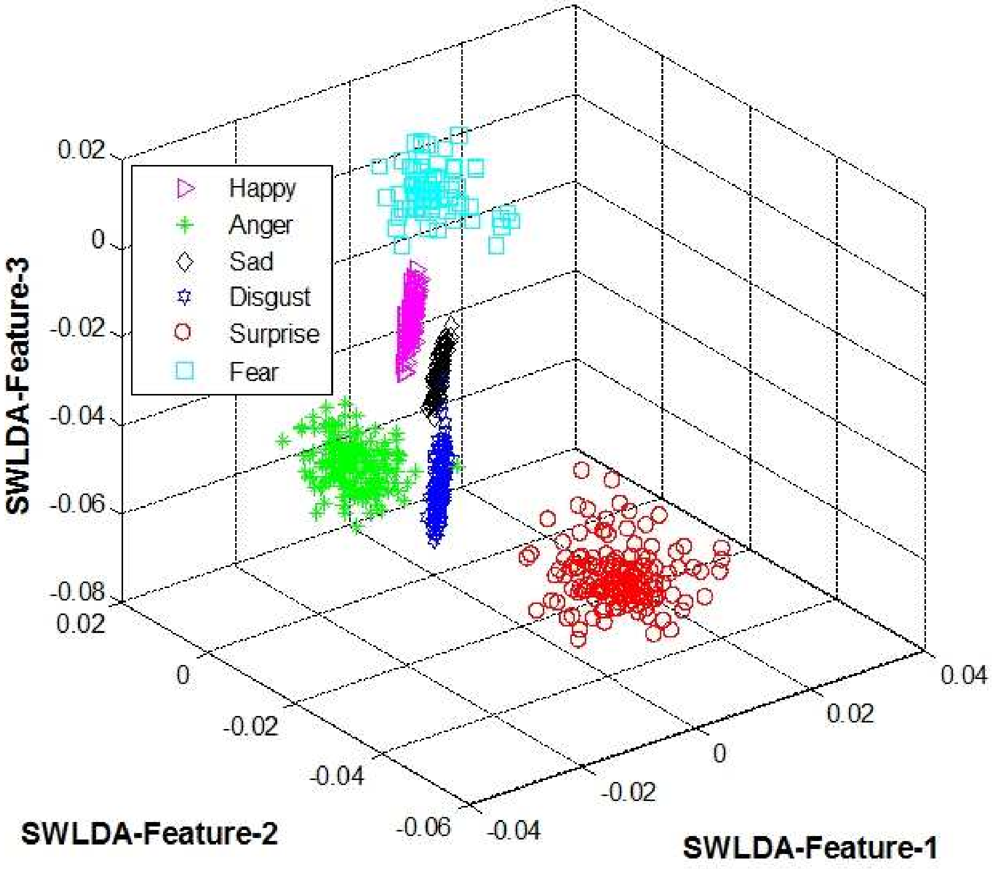


Figure 5.3: 3D-feature plot of the proposed techniques for six facial expressions. It is indicated that the proposed techniques provided better classification rate on USTC-NVIE dataset.

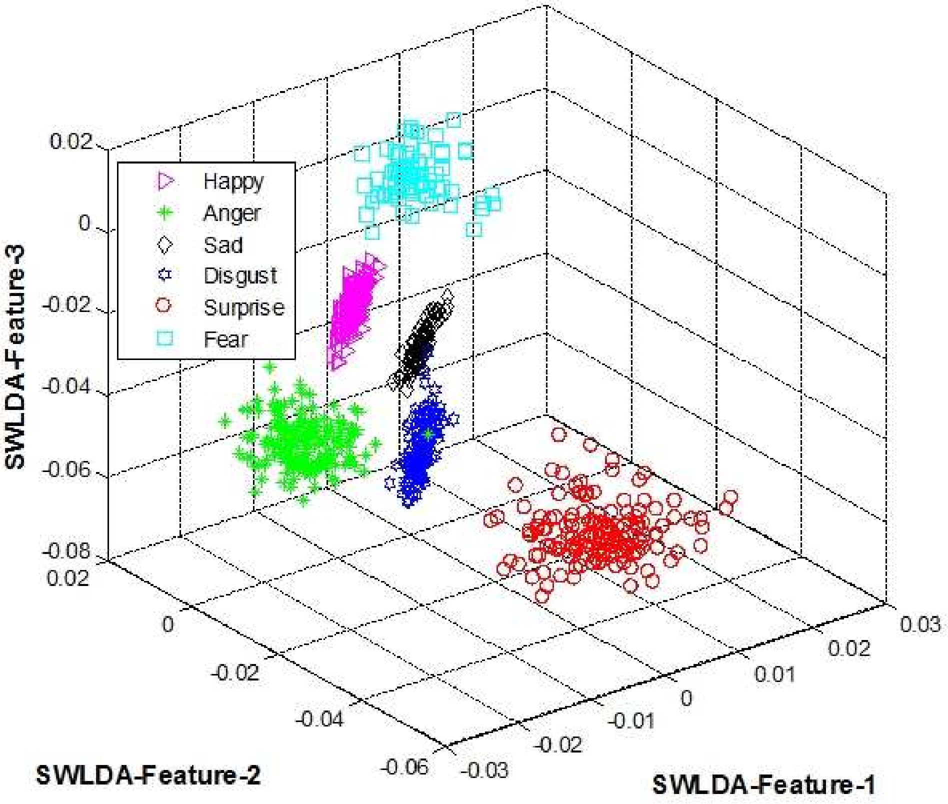


Figure 5.4: 3D-feature plot of the for six facial expressions. It is indicated that the proposed techniques provided better classification rate on MUG dataset.

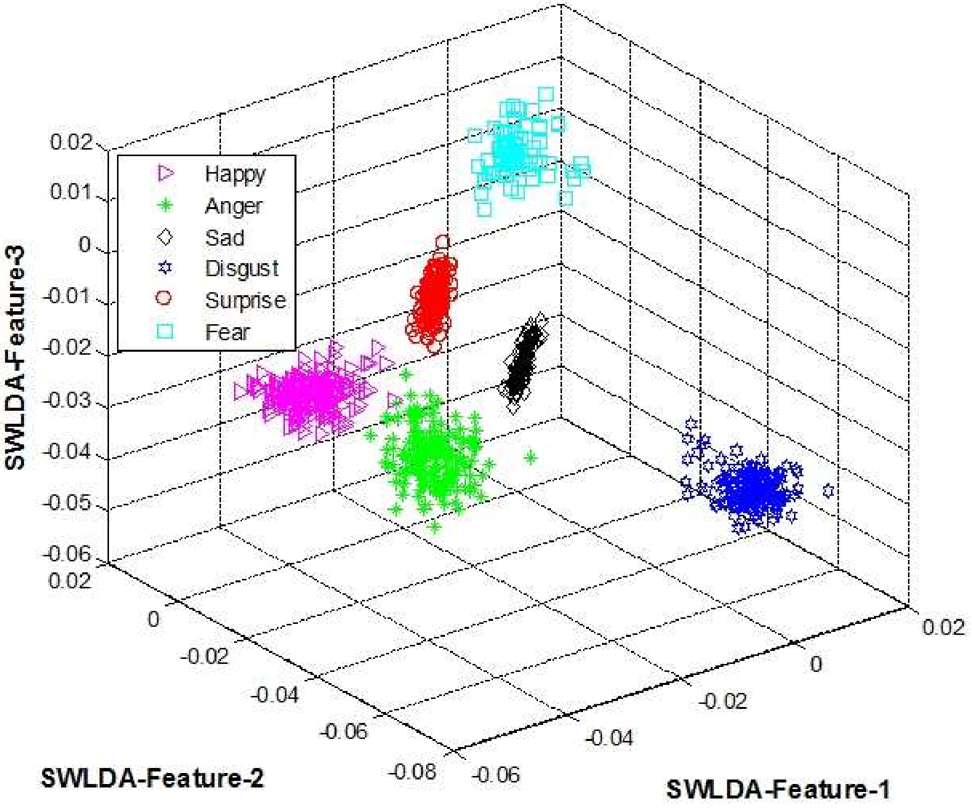


Figure 5.5: 3D-feature plot of the proposed techniques for six facial expressions. It is indicated that the proposed methods provided better classification rate on MMI dataset.

Table 5.7: Confusion matrix of the proposed techniques using MUG dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	96	4	0	0	0	0
Sad	3	97	0	0	0	0
Anger	0	0	95	0	0	5
Disgust	0	0	0	99	1	0
Surprise	0	0	0	3	97	0
Fear	0	0	2	0	0	98
Average			97.0			

Table 5.8: Confusion matrix of the proposed techniques using MMI dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	98	0	1	0	1	0
Sad	1	97	0	1	0	1
Anger	0	0	98	0	2	0
Disgust	0	2	0	97	1	0
Surprise	0	1	0	0	99	0
Fear	0	1	0	1	0	98
Average			97.83			

5.2.3 Third Experiment: Classification Results Across the Datasets (Robustness)

In this experiment, the proposed methods were validated using a cross-validation scheme based on datasets. The overall results are shown in Tables 5.9, 5.10, 5.11, and 5.12, respectively.

It is clear from Tables 5.9, 5.10, and 5.11 that the proposed methods achieved a high recognition rate when it was trained on CK+, USTC-NVIE, and MUG datasets. However, the methods achieved low accuracy of classification when it was trained on the MMI dataset (shown in Table 5.12). This may be because the datasets have different facial features; for instance, some of the subjects in the MMI face dataset have worn glasses, while subjects in the CK+, USTC-NVIE, and MUG datasets did not wear glasses. Furthermore, eye features in the MMI dataset are very

Table 5.9: Confusion matrix of the proposed techniques training on Extended Cohn-Kanade (CK+) dataset and testing on USTC-NVIE, MUG, and MMI datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	87	3	2	3	2	3
Sad	3	86	2	4	3	2
Anger	0	3	88	4	3	2
Disgust	1	2	2	90	4	1
Surprise	5	3	3	1	85	3
Fear	3	2	1	1	4	89
Average			87.50			

Table 5.10: Confusion matrix of the proposed techniques training on USTC-NVIE dataset and testing on CK+, MUG, and MMI datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	89	2	1	2	4	2
Sad	3	86	2	4	3	2
Anger	4	3	87	1	2	3
Disgust	2	1	2	89	3	3
Surprise	2	1	3	2	90	2
Fear	3	3	3	1	4	86
Average			87.83			

different from those in the eyes of the CK+, USTC-NVIE, and MUG datasets. Similarly, some expressions in MMI datasets were taken by using different angles of the camera, while other datasets have the expressions only on frontal view of the camera, and therefore achieved low accuracy when it was trained on MMI dataset (shown in Table 5.12). Nevertheless, the results are very encouraging and this suggests that the proposed techniques is robust, i.e., the methods not only achieved a high recognition rate on one dataset, but also provided good recognition rates when used across multiple datasets.

Table 5.11: Confusion matrix of the proposed techniques training on MUG dataset and testing on CK+, USTC-NVIE, and MMI datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	90	2	1	2	3	2
Sad	2	86	3	3	4	2
Anger	3	1	87	4	3	2
Disgust	1	3	2	89	3	2
Surprise	2	2	4	2	88	2
Fear	2	1	3	4	1	89
Average			88.17			

5.2.4 Fourth Experiment: Results Under the Absence of Each Module

In this experiment, a set of sub-experiments was performed in order to show the importance of each module (feature extraction, feature selection, and recognition model). For this purpose, all the experiments were performed using all the datasets under the absence of each respective module.

5.2.4.1 Experimental Results Under the Absence of the Proposed Feature Extraction Technique

For the first case, ICA (a well-known local feature extraction technique) was utilized with SWLDA and HCRF instead of using the proposed feature extraction (i.e., wavelet transform with optical).

Table 5.12: Confusion matrix of the proposed techniques training on MMI dataset and testing on CK+, USTC-NVIE, and MUG datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	82	4	3	2	4	5
Sad	3	83	3	5	4	2
Anger	2	2	85	3	4	4
Disgust	2	0	3	87	4	4
Surprise	4	5	3	4	80	4
Fear	3	5	4	3	3	82
Average			83.17			

The overall results for the first case using all the four datasets are shown in Tables 5.13, 5.14, 5.15, and 5.16, respectively.

Table 5.13: Confusion matrix of the ICA+SWLDA with HCRF using Extended Cohn-Kanade (CK+) dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	93	2	1	2	1	1
Sad	2	94	0	2	1	1
Anger	0	0	96	0	3	1
Disgust	0	3	0	97	0	0
Surprise	0	2	3	2	93	0
Fear	0	0	1	2	0	96
Average			94.83			

It is clear from Tables 5.13, 5.14, 5.15, and 5.16 that without using the proposed feature extraction method (wavelet transform with optical flow), the system was not able to achieve best accuracy of classification. The existing SWLDA is a robust feature selection technique that addresses the limitations of previous feature selection techniques, especially PCA, LDA, KDA, and GDA. However, still it is unable to get high recognition rate without the proposed feature extraction method. This is because, the proposed feature extraction method is a compactly supported

Table 5.14: Confusion matrix of the ICA+SWLDA with HCRF using USTC-NVIE dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	92	2	1	3	1	1
Sad	1	93	3	2	1	0
Anger	0	1	94	2	0	3
Disgust	0	3	0	91	2	4
Surprise	1	4	0	3	92	0
Fear	1	4	2	2	1	90
Average			92.00			

Table 5.15: Confusion matrix of the ICA+SWLDA with HCRF using MUG dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	94	3	0	2	0	1
Sad	1	93	1	3	2	0
Anger	0	3	94	0	0	3
Disgust	2	2	0	92	4	0
Surprise	2	0	0	2	95	1
Fear	1	2	3	0	1	93
Average			93.50			

Table 5.16: Confusion matrix of the ICA+SWLDA with HCRF using MMI dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	92	4	0	2	0	2
Sad	2	90	2	3	1	2
Anger	1	1	93	2	3	0
Disgust	2	3	0	91	1	3
Surprise	2	1	2	2	89	4
Fear	2	0	3	2	1	92
Average			91.16			

wavelet on gray scale images with the least asymmetry and highest number of vanishing moments for a given support width. The symlet wavelet has the capability to support the characteristics of orthogonal, biorthogonal, and reverse biorthogonal of gray scale images, that's why it provides better classification results. The frequency-based assumption is supported in our experiments. We measure the statistic dependency of wavelet coefficients for all the facial frames of gray scale. Joint probability of a gray scale frame is computed by collecting geometrically aligned frames of the expression for each wavelet coefficient. Symlet wavelet transform is capable to extract prominent features from gray scale images with the aid of locality in frequency, orientation, and in space as well. Since wavelet is a multi-resolution that helps us to efficiently find the images in coarse-to-find way. The motion of the pixels in some parts of the face have much contribution in making the expression. Therefore, in the proposed feature extraction method, we also incorporated the optical flow, due to which we can find the motion information of the pixels that improve the recognition rate.

In the second case, wavelet with optical flow was coupled with LDA (a well-known discriminant analysis approach) before feeding the features to HCRF. The results for this case are presented in Tables 5.17, 5.18, 5.19, and 5.20, respectively.

Table 5.17: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using Extended Cohn-Kanade (CK+) dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	93	2	2	0	3	0
Sad	2	90	2	4	2	0
Anger	2	0	95	1	2	0
Disgust	3	4	0	90	2	1
Surprise	1	2	3	3	91	0
Fear	0	0	2	5	1	92
Average			91.83			

Table 5.18: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using USTC-NVIE dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	93	3	2	1	0	1
Sad	0	95	3	0	0	2
Anger	2	0	90	2	3	3
Disgust	2	1	3	91	2	1
Surprise	1	2	2	0	92	3
Fear	0	0	2	4	0	94
Average			92.50			

Similarly, it is also apparent from Tables 5.17, 5.18, 5.19, and 5.20 that without using SWLDA,

Table 5.19: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using MUG dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	95	0	2	0	2	1
Sad	0	96	2	0	2	0
Anger	0	0	97	3	0	0
Disgust	0	0	3	92	2	3
Surprise	0	3	3	0	93	1
Fear	0	2	2	0	2	94
Average			94.50			

Table 5.20: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using MMI dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	90	3	4	2	1	0
Sad	3	94	3	0	0	0
Anger	1	1	89	2	4	3
Disgust	2	3	2	91	2	0
Surprise	3	2	2	0	91	2
Fear	0	3	3	0	0	94
Average			91.50			

the system was also unable to achieve high recognition rate. Even though, the proposed feature extraction method is more robust and accurate than of the previous methods; however, still it has lack of accuracy because of mixing some informative features at the feature space, and also there might still be some redundancy among these features. Therefore, the method has been utilized in order to solve this problem. SWLDA is a robust feature selection technique that addresses the limitations of previous techniques, especially those of well-known statistical techniques such as PCA, lDA, KDA, and GDA. As shown earlier that the features for the six classes are highly merged, which can result later in a high misclassification rate. This is due to the similarity among the expressions that results in high within-class variance and low between-class variance. Therefore, SWLDA is not only provides dimension reduction, but also increases the low between-class variance to increase the class separation before the features are fed to the classifier. The proposed single-level system provided a significant improvement in the class separation, and a very low within-class variance was observed. The low within/between variations are based on forward and backward regression models in SWLDA.

5.2.4.2 Experimental Results Under the Absence of the Proposed Recognition Model

Finally, in the third case, the existing HCRF [2] was used with wavelet transform with optical flow+SWLDA instead of using proposed HCRF model. While, the results for this case are presented by Tables 5.21, 5.22, 5.23, and 5.24, respectively.

Table 5.21: Confusion matrix of the wavelet transform with optical flow+SWLDA with existing HCRF [2] using Extended Cohn-Kanade (CK+) dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	88	2	0	3	4	3
Sad	0	91	1	2	4	2
Anger	2	1	92	2	0	3
Disgust	0	2	3	93	0	2
Surprise	3	1	3	2	90	1
Fear	2	1	3	2	3	89
Average			90.50			

Likewise, it is also obvious from Tables 5.21, 5.22, 5.23, and 5.24 that when HCRF model was replaced with existing HCRF model, the system also was not able to achieve good recognition rate. Thus the proposed HCRF model successfully addresses the limitations of HMM, which has widely been used for sequential FER, and overcome the shortcomings of existing HCRF model as well.

5.2.5 Fifth Experiment: Comparison of the Proposed FER System with Existing Systems

In this experiment, performance of the proposed FER system was compared with some of the existing FER systems [33, 35–37, 127] using all the datasets, i.e., CK+, USTC-NVIE, MUG, and MMI datasets of facial expressions. We implemented all these methods using the instructions provided in their respective papers. For each dataset, 10−fold cross-validation scheme (based on subjects) was applied. The average recognition rate for each method along with the proposed FER system are presented in Table 5.25.

It can be seen from Table 5.25 that the proposed techniques outperformed the state-of-the-art. Thus, the proposed methods show significant potential in its ability to accurately and robustly recognize human facial expressions using video data.

5.3 Experimental Results on Spontaneous Datasets

As we utilized the previous spontaneous datasets of facial expressions; therefore, from now on, we will call our proposed system "proposed spontaneous methods".

This section aims at evaluating the capabilities of the proposed spontaneous methods when operated on the diverse scenarios of increasing complexity and realism devised in this work. For a

Table 5.22: Confusion matrix of the wavelet transform with optical flow+SWLDA with existing HCRF [2] using USTC-NVIE dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	89	2	4	1	3	1
Sad	1	90	2	2	2	3
Anger	2	2	92	1	1	2
Disgust	0	5	0	93	1	1
Surprise	3	2	2	3	88	2
Fear	2	2	2	3	4	87
Average			89.83			

Table 5.23: Confusion matrix of the wavelet transform with optical flow+SWLDA with existing HCRF [2] using MUG dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	88	3	4	2	2	1
Sad	4	86	2	2	3	3
Anger	2	3	89	3	1	2
Disgust	0	2	3	93	2	0
Surprise	5	4	3	1	85	2
Fear	3	2	2	1	2	90
Average			88.50			

Table 5.24: Confusion matrix of the wavelet transform with optical flow+SWLDA with existing HCRF [2] using MMI dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	86	3	5	4	2	0
Sad	2	85	2	4	3	4
Anger	0	3	87	3	5	2
Disgust	2	3	4	83	3	5
Surprise	4	0	2	3	88	3
Fear	2	1	3	2	2	90
Average			86.50			

Table 5.25: The weighted average classification results of the proposed methods with some state-of-the-art (Unit: %).

Existing Works	[35]	[36]	[33]	[37]	[127]	[75]	[128]	[129]	Proposed Approaches
Accuracy Rate	90	69	81	72	78	83	82	88	**97**

thorough validation, the following four experiments were performed.

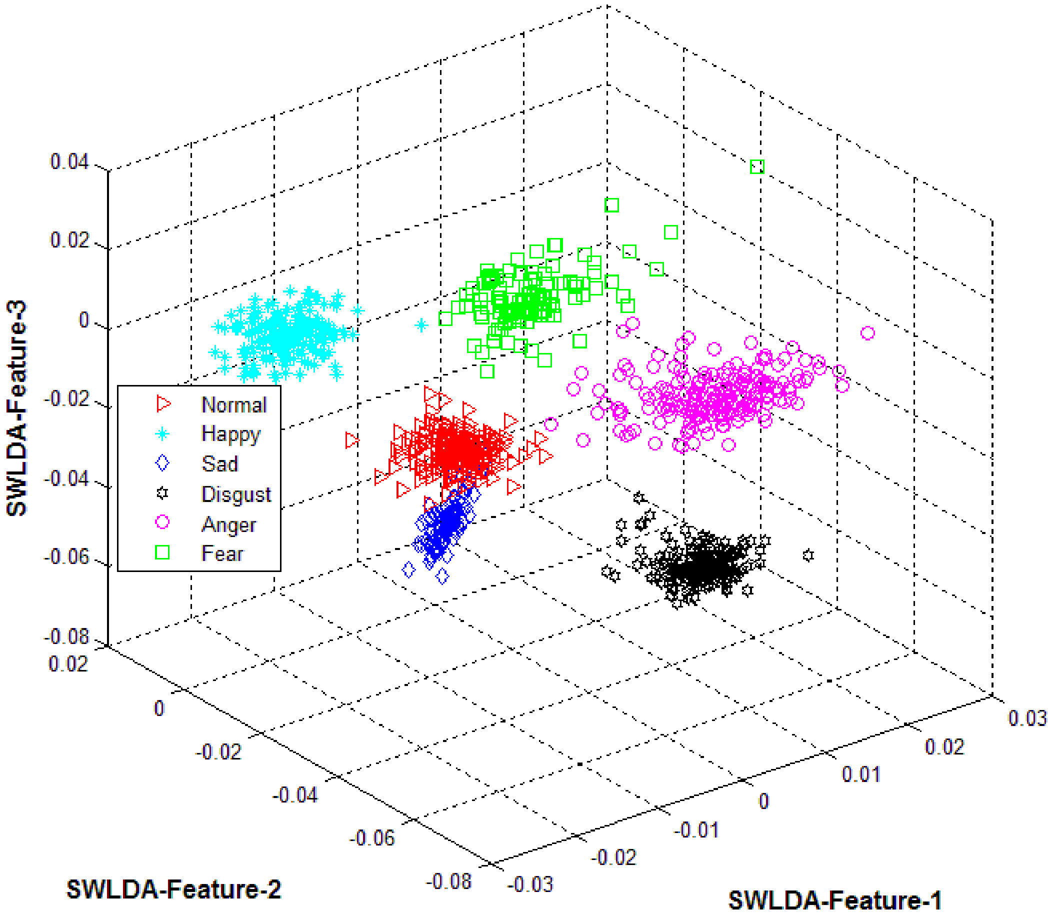


Figure 5.6: 3D feature plots for the six expressions after applying the proposed spontaneous methods on USTC-NVIE spontaneous dataset.

5.3.1 First Experiment: Recognition Rates on Individual Dataset of Facial Expressions

This experiments show the performance of the proposed spontaneous methods in naturalistic environments. Therefore, the methods were tested and validated on three publicly available spontaneous datasets of facial expressions such as USTC-NVIE and IMFDB, and Radboud Faces Database (RaFD) separately. The overall results are presented in Figure 5.6–Table 5.26 (using USTC-NVIE dataset), Figure 5.7–Table 5.27 (using IMFDB dataset), and Figure 5.8–Table 5.28 (using RaFD dataset), respectively.

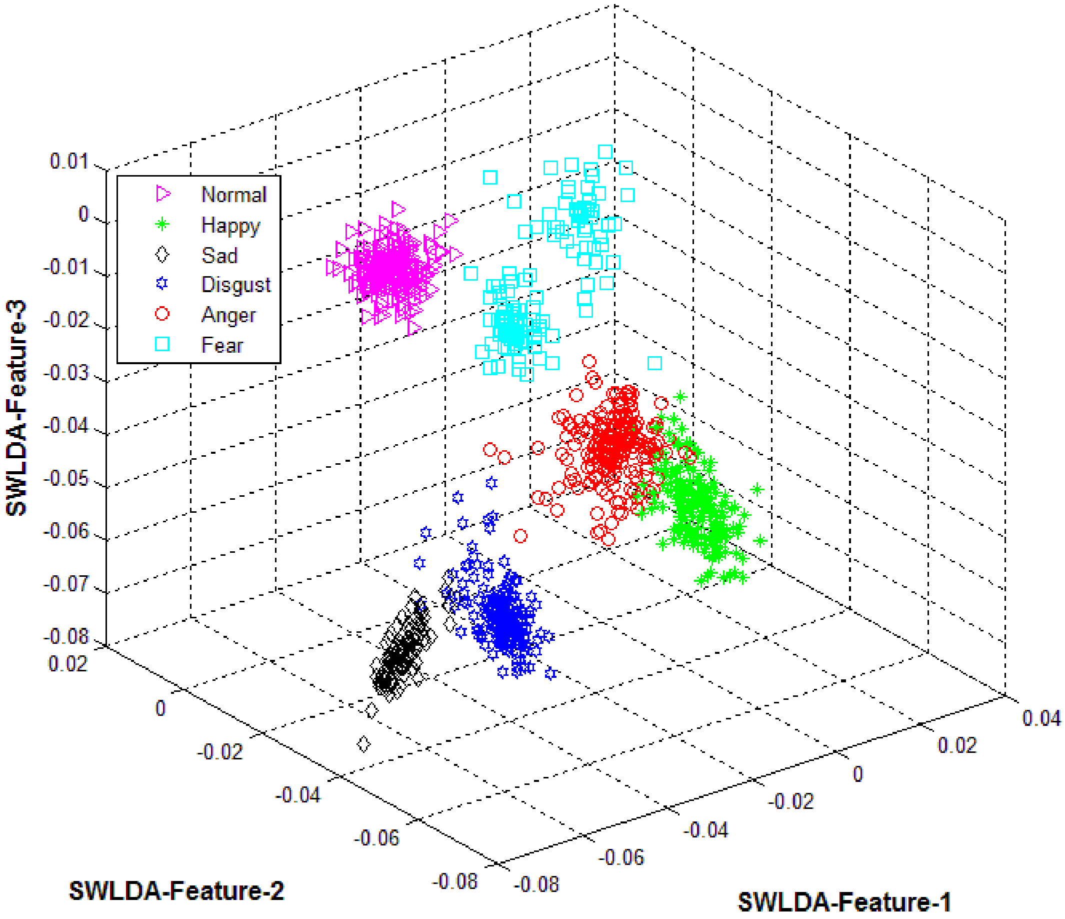


Figure 5.7: 3D feature plots for the six expressions after applying the proposed spontaneous methods on IMFDB dataset.

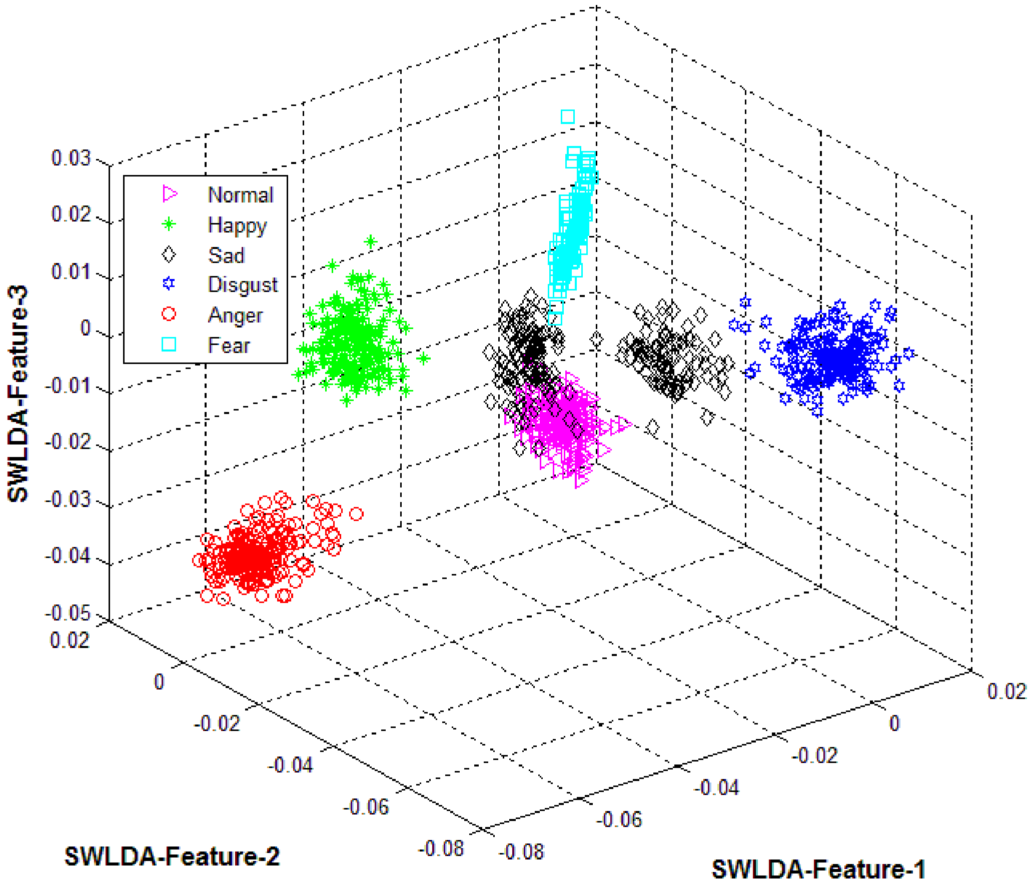


Figure 5.8: 3D feature plots for the six expressions after applying the proposed spontaneous methods on RaFD dataset.

Table 5.26: Confusion matrix of the proposed spontaneous methods on USTC-NVIE spontaneous dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	95	0	1	2	1	1
Sad	1	94	2	1	1	1
Anger	0	1	93	3	1	2
Disgust	1	1	2	93	2	1
Surprise	0	2	0	0	96	2
Fear	0	2	2	1	1	94
Average			94.16			

Table 5.27: Confusion matrix of the proposed spontaneous methods on IMFDB dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	94	1	0	2	2	1
Sad	1	97	0	1	0	1
Anger	0	3	92	2	1	2
Disgust	2	2	1	91	2	2
Surprise	1	2	2	3	90	2
Fear	0	0	3		0	97
Average			93.50			

Table 5.28: Confusion matrix of the proposed spontaneous methods on RaFD dataset of facial expressions (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	92	0	2	3	1	2
Sad	2	91	2	3	2	0
Anger	0	2	94	2	1	1
Disgust	2	2	3	91	0	2
Surprise	0	2	2	2	90	4
Fear	2	1	2	2	1	92
Average			91.67			

It is also clear from Figures 5.6, 5.7, and 5.8, and Tables 5.26, 5.27, and 5.28 that the proposed spontaneous methods consistently achieved a high recognition rate when applied on spontaneous datasets separately.

This means that, unlike existing methods, the proposed spontaneous methods is more robust, i.e., it provided high recognition rate not just for one dataset but all the three spontaneous datasets. The reason is that the proposed feature extraction and recognition method are more robust to the real life scenarios.

5.3.2 Second Experiment: Recognition Rate under the Absence of Each Module

In this experiment, a set of sub-experiments were performed in order to show the importance of sub-components in the proposed spontaneous methods, i.e., wavelet transform with optical flow, SWLDA, and HCRF. For this purpose, nine sub-experiments were performed on the spontaneous datasets using the 10-fold validation rule.

5.3.2.1 Results While Removing the Proposed Feature Extraction Technique

In the first three sub-experiments, ICA (a well-known local feature extraction technique) was utilized with SWLDA and HCRF instead of the proposed feature extraction method (i.e., wavelet transform with optical flow). The overall results for the these sub-experiments on each dataset are shown in Tables 5.29, 5.30, and 5.31, respectively.

Table 5.29: Confusion matrix of the ICA+SWLDA with HCRF using the USTC-NVIE spontaneous dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	93	1	3	1	2	0
Sad	1	90	1	2	2	4
Anger	2	1	89	3	2	3
Disgust	1	2	2	90	3	2
Surprise	2	3	2	0	91	2
Fear	2	2	3	3	1	89
Average			90.33			

For the feature extraction, we utilized symlet wavelet transform with optical flow. So, it can be seen from Tables 5.29, 5.30, and 5.31 that without the proposed feature extraction method, the system did not show better performance. This is because, the proposed idea of noise reduction using symlet wavelet transform relies on the operation of the wavelet coefficients using a filter that takes into account the local regularity of the coefficients in the transform domain. Similarly, the estimation of the threshold is not required for this method. The initial probabilities have been assigned in order to show how noisy the coefficients are [130]. Also, the symlet wavelet has the capability to support the characteristics of orthogonal, biorthogonal, and reverse biorthogonal of gray scale images, thats why it provides better results. Moreover, we measure the statistical dependency of wavelet coefficients for all the facial frames of gray scale. Joint probability of a gray scale frame is computed by collecting geometrically aligned frames of the expression for each wavelet coefficient.

Furthermore, the motion of pixels in contributing parts of the face could help in recognizing expressions. Therefore, once, the noise has been diminished from the expression frames, then the optical flow is employed in order to extract the frequency-based features from the expression frames. High recognition results show that optical flow is capable of extracting prominent features from the enhanced expression frames with the aid of locality in frequency, orientation, and in space as well.

Table 5.30: Confusion matrix of the ICA+SWLDA with HCRF using IMFDB dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	88	4	2	3	2	1
Sad	2	90	3	2	1	2
Anger	3	2	87	1	4	3
Disgust	2	1	2	90	2	3
Surprise	0	3	2	2	91	3
Fear	4	0	3	1	2	90
Average			89.33			

Table 5.31: Confusion matrix of the ICA+SWLDA with HCRF using RaFD dataset of facial expressions, while removing the proposed feature extraction (wavelet transform coupled with optical flow) method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	89	4	5	3	4	5
Sad	6	91	5	7	5	4
Anger	4	6	92	7	7	6
Disgust	6	5	5	89	4	6
Surprise	5	4	4	4	88	5
Fear	4	3	4	5	4	94
Average			90.50			

5.3.2.2 Results While Removing SWLDA Technique

In the next three sub-experiments, wavelet transform with optical flow was coupled with LDA (a well-known discriminant analysis approach) before feeding the features to the proposed HCRF. The results for the these sub-experiments on each dataset are presented in Tables 5.32, 5.33, and 5.34, respectively.

Table 5.32: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using USTC-NVIE spontaneous dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	95	0	1	2	1	1
Sad	2	91	0	2	3	2
Anger	2	0	93	3	0	2
Disgust	1	2	3	91	1	2
Surprise	0	1	1	2	96	0
Fear	3	2	2	0	1	92
Average			93.00			

Similarly, it is also apparent from Tables 5.32, 5.33, and 5.34 that without SWLDA technique, the system was also unable to achieve high classification rate. This is because SWLDA not only provides dimension reduction, it also increases the low between-class variance to increase the

Table 5.33: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using IMFDB dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	91	0	2	3	2	2
Sad	0	94	2	1	1	2
Anger	3	3	90	1	1	2
Disgust	1	1	1	93	2	2
Surprise	3	3	1	2	89	2
Fear	1	2	4	2	0	91
Average			91.33			

Table 5.34: Confusion matrix of the wavelet transform with optical flow+LDA with HCRF using RaFD dataset of facial expressions, while removing SWLDA method (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	90	2	0	4	1	3
Sad	2	93	0	2	2	1
Anger	2	2	89	3	2	2
Disgust	2	2	3	91	2	0
Surprise	1	1	1	1	94	2
Fear	0	1	3	2	2	92
Average			91.50			

class separation before the features are fed to the classifier. The low within class variance and high between class variance are achieved because of the forward and backward regression models in the SWLDA.

5.3.2.3 Results While Removing the Proposed Recognition Model

Finally, in the last sub-experiments, the existing HCRF [2] was used with wavelet transform coupled with optical flow and SWLDA instead of using the proposed HCRF. The experimental results for these sub-experiments on USTC-NVIE spontaneous, IMFDB, and RaDB datasets are shown in Tables 5.35, 5.36, and 5.37, respectively.

Likewise, it is also clear from Tables 5.35, 5.36, and 5.37 that when HCRF was replaced

Table 5.35: Confusion matrix of the wavelet transform with optical flow+SWDA with the existing HCRF [2] model using USTC-NVIE spontaneous dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	88	3	2	2	2	3
Sad	2	91	2	3	1	1
Anger	1	3	90	2	1	3
Disgust	0	3	1	89	4	3
Surprise	0	3	2	2	90	3
Fear	2	0	4	2	3	89
Average			89.50			

Table 5.36: Confusion matrix of the wavelet transform with optical flow+SWDA with the existing HCRF [2] model using IMFDB dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	90	2	4	0	3	1
Sad	1	93	4	2	0	0
Anger	3	2	89	4	0	2
Disgust	1	1	2	95	2	0
Surprise	0	2	3	1	91	3
Fear	3	0	2	3	4	88
Average			91.00			

with existing HCRF [2], the system was also unable to achieve good recognition rate. Thus the proposed HCRF model successfully addresses the limitations of HMM and existing HCRFs, which has widely been used for the sequential FER.

5.3.3 Third Experiment: Recognition Rate of the Proposed Spontaneous Methods Across the Datasets (Robustness)

For this experiment, $n-$fold cross-validation rule based on dataset was performed (in our case n =3). The overall results for this experiment are presented in Tables 5.38, 5.39, and 5.40, respec-

Table 5.37: Confusion matrix of the wavelet transform with optical flow+SWDA with the existing HCRF [2] model using RaFD dataset of facial expressions, while removing the proposed recognition model (HCRF) (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	88	3	4	1	0	4
Sad	2	94	3	0	1	0
Anger	3	2	92	0	3	0
Disgust	3	2	3	87	1	4
Surprise	1	4	0	3	90	2
Fear	0	1	3	3	0	93
Average			90.67			

Table 5.38: Confusion matrix of the proposed Spontaneous FER system training on USTC-NVIE spontaneous dataset and testing on IMFDB and RaFD datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	85	4	2	3	4	2
Sad	3	89	2	3	1	2
Anger	4	3	82	5	2	4
Disgust	2	4	2	83	3	6
Surprise	1	2	4	2	90	1
Fear	1	3	2	2	1	91
Average			86.67			

tively.

It is clear from tables that the proposed spontaneous methods achieved a high recognition rate when it was trained on individual dataset; however, the recognition rate is still low. This might be because the datasets have different facial features and different environment. For instance, the subjects of USTC-NVIE spontaneous dataset performed the expressions in a posed manner, that is, each subject tried to copy or mimic the instructor, so there were little variations from subject-to-subject and in timings. However, the variation in capturing of expression from various angles (placing camera at variant angles) gave us the ability to test the proposed algorithm on the maximum possible alterations/variations in the images. Moreover, USTC-NVIE spontaneous

Table 5.39: Confusion matrix of the proposed spontaneous methods training on IMFDB dataset and testing on USTC-NVIE spontaneous and RaFD datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	79	5	4	6	2	4
Sad	1	90	2	2	3	2
Anger	5	3	80	6	2	4
Disgust	2	4	4	85	3	2
Surprise	5	0	4	1	88	2
Fear	0	2	4	2	3	89
Average			85.17			

Table 5.40: Confusion matrix of the proposed spontaneous methods training on RaFD dataset and testing on USTC-NVIE spontaneous and IMFDB datasets (Unit: %).

Expressions	Happy	Sad	Anger	Disgust	Surprise	Fear
Happy	87	2	3	3	4	1
Sad	3 83	2	4	3	5	
Anger	1	3	85	4	4	3
Disgust	2	2	2	89	4	1
Surprise	4	6	3	2	82	3
Fear	2	2	3	2	1	90
Average			86.00			

datasett images are mostly front-faced, right-sided, and left-sided with up and down orientations. Likewise, the expressions in IMFDB dataset were collected from the movie/drama scenes of professional actors and actresses, where we had no control on expression timings, camera, lighting and background settings. Hence, these expressions are semi-naturalistic expressions collected under dynamic settings. The performance of the methods also degrade when trained on RaFD dataset. This is because the expressions in this dataset are spontaneous expressions collected in natural and dynamic settings. The dataset includes both indoor and outdoor subjects with varying and dynamic backgrounds. In this dataset, different views from different angles with glasses, hair open, wearing hat, and other complex scenarios with obvious actions and things were included. Moreover, the images in this dataset were collected in real life setting such as a variety of back-

Table 5.41: The weighted average classification results of the proposed spontaneous methods with some existing state-of-the-art systems (Unit: %).

Existing Works	[131]	[132]	[33]	[133]	[134]	[135]	[136]	[35]	Proposed Methods
Accuracy Rate	75	81	66	77	82	65	68	76	**93**

grounds, unintentional expressions of the subjects, some variant/orientation angles of the face of the subjects, and lighting variations. These were some factors which may cause misclassification. Nevertheless, the results are very encouraging and this suggests that the proposed spontaneous methods is robust, i.e., the methods not only achieved a high recognition rate on one dataset, but also provided good recognition rates when used across multiple datasets.

5.3.4 Fourth Experiment: Comparison with the Previous Systems on Spontaneous Datasets

In this experiment, the recognition rates for the proposed spontaneous methods was compared against some of the existing FER systems. The overall results of these systems along with the proposed spontaneous methods are summarized in Table 5.41.

It can be seen from Table 5.41 that the proposed spontaneous methods outperformed the existing methods. Thus, the proposed techniques show significant potential in its ability to accurately and robustly recognize human facial expressions in naturalistic scenarios.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

Over the last decade, automatic human FER has become an important research area for many applications. Several factors make FER a challenging research problem, such as varying light conditions in training and test images, the need for automatic and accurate detection of faces before feature extraction, and the high similarity among different expressions resulting in overlaps among feature values of different classes in the feature space. Though several FER systems have been proposed that showed promising results for a certain dataset, their performance was significantly reduced when tested with different datasets. Moreover, one limitation seen among most of these systems is that they were evaluated under controlled settings that are far from real-life scenarios. The reason is that the existing publicly available FER datasets are mostly pose-based and assume a predefined setup. The facial expressions in these datasets are recorded using a fixed camera deployment with a constant background and static ambient settings. In a real-life scenario, FER systems are expected to deal with changing ambient conditions, dynamic background, varying camera angles, different face size, and other human-related variations. Accordingly, in this work, we have proposed two FER systems which solved the limitations of the existing FER systems.

- A hierarchical recognition systems that is capable of providing a high recognition accuracy even when images are captured under different lighting conditions and subjects' facial features vary. In the proposed system, the recognition starts by employing ICA to extract both the global and the local features. Then, the dimensions of the feature space have been reduced by employing linear discriminant analysis (LDA). Finally, we used a hierarchical recognition scheme to overcome the problem of high similarity among different expressions. In this system, the expressions were divided into three categories based on different parts of the face. At the first level, LDA was used with an HMM to recognize the expression category. While, at the second level, the label for an expression within the recognized category is recognized using a separate set of LDA and HMM, trained just for that category. In order to evaluate the performance of the proposed hierarchical recognition system detailed experiments were conducted on two publicly available datasets with respect to different experimental settings. The proposed hierarchical system achieved a weighted average recognition accuracy of 98.7% using three HMMs, one for per category expression across two different datasets (the Cohn-Kanade dataset has 5,400 images, and the JAFFE dataset has 193 images), illustrating the successful use of the hierarchical recognition scheme for automatic FER.

However, the proposed hierarchical recognition system used two level classifications; therefore, took a bit more time and also computational wise much expensive than of the existing system.

- Therefore, in order to maintain the same accuracy using a single level classification with little computational cost, we proposed accurate and robust methods for a single-level system, capable of providing high recognition accuracy even in a single level of classification. In order to achieve high classification rate, accurate feature extraction, feature selection, and recognition methods are required. In the proposed FER system, we have contributed in each module. For the feature extraction, we have proposed a robust feature extraction technique based on the facial movement features is proposed. The technique is based on symlet wavelet transform coupled with optical flow to get the facial movement features. The reason for using the wavelet transform is to diminish the noise before extracting the facial movement features. Even though the proposed feature extraction method extracts good features, there might still be some redundancy among these features. Therefore, for the feature selection, the we used the existing robust method named Stepwise Linear Discriminant Analysis (SWLDA) is applied to the selected feature space. SWLDA selects the most informative features taking the advantage of the forward selection model and removes irrelevant features by taking the advantage of the backward regression model. For expression classification, we have proposed the improved version of the hidden conditional random fields (HCRF) model. This model inherits the advantages of existing HCRF model and completely tackle the limitations of the existing work. In the proposed HCRF model, we explicitly utilized mixture of full covariance Gaussian distribution. The system was validated and tested on four publicly available standard datasets such as Extended Cohn-Kanade CK+, USTC-NVIE, MUG, and MMI datasets. The system achieved weighted average recognition rate (97%) using the four datasets. Moreover, the proposed system was also a robust system, which means that when the system was trained on one dataset and tested on the remaining datasets, the proposed system showed a significant performance.

In these proposed FER systems, we utilized the existing standard datasets of facial expressions. However, most of these datasets were collected under predefined setups. Several factors that effect the accuracy of the FER methodologies include varying light conditions and dynamic variation of the background and the subject positions. Therefore, we have evaluated the proposed methodology on some previous spontaneous datasets such as USTC-NVIE (Spontaneous-based), IMFDB, and RaFD datasets. Based the experimental results, the system showed a significant performance in realistic scenarios.

6.2 Future Directions

Healthcare applications that employ video technologies raise privacy concerns since it can lead to situations where subjects may not know that their private information is being shared and thus become exposed to a threat. Unlike RGB-cameras, depth-cameras only capture the depth information and do not reveal the identification of the subject or other sensitive information, which makes them a superior choice over RGB-cameras. Therefore, in our future work, we will choose the depth-camera over RGB-cameras for the human FER for such domains.

Moreover, in real life scenarios (outdoor environments), any kind of camera might not be applicable. So, smartphone is one of the best candidates for such situations. Therefore, our alternative goal is to propose an accurate and efficient FER system using smartphone for real life scenarios.

Bibliography

[1] M. Z. Uddin, T.-S. Kim, and B. C. Song, "An optical flow feature-based robust facial expression recognition with hmm from video," *International Journal of Innovative Computing, Information and Control*, vol. 9, no. 4, pp. 1409–1421, 2013.

[2] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, vol. 2, no. 1. Citeseer, 2005, pp. 1117–1120.

[3] V. Bettadapura, "Face expression recognition and analysis: the state of the art," *arXiv preprint arXiv:1203.6722*, 2012.

[4] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.

[5] F. Dornaika and B. Raducanu, "Facial expression recognition for hci applications." 2009.

[6] A. A. M. Al-modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, "Facial expression recognition intelligent security system for real time surveillance," 2012.

[7] J. N. Liu, M. Wang, and B. Feng, "ibotguard: an internet-based intelligent robot security system using invariant face recognition against intruder," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 1, pp. 97–105, 2005.

[8] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system–a case study," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 132–137.

[9] X. Wu and J. Zhao, “Curvelet feature extraction for face recognition and facial expression recognition,” in *Natural Computation (ICNC), 2010 Sixth International Conference on*, vol. 3. IEEE, 2010, pp. 1212–1216.

[10] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Measuring facial expressions by computer image analysis,” *Psychophysiology*, vol. 36, no. 2, pp. 253–263, 1999.

[11] S. Moore and R. Bowden, “The effects of pose on facial expression recognition,” in *Proceedings of the British Machine Vision Conference*, 2009, pp. 1–11.

[12] R. Gross, S. Baker, I. Matthews, and T. Kanade, “Face recognition across pose and illumination,” in *Handbook of Face Recognition*. Springer, 2005, pp. 193–216.

[13] Z. Zhu and Q. Ji, “Robust real-time face pose and facial expression recovery,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 681–688.

[14] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Recognizing facial expression: machine learning and application to spontaneous behavior,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 568–573.

[15] Y.-L. Tian, T. Kanade, and J. F. Cohn, “Facial expression analysis,” in *Handbook of face recognition*. Springer, 2005, pp. 247–275.

[16] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[17] C. Shan and R. Braspenning, “Recognizing facial expressions automatically from video,” in *Handbook of ambient intelligence and smart environments*. Springer, 2010, pp. 479–509.

[18] M. H. Siddiqi, S. Lee, Y.-K. Lee, A. M. Khan, and P. T. H. Truc, “Hierarchical recognition scheme for human facial expression recognition systems,” *Sensors*, vol. 13, no. 12, pp. 16 682–16 713, 2013.

[19] B. M. Demaerschalk, M. L. Miley, T.-E. J. Kiernan, B. J. Bobrow, D. A. Corday, K. E. Wellik, M. I. Aguilar, T. J. Ingall, D. W. Dodick, K. Brazdys *et al.*, "Stroke telemedicine," in *Mayo Clinic Proceedings*, vol. 84, no. 1. Elsevier, 2009, pp. 53–64.

[20] "Telemedicine: Extending specialist care to rural areas, newsletter article, cisco," http://www.cisco.com/web/strategy/docs/gov/fedbiz081810healthpresence.pdf, accessed: 2015-04-23.

[21] I. H. Kraai, M. Luttik, R. M. de Jong, T. Jaarsma, and H. Hillege, "Heart failure patients monitored with telemedicine: patient satisfaction, a review of the literature," *Journal of cardiac failure*, vol. 17, no. 8, pp. 684–690, 2011.

[22] W. Q. Bogdan J. Matuszewski and L.-K. Shark, "Facial expression recognition," in *Oxford Handbook of Innovation*, A. Midori, Ed. InTech, 2011.

[23] Z. Zeng, Y. Hu, G. I. Roisman, Z. Wen, Y. Fu, and T. S. Huang, "Audio-visual spontaneous emotion recognition," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 72–90.

[24] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 12, pp. 2067–2083, 2008.

[25] K. H. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[26] H. Lee, Y. S. Choi, S. Lee, and I. Park, "Towards unobtrusive emotion recognition for affective social communication," in *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*. IEEE, 2012, pp. 260–264.

[27] R. B. Knapp, J. Kim, and E. Andre, "Physiological signals and their use in augmenting emotion recognition for human–machine interaction," in *Emotion-oriented systems*. Springer, 2011, pp. 133–159.

[28] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 974–989, 1999.

[29] C. C. Chibelushi and F. Bourel, "Facial expression recognition: A brief tutorial overview," *CVonline: On-Line Compendium of Computer Vision*, vol. 9, 2003.

[30] A. Mehrabian, "Communication without words," *Psychological today*, vol. 2, pp. 53–55, 1968.

[31] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[32] W. E. Rinn, "The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions." *Psychological bulletin*, vol. 95, no. 1, p. 52, 1984.

[33] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI journal*, vol. 32, no. 5, pp. 784–794, 2010.

[34] M. B. Mariappan, M. Suk, and B. Prabhakaran, "Facial expression recognition using dual layer hierarchical svm ensemble classification." in *ISM*, 2012, pp. 104–107.

[35] A. Ramirez Rivera, J. Rojas Castillo, and O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1740–1752, 2013.

[36] Y. Rahulamathavan, R. Phan, J. Chambers, and D. Parish, "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis," vol. 4, no. 1, pp. 83–92, 2013.

[37] T. Ahsan, T. Jabid, U. Chong *et al.*, "Facial expression recognition using local transitional pattern on gabor filtered facial images," *IETE Technical Review*, vol. 30, no. 1, pp. 47–52, 2013.

[38] Y. Pang, Y. Yuan, and X. Li, “Iterative subspace analysis based on feature line distance,” *Image Processing, IEEE Transactions on*, vol. 18, no. 4, pp. 903–907, 2009.

[39] S. R. V. Kittusamy and V. Chakrapani, “Facial expressions recognition using eigenspaces,” *Journal of Computer Science*, vol. 8, no. 10, pp. 1674–1679, 2012.

[40] J. Kalita and K. Das, “Recognition of facial expression using eigenvector based distributed features and euclidean distance based decision making technique,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, no. 2, pp. 196–202, 2013.

[41] Z. Abidin and A. Harjoko, “A neural network based facial expression recognition using fisherface,” *International Journal of Computer Applications*, vol. 59, no. 3, 2012.

[42] V. J. Mistry and M. M. Goyani, “A literature survey on facial expression recognition using global features,” *International Journal of Engineering and Advanced Technology*, vol. 2, no. 4, pp. 653–657, 2013.

[43] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, “Learning spatiotemporal features by using independent component analysis with application to facial expression recognition,” *Neurocomputing*, vol. 2, no. 1, pp. 126–132, 2012.

[44] A. Halder, A. Jati, G. Singh, A. Konar, A. Chakraborty, and R. Janarthanan, “Facial action point based emotion recognition by principal component analysis,” in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011.* Springer, 2012, pp. 721–733.

[45] L. Ma, “Facial expression recognition using 2-d dct of binarized edge images and constructive feedforward neural networks,” in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on.* IEEE, 2008, pp. 4083–4088.

[46] M. Kumbhar, A. Jadhav, and M. Patil, “Facial expression recognition based on image feature,” *International Journal of Computer and Communication Engineering*, vol. 1, no. 2, pp. 117–119, 2012.

[47] S. Joseph, “Ecological and biochemical studies on cyanobacteria of cochin estuary and their application as source of antioxidants,” 2005.

[48] S. Chitra and D. G. Balakrishnan, “A survey of face recognition on feature extraction process of dimensionality reduction techniques,” *Journal of Theoretical and Applied Information Technology*, vol. 36, no. 1, pp. 92–100, 2012.

[49] S. C. Christabel, M. Annalakshmi, M. D. P. Winston *et al.*, “Facial feature extraction based on local color and texture for face recognition using neural network,” *International Journal of Science and Engineering Applications*, vol. 2, no. 4, pp. 78–82, 2013.

[50] L. Zhang and D. Tjondronegoro, “Feature extraction and representation for face recognition,” *Face Recognition, Milos Oraves (Ed.),*, pp. 1–20, 2010.

[51] M. Gargesha and P. Kuchi, “Facial expression recognition using artificial neural networks,” *Artif. Neural Comput. Syst*, pp. 1–6, 2002.

[52] W.-f. Liu, J.-l. Lu, Z.-f. Wang, and H.-j. Song, “An expression space model for facial expression analysis,” in *Image and Signal Processing, 2008. CISP’08. Congress on*, vol. 2. IEEE, 2008, pp. 680–684.

[53] M. Schels and F. Schwenker, “A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors,” in *Pattern Recognition (ICPR), 2010 20th International Conference on.* IEEE, 2010, pp. 4251–4254.

[54] T. Bouwmans, F. El Baf *et al.*, “Modeling of dynamic backgrounds by type-2 fuzzy gaussians mixture models,” *MASAUM Journal of of Basic and Applied Sciences*, vol. 1, no. 2, pp. 265–276, 2009.

[55] M. Z. Uddin, J. Lee, and T.-S. Kim, “An enhanced independent component-based human facial expression recognition from video,” *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 4, pp. 2216–2224, 2009.

[56] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” pp. 282–289, 2001.

[57] X. Wang and K. K. Paliwal, “Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition,” *Pattern recognition*, vol. 36, no. 10, pp. 2429–2439, 2003.

[58] M. T. Mahmood, “Face detection by image discriminating,” *Blekinge Institute of Technology Box*, vol. 520, 2006.

[59] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, “Two-dimensional pca: a new approach to appearance-based face representation and recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.

[60] R. Cendrillon and B. C. Lovell, “Real-time face recognition using eigenfaces,” in *Proceedings-Spie The International Society for Optical Engineering*, no. 1. International Society for Optical Engineering; 1999, 2000, pp. 269–276.

[61] R. Jafri and H. R. Arabnia, “A survey of face recognition techniques,” *Journal of Information Processing Systems*, vol. 5, no. 2, pp. 41–68, 2009.

[62] E. Hjelmaas and B. K. Low, “Face detection: A survey,” *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[63] N. M. Duc and B. Q. Minh, “Your face is not your password face authentication bypassing lenovo–asus–toshiba,” *Black Hat Briefings*, 2009.

[64] C. Kotropoulos and I. Pitas, “Rule-based face detection in frontal views,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 4. IEEE, 1997, pp. 2537–2540.

[65] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[66] P. F. de Carrera, “Face recognition algorithms,” 2010.

[67] S. K. Singh, D. Chauhan, M. Vatsa, and R. Singh, “A robust skin color based face detection algorithm,” *Tamkang Journal of Science and Engineering*, vol. 6, no. 4, pp. 227–234, 2003.

[68] X. Zhang and Y. Gao, “Face recognition across pose: A review,” *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[69] G. Aguilar-Torres, K. Toscano-Medina, G. Sanchez-Perez, M. Nakano-Miyatake, and H. Perez-Meana, “Eigenface-gabor algorithm for feature extraction in face recognition,” *International Journal of Computers*, vol. 3, no. 1, pp. 20–30, 2009.

[70] K. Karande, S. Talbar, and S. Inamdar, “Face recognition using oriented laplacian of gaussian (olog) and independent component analysis (ica),” in *Digital Information and Communication Technology and it’s Applications (DICTAP), 2012 Second International Conference on*. IEEE, 2012, pp. 99–103.

[71] M. Kaur, R. Vashisht, and N. Neeru, “Recognition of facial expressions with principal component analysis and singular value decomposition,” *International Journal of Computer Applications*, vol. 9, no. 12, pp. 36–40, 2010.

[72] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, “Learning spatially localized, parts-based representation,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–207.

[73] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, “Facial expression recognition using radial encoding of local gabor features and classifier synthesis,” *Pattern Recognition*, vol. 45, no. 1, pp. 80–91, 2012.

[74] I. Buciu and I. Pitas, “Application of non-negative and local non negative matrix factorization to facial expression recognition,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1. IEEE, 2004, pp. 288–291.

[75] S. Zhang, X. Zhao, and B. Lei, “Facial expression recognition based on local binary patterns and local fisher discriminant analysis,” *WSEAS Transactions on Signal Processing*, vol. 8, no. 1, pp. 21–31, 2012.

[76] M. H. Siddiqi, F. Farooq, and S. Lee, “A robust feature extraction method for human facial expressions recognition systems,” in *Proceedings of the 27th Conference on Image and Vision Computing New Zealand.* ACM, 2012, pp. 464–468.

[77] M. H. Kabir, T. Jabid, and O. Chae, “A local directional pattern variance (ldpv) based face descriptor for human facial expression recognition,” in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on.* IEEE, 2010, pp. 526–532.

[78] Z. Wang and X. Sun, “Manifold adaptive kernel local fisher discriminant analysis for face recognition,” *Journal of Multimedia*, vol. 7, no. 6, pp. 387–393, 2012.

[79] C. S Patil and A. J Patil, “A review paper on facial detection technique using pixel and color segmentation,” *International Journal of Computer Applications*, vol. 62, no. 1, pp. 21–24, 2013.

[80] C. Boutsidis, M. W. Mahoney, and P. Drineas, “Unsupervised feature selection for principal components analysis,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 61–69.

[81] S. Mika, “Kernel fisher discriminants,” Ph.D. dissertation, Universita tsbibliothek, 2002.

[82] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, “Fisher discriminant analysis with kernels,” in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.* IEEE, 1999, pp. 41–48.

[83] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[84] G.-C. Feng, P. C. Yuen, and D.-Q. Dai, “Human face recognition using pca on wavelet subband,” *Journal of Electronic Imaging*, vol. 9, no. 2, pp. 226–233, 2000.

[85] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Face recognition using lda-based algorithms,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 195–200, 2003.

[86] S. W. Park and M. Savvides, “A multifactor extension of linear discriminant analysis for face recognition under varying pose and illumination,” *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 6, 2010.

[87] W. Zheng, L. Zhao, and C. Zou, “A modified algorithm for generalized discriminant analysis,” *Neural Computation*, vol. 16, no. 6, pp. 1283–1297, 2004.

[88] J. V. Tu, “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes,” *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[89] H.-H. Tsai, Y.-S. Lai, and Y.-C. Zhang, “Using svm to design facial expression recognition for shape and texture features,” in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, vol. 5. IEEE, 2010, pp. 2697–2704.

[90] “Support vector machines (svms),” http://scikit-learn.org/stable/modules/svm.html, 2011, (Last visited in Sunday 15 December 2013).

[91] S. Kumar and M. Hebert, “Discriminative fields for modeling spatial dependencies in natural images.” NIPS, 2003.

[92] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, “Hidden conditional random fields,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, pp. 1848–1852, 2007.

[93] M. Mahajan, A. Gunawardana, and A. Acero, “Training algorithms for hidden conditional random fields,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–273–I–276.

[94] S. Reiter, B. Schuller, and G. Rigoll, “Hidden conditional random fields for meeting segmentation,” in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 639–642.

[95] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[96] R. D. Howe and Y. Matsuoka, “Robotics for surgery,” *Annual Review of Biomedical Engineering*, vol. 1, no. 1, pp. 211–240, 1999.

[97] A. A. M. Al-modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, “Facial expression recognition intelligent security system for real time surveillance,” in *Proceedings of the International Conference on Computer Graphics and Virtual Reality (CGVR)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012, p. 1.

[98] P. W. Yu, “Security technology–facial recognition,” 2008.

[99] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, “Emotion recognition from physiological signals using wireless sensors for presence technologies,” *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, 2004.

[100] P. C. Petrantonakis and L. J. Hadjileontiadis, “Emotion recognition from eeg using higher order crossings,” *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 186–197, 2010.

[101] K. W. Bowyer, K. Chang, and P. Flynn, “A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition,” *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[102] M. Hossny, D. Filippidis, W. Abdelrahman, H. Zhou, M. Fielding, J. Mullins, L. Wei, D. Creighton, V. Puri, and S. Nahavandi, “Low cost multimodal facial recognition via kinect sensors,” in *Proceedings of the land warfare conference (LWC): potent land force for a joint maritime strategy. Commonwealth of Australia*, 2012, pp. 77–86.

[103] “The google image search,” http://searchengineland.com/up-close-with-google-search-by-image-82313, accessed: 2014-05-26.

[104] "Solve puzzle," https://support.google.com/websearch/answer/166331, accessed: 2014-05-26.

[105] T. F. Chan and L. A. Vese, "Active contours without edges," *Image Processing, IEEE Transactions on*, vol. 10, no. 2, pp. 266–277, 2001.

[106] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.

[107] H. Ibrahim and N. S. P. Kong, "Image sharpening using sub-regions histogram equalization," *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 2, pp. 891–895, 2009.

[108] P. Shanmugavadivu and K. Balasubramanian, "Image inversion and bi level histogram equalization for contrast enhancement," *International Journal of Computer Applications (0975-8887) Volume*, 2010.

[109] H. Yoon, Y. Han, and H. Hahn, "Image contrast enhancement based sub-histogram equalization technique without over-equalization noise," *World Academy of Science, Engineering and Technology*, vol. 3.

[110] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[111] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. Wiley-interscience, 2001, vol. 26.

[112] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[113] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[114] L. E. Baum, "An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[115] F. Samaria, “Face recognition using hidden markov models,” Ph.D. dissertation, University of Cambridge, 1994.

[116] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bulthoff, “The contribution of different facial regions to the recognition of conversational expressions,” *Journal of vision*, vol. 8, no. 8, p. 1, 2008.

[117] K. Kaulard, D. W. Cunningham, H. H. Bulthoff, and C. Wallraven, “The mpi facial expression databasea validated database of emotional and conversational facial expressions,” *PloS one*, vol. 7, no. 3, p. e32321, 2012.

[118] K. L. Schmidt and J. F. Cohn, “Human facial expressions as adaptations: Evolutionary questions in facial expression research,” *American journal of physical anthropology*, vol. 116, no. S33, pp. 3–24, 2001.

[119] L. He, W. G. Wee, S. Zheng, and L. Wang, “A level set model without initial contour,” in *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009, pp. 1–6.

[120] J. Turunen *et al.*, “A wavelet-based method for estimating damping in power systems,” 2011.

[121] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques,” *International journal of computer vision*, vol. 12, no. 1, pp. 43–77, 1994.

[122] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “Toward enhanced p300 speller performance,” *Journal of neuroscience methods*, vol. 167, no. 1, pp. 15–21, 2008.

[123] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *Communications, IEEE Transactions on*, vol. 28, no. 1, pp. 84–95, 1980.

[124] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, “A natural visible and infrared facial expression database for expression recognition and emotion inference,” *Multimedia, IEEE Transactions on*, vol. 12, no. 7, pp. 682–691, 2010.

[125] P. B. J. G. M. K. R. V. V. H. J. C. K. R. R. R. V. K. Shankar Setty, Moula Husain and C. V. Jawahar, “Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations,” in *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Dec 2013.

[126] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, “Presentation and validation of the radboud faces database,” *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[127] D. Ghimire and J. Lee, “Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines,” *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[128] P. Shen, S. Wang, and Z. Liu, “Facial expression recognition from infrared thermal videos,” in *Intelligent Autonomous Systems 12*. Springer, 2013, pp. 323–333.

[129] S. Mohseni, H. M. Kordy, and R. Ahmadi, “Facial expression recognition using dct features and neural network based decision tree,” in *ELMAR, 2013 55th International Symposium*. IEEE, 2013, pp. 361–364.

[130] D. S. Chandra, “Image enhancement and noise reduction using wavelet transform,” in *Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on*, vol. 2. IEEE, 1997, pp. 989–992.

[131] R. Kapoor and R. Gupta, “Morphological mapping for non-linear dimensionality reduction,” *IET Computer Vision*, vol. 9, no. 2, pp. 226–233, 2014.

[132] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, “Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,” *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1386–1398, 2015.

[133] M. H. Siddiqi, R. Ali, M. Idris, A. M. Khan, E. S. Kim, M. C. Whang, and S. Lee, “Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection,” *Multimedia Tools and Applications*, pp. 1–25, 2014.

[134] Q. Jia, X. Gao, H. Guo, Z. Luo, and Y. Wang, “Multi-layer sparse representation for weighted lbp-patches based facial expression recognition,” *Sensors*, vol. 15, no. 3, pp. 6719–6739, 2015.

[135] X. Zhao, X. Shi, and S. Zhang, “Facial expression recognition via deep learning,” *IETE Technical Review*, no. ahead-of-print, pp. 1–9, 2015.

[136] S. Happy and A. Routray, “Robust facial expression classification using shape and appearance features,” in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015, pp. 1–5.

Appendix A

List of Publications

International Journal Papers:

[1] **Muhammad Hameed Siddiqi**, Rahman Ali, Adil Mehmood Khan, Sungyoung Lee, ”Spontaneous-based Facial Expression Recognition System using Real-time YouTube Dataset,” IEEE Transactions on Pattern Analysis and Machine Intelligence (SCI, IF: 5.781), 2015 (Under Review).

[2] **Muhammad Hameed Siddiqi**, Oresti Banos, Rahman Ali, Maqbool Ali, Muhammad Idris, Mohamed Elsayed Abdelrahman Eldib, Adil Mehmood Khan, Sungyoung Lee, ”Evaluating Facial Expression Recognition Methodologies in Real-World Situations by using YouTube-based Datasets,” Multimedia Systems (SCI, IF: 0.619), 2015 (Under Review).

[3] **Muhammad Hameed Siddiqi**, Oresti Banos, Irshad Ahmad, Asad Masood Khattak, Junaid Ahsenali Chaudhry, Adil Mehmood Khan, Sungyoung Lee, ”Evaluating Real-life Performance of the State-of-the-art in Facial Expression Recognition using a Novel Dataset,” Multimedia Tools and Applications (SCIE, IF:1.346), 2015 (Under Review).

[4] **Muhammad Hameed Siddiqi**, Rahman Ali, Adil Mehmood Khan, Young-Tack Park, Sungyoung Lee, ”Human Facial Expression Recognition using Stepwise Linear Discriminant Analysis and Hidden Conditional Random Fields,” IEEE Transactions on Image Processing (SCI, IF 3.625), Vol. 24, Issue 4, pp. 1386–1398, 2015.

[5] **Muhammad Hameed Siddiqi**, Rahman Ali, Muhammad Idris, Adil Mehmood Khan, Eun Soo Kim, Min Cheol Whang, Sungyoung Lee, ”Facial Expression Recognition using Curvelet Feature Extraction and Normalized Mutual Information Feature Selection,” Multimedia Tools and Applications (SCIE, IF:1.346) (Accepted), 2014.

[6] **Muhammad Hameed Siddiqi**, Rahman Ali, Adil Mehmood Khan, Eun Soo Kim, Min Cheol Whang, Gerard Junghyun Kim, Sungyoung Lee, ”Facial Expression Recognition using Active Contour-based Face Detection, Facial Movement-based Feature Extraction, and Non-Linear Feature Selection,” Multimedia Systems (SCI, IF: 0.619),, Vol. 21, Issue 6, pp. 541–555, 2014.

[7] **Muhammad Hameed Siddiqi**, Rahman Ali, Abdul Sattar, Adil Mehmood Khan, Sungyoung Lee, ”Depth Camera-based Facial Expression Recognition System using Multilayer Scheme,” IETE Technical Review (SCIE, IF: 0.888), Vol. 31, Issue 4, pp. 277–286, 2014.

[8] **Muhammad Hameed Siddiqi**, Rahman Ali, Md. Sohel Rana, Een-Kee Hong, Eun Soo Kim, Sungyoung Lee, ”Video-based Human Activity Recognition using Multilevel Wavelet Decomposition and Stepwise Linear Discriminant Analysis,” Sensors (SCIE, IF: 2.245), Vol. 14, Issue 4, pp. 6370–6392, 2014.

[9] **Muhammad Hameed Siddiqi**, Sungyoung Lee, Young-Koo Lee, Adil Mehmood Khan, Phan Tran Ho Truc, ”Hierarchical Recognition Scheme for Human Facial Expression Recognition System,” Sensors (SCIE, IF: 2.245), Vol. 13, Issue 12, pp. 16682–16713, 2013.

[10] Rahman Ali, Muhammad Afzal, Maqbool Hussain, Maqbool Ali, **Muhammad Hameed Siddiqi**, Byeong Ho Kang, Sungyoung Lee, ”Multimodal Hybrid Reasoning Methodology for Personalized Wellbeing Services,” Computers in Biology and Medicine (SCI, IF: 1.240), 2015 (Minor Revisions).

[11] Muhammad Idris, Shujaat Hussain, Waseem Ahmad, **Muhammad Hameed Siddiqi**, Maqbool Ali, Sungyoung Lee, ”MRPack: Multi-algorithm Execution using Compute-intensive Approach in MapReduce,” Journal of PLOS ONE (SCIE, IF: 3.7), (Accepted), 2015.

[12] Muhammad Idris, Shujaat Hussain, **Muhammad Hameed Siddiqi**, Maqbool Ali, Byeong Ho Kang, Sungyoung Lee, ”Context-aware Scheduling in MapReduce: A Compact Review,” Journal of Concurrency and Computation: Practice and Exercise (SCIE, IF: 0.784), (Accepted) 2015.

[13] Rahman Ali, Jamil Hussain, **Muhammad Hameed Siddiqi**, Maqbool Hussain, Sungyoung Lee, ”H2RM: A Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus,” Sensors (SCIE, IF: 2.245), Vol. 15, Issue 7, pp. 15921–15951, 2015.

[14] Rahman Ali, **Muhammad Hameed Siddiqi**, Muhammad Idris, Shujaat Hussain, Eui-Nam Huh, Byong Ho Kang, Sungyoung Lee, ”GUDM: Automatic Generation of Unified Datasets for Learning and Reasoning in Healthcare,” Sensors (SCIE, IF: 2.245), Vol. 15, Issue 7, pp. 15772–15798, 2015.

[15] Rahman Ali, **Muhammad Hameed Siddiqi**, Sungyoung Lee, ”Rough Set-based Approaches for Discretization: A Compact Review”, Artificial Intelligence Review,” (SCI, IF: 2.111) (Accepted), 2014.

[16] Maqbool Hussain, Asad Masood Khattak, Wajahat Ali Khan, Iram Fatima, Muhammad Bilal Amin, Zeeshan Pervez, Rabia Batool, Muhammad Amir Saleem, Muhammad Afzal, Muhammad Fahim, **Muhammad Hameed Siddiqi**, Sungyoung Lee, and Khalid Latif, ”Cloud-based Smart CDSS for Chronic Diseases”, Journal of Health and Technology, Vol. 3, Issue. 2, pp. 153-175, 2013.

International Conference Papers:

[17] **Muhammad Hameed Siddiqi**, Rahman Ali, Byeong Ho Kang, Sungyoung Lee, ”A New Feature Extraction Technique for Human Facial Expression Recognition Systems using Depth Camera”. Proc. Of the 6th International Work-conference on Ambient Assisted Living (IWAAL’14), pp. 131–138, UK, 2014.

[18] **Muhammad Hameed Siddiqi**, Rahman Ali, Ibrahiem M. M. El Emary, Sungyoung Lee, ”An Unsupervised and Robust Technique for Human Face Detection and Extraction”, Proc. of IEEE International Conference on Information Science, Electronics and Electrical Engineering (ISEEE14), pp. 1756–1760, Sapporo City, Hokkaido, Japan, 2014.

[19] **Muhammad Hameed Siddiqi**, Sungyoung Lee, ”Human Facial Expression Recognition using Wavelet Transform and Hidden Markov Model”. Proc. Of the 5th International Work-

conference on Ambient Assisted Living (IWAAL'13), pp. 112–119, UK, 2013.

[20] **Muhammad Hameed Siddiqi**, Adil Mehmood Khan, Tae Choong Chung, Sungyoung Lee, ”A Precise Recognition Model for Human Facial Expressions Recognition Systems”. Proc. of the 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'13), Regina, Saskatchewan, Canada, 2013.

[21] **Muhammad Hameed Siddiqi**, Faisal Farooq, Sungyoung Lee, ”A Robust Feature Extraction Method for Human Facial Expressions Recognition Systems”. Proc. of the 27th Image and Vision Computing New Zealand (IVCNZ'12), pp. 464–468, Dunedin, New Zealand, 2012.

[22] Rahman Ali, **Muhammad Hameed Siddiqi**, Taqdir Ali, Sungyoung Lee, ”An Integrated Data Model for Healthcare Applications”, 9th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2015), Puerto Varas, Patagonia, Chile, December 1–4, 2015.

[23] Oresti Banos, Jaehun Bang, Taeho Hur, **Muhammad Hameed Siddiqi**, Huynh-The Thien, Le-Ba Vui, Wajahat Ali Khan, Taqdir Ali, Claudia Villalonga, Sungyoung Lee, ”Mining Human Behavior for Health Promotion”, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, August 25–29, 2015.

[24] Banos, O., Bilal Amin, M., Ali Khan, W., Afzel, M., Ahmad, M., Ali, M., Ali, T., Ali, R., Bilal, M., Han, M., Hussain, J., Hussain, M., Hussain, S., Hur, T. H., Bang, J. H., Huynh-The, T., Idris, M., Kang, D. W., Park, S. B., **Siddiqi, M.**, Vui, L. B., Fahim, M., Khattak, A. M., Kang, B. H. and Lee, S, ”An Innovative Platform for Person-Centric Health and Wellness Support”, Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015), pp. 131–140, Granada, Spain, April 15–17, 2015.

[25] Rahman Ali, **Muhammad Hameed Siddiqi**, Sungyoung Lee, ”KARE: A Hybrid Reasoning Approach for Promoting Active Lifestyle”, ACM-IMCOM (ICUIMC) 2015, Jan 8-10, 2015.

[26] Rahman Ali, **Muhammad Hameed Siddiqi**, Muhammad Idris, Byeong Ho Kang, and Sungyoung Lee, ”The Prediction of Diabetes Mellitus Based on Boosting Ensemble Mod-

eling", 8th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2014), Belfast, United Kingdom, December 2–5, 2014.