T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

A NEW APPROACH FOR FACIAL EXPRESSION RECOGNITION WITH AN ADAPTIVE CLASSIFICATION

Abubakar ASHIR

PhD THESIS

Electrical-Electronics Engineering

August-2018
KONYA
All rights reserved

Thesis Acceptance and Approval

Submitted by ABUBAKAR MUHAMMAD ASHIR in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical-Electronics Engineering Department, Selçuk University** by:

Examining Committee Members:	**Signatures**
Başkan Assoc. Prof. Dr. Murat CEYLAN	.....
Danışman Assist. Prof. Dr. Bayram AKDEMİR	.....
Üye Assist. Prof. Dr. Ömer Kaan BAYKAN	.....
Üye Assist. Prof. Dr. Sabri ALTUNKAYA	.....
Üye Assist. Prof. Dr. Alaa ELEYAN	.....

I approve the results above.

Prof. Dr. Mustafa YILMAZ
FBE Müdürü

Date of Approval: 26.07.2018

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Abubakar Muhammad **ASHIR**

Signature:

Date : 19-09-2018

ABSTRACT

A NEW APPROACH FOR FACIAL EXPRESSION RECOGNITION WITH AN ADAPTIVE CLASSIFICATION

Abubakar Muhammad ASHIR

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF DOCTOR OF PHILOSOPHY IN ELECTRIC-ELECTRONICS ENGINEERING

Thesis Advisor: Assist. Prof. Dr. Bayram AKDEMIR

2018, 103 Pages

Examining Committee Members
Assist. Prof. Dr. Bayram AKDEMİR
Assoc. Prof. Dr. Murat CEYLAN
Assist. Prof. Dr. Ömer Kaan BAYKAN
Assist. Prof. Dr. Sabri ALTUNKAYA
Assist. Prof. Dr. Alaa ELEYAN

In recent time Facial Expression Recognition (FER) applications are increasingly becoming more relevant than has been witnessed before. Such applications are visible in many fields such as psychology, medical diagnosis, security, human-machine interaction and entertainments. However, Emotions recognitions presents enormously difficulties and challenges which are tricky and diverse in their various forms. This is because apart from the conventional challenges of a face recognition, FER is affected by the human race, culture and variation in intensity of the same emotion which may present a big challenge.

In this thesis, new holistic approaches to FER have been proposed for improved accuracy and adaptability. In the research two major stages (features extraction and classification) of FER have been identified for improvement. In the realm of feature extraction, a number of new approaches have been proposed one of which got its inspiration from the multiresolution image processing and compressive sensing theory. The multiresolution processing utilizes image Pyramid Processing where each image sample is decomposed into a number of predetermined image pyramids at different sizes and resolutions. Robust features are extracted from all pyramid levels and then get concatenated to extract the pyramid feature vector. compressive sensing theory is used to further reinforce the feature vectors and significantly trim down its dimension. Another variant of the feature extractor proposed is monogenic-Local Binary Pattern (mono-LBP) which operates on Local Gabor Pattern (LGP). Basically, mono-LBP extracts its features from the LGP of inputs data at different orientation and scales. These LGP features are improved by being decomposed into tri-monogenic LBP features which are subsequently recombined to make up the final feature vector. This feature vector is normalized using different normalization schemes.

For classification, two state-of-the art classifiers; Single Branch Decision Tree (SBDT) and Dynamic Cascaded Classifier (DCC) were proposed for adaptivity and overall performance improvement. SBDT is a multi-level classifier based on single-branch decision tree. SBDT essentially uses binary support vector machines during training. To determine a test data class, the outputs of the trained binary SVMs are evaluated through the nodes of the tree from the root to its apex. On the other hand, DCC presents an entirely different new frontier. DCC uses SVM as well is its framework with an automatic kernel parameter optimization with Radial Basis Function. As a prerequisite to its use, features vectors are projected to a Gaussian space using compressive sensing techniques. The Receiver operating characteristics of the SVM classifier provides the basis upon which optimum kernel parameters are automatically determined. Promising and impressive results were recorded from the experiments. The

experimental results obtained from the popular facial expression databases which deployed various state-of-the-art cross-validation techniques, indicated that the proposed method outperforms most of its counterparts under the same experimental set-up and databases. The results further confirmed the validity and advantages of the proposed approach over other approaches currently used in the literature**.**

Keywords: Facial expression recognition, Compressive sensing, Image pyramid, Sigle-Branch decision tree, Dynamic Cascaded Classifier.

ÖZET

UYARLANABİLİR SINIFLANDIRMA iLE YÜZ İFADE TANIMA İÇİN YENİ BİR YAKLAŞIM

Abubakar Muhammad ASHIR

ŞELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ

Danışman: Dr. Öğr. Üye. Bayram AKDEMIR

2018, 103 sayfa

Jüri
Dr. Öğr. Üye. Bayram AKDEMİR
Doç. Dr. Murat CEYLAN
Dr. Öğr. Üye. Ömer Kaan BAYKAN
Dr. Öğr. Üye. Sabri ALTUNKAYA
Dr. Öğr. Üye. Alaa ELEYAN

Son zamanlarda Yüz İfade Tanıma (FER) uygulamaları, daha önce olduğundan giderek daha ilgi çeken hale gelmektedir. Bu tür uygulamalar psikoloji, tıbbi teşhis, güvenlik, insan-makine etkileşimi ve eğlence gibi birçok alanda görülmektedir. Bununla birlikte, duygu tanımaları, çeşitli şekilde zorluk ve bazen de çok hassas olmaktadır. Bunun nedeni ise, yüz tanıma konusunun sıradan zorluklarından ayrı olarak, FER'in, insan ırkının, kültürünün ve aynı duyguların yoğunluğunun büyük bir zorluğa yol açabilecek çeşitliliğinden etkilenmesidir.

Bu tez çalışmasında, daha iyileştirilmiş hassasiyet ve uyarlana bilirlik için FER'e yeni bütüncü yaklaşımlar önerilmiştir. Araştırmada, FER'in iki ana aşaması olan (öznitelik çıkarımı ve sınıflandırma özellikleri) iyileştirme için belirlenmiştir. Öznitelik çıkarma alanında, yeni yaklaşımlar önerilmiş ve bunlardan biri de çoklu-çözünürlük ayrışımı ve sıkıştırılmalı algılama teorisinden ilham alınmıştır. Çoklu çözünürlük ayrışımı, her bir görüntü örneğinin, farklı boyutlarda ve çözünürlüklerde önceden belirlenmiş bir dizi görüntü piramidine ayrıştırıldığı görüntü Piramit İşlemini kullanır. Sağlam özellikler, tüm piramit seviyelerinden çıkarılır ve daha sonra piramit özellik vektörünü çıkarmak için birleştirilir. Sıkıştırılmalı algılama teorisi, özellik vektörlerini daha da güçlendirmek ve boyutunu önemli ölçüde azaltmak için kullanılır. Önerilen özellik çıkarıcının bir başka şekli, Yerel Gabor Örüntü (LGP) üzerinde çalışan monogenik-Yerel İkili Örüntüdür (mono-LBP). Esas olarak, monogenic- Yerel İkili Örüntüsü (mono-LBP), farklı yönlendirme ve ölçeklerdeki giriş verilerinin Yerel Gabor Örüntüsü'nden (LGP) özelliklerini çıkarır. Bu LGP özellikleri, son özellik vektörünü oluşturmak için daha sonra yeniden birleştirilen üçlü-monojenik LBP özelliklerine ayrıştırılarak geliştirilmektedir. Bu özellik vektörü farklı normalizasyon şemaları kullanılarak normalleştirilmiştir.

Sınıflandırma için, iki son teknoloji ürünü sınıflandırıcı; Uyarlana bilirlik ve genel performans iyileştirmesi için Tek Dallı Karar Ağacı (SBDT) ve Dinamik Basamaklı Sınıflandırıcı (DCC) önerilmiştir. SBDT, tek dallı karar ağacına dayanan çok düzyeli bir sınıflandırıcıdır. SBDT, talimat sırasında esas olarak ikili destek vektörü makineleri kullanmaktadır. Bir deneme veri sınıfını belirlemek için, talımatlandırılmış ikili SVM'lerin çıktıları, ağacın köklerinden tepesine kadar olan düğümleri aracılığıyla değerlendirilir. Öte yandan, Dinamik Basamaklı Sınıflandırıcı (DCC) tamamen farklı bir yeni sınır sunuyor. DCC, SVM'yi de kullanır ve Dairesel Temel Fonksiyonu ile otomatik bir çekirdek parametresi optimizasyonu ile çerçevesini oluşturur. Kullanımının bir ön koşulu olarak, özellikler vektörleri, sıkıştırma algılama teknikleri kullanılarak bir Gauss uzayına yansıtılır. SVM sınıflandırıcısının özelliklerini çalıştıran alıcı, optimum çekirdek parametrelerinin otomatik olarak belirlendiği temeli sağlar.

Umut verici ve etkileyici sonuçlar deneylerden elde edilmektedir. Yaygın yüz ifadesi veri tabanlarından elde edilen deneysel sonuçlar, son teknoloji ürünü çapraz doğrulama tekniklerinin kullanıldığı alanda, önerilen yöntemin, benzer deney düzeneği ve veri tabanları altında muadillerinin çoğunu geride bıraktığını göstermiştir. Sonuçlar ayrıca, önerilen yaklaşımın literatürde halihazırda kullanılan diğer yaklaşımlar üzerindeki geçerliliğini ve avantajlarını doğrulamıştır

Anahtar Kelimeler: Yüz İfade Tanıma, Sıkıştırılmalı algılama, Görüntü piramidi, Tek Dallı Karar Ağacı, Dinamik Basamaklı Sınıflandırıcı.

DEDICATION

To my family

ACKNOWLEGMENT

I wish to express my deepest gratitude to my current supervisor Assist. Prof. Dr. Bayram AKDEMIR and Asst. Prof. Dr. Alaa ELEYAN for their guidance, advice, criticisms, encouragement and insight throughout the research period.

My appreciations and heartfelt gratitude go out to all who particularly participated or contributed in one way or the other towards my success in this life, and in the course of this thesis, in particular. Worthy to mention are my parents, immediate family, friends, relatives (both close and distance) and associates. Your advices, goodwill messages and prayers would be remembered for always. I love you all.

I am particularly indebted in all sincerity to an icon, father and a role model, Sen. Rabiu Musa Kwankwaso for his foresight in piloting the scholarship schemes which provided us with opportunities and laid down the foundation upon which I attained this remarkable feat. You will forever continue to endear in the bosom of our hearts for your unparalleled contributions towards humanity and pursuit of dignity, equality and prosperity to all.

Abubakar Muhammad ASHIR
KONYA-2018

TABLE OF CONTENTS

ABSTRACT ..... **iv**

DEDICATION ..... **viii**

ACKNOWLEGMENT ..... **ix**

TABLE OF CONTENTS ..... **x**

LIST OF FIGURES ..... **xii**

LIST OF TABLES ..... **xiii**

LIST OF SYMBOLS/ABBREVIATIONS ..... **xiv**

1. INTRODUCTION **1**

1.1. Overview 1
1.2. Dearth and Resurgence of FER 4
1.3. Objectives 6
1.4. Methodology 7

2. LITERATURE REVIEW **9**

3. THEORETICAL BACKGROUND **20**

3.1. Anatomy of Facial Expression Muscles 20
3.1.1. Orbital Muscles Group 20
3.1.2. Nasal Muscles Group 21
3.1.3. Oral Muscles Group 22
3.1.4. Facial Action Coding Systems (FACS) 24
3.2. Holistic-based Approach to FER 29
3.3. Feature Extraction Algorithms 31
3.3.1. Local Binary Pattern (LBP) 32
3.3.2. Gabor Wavelets Transforms 37
3.3.3. Image Pyramid Processing 38
3.3.4. Compressive sensing Theory 40
3.4. Classification Algorithms 41
3.4.1. K-Nearest Neighbor 42
3.4.2. Support Vector Machine (SVM) 43
3.4.3. Artificial Neural Networks (ANN) 50
3.5. Normalization Schemes 51
3.5.1. Z-Score Normalization 51
3.5.2. MIN-MAX Normalization 52
3.5.3. Median-MAD Normalization 52
3.5.4. Tangent-hyperbolic (Tanh) Normalization 53

4. CONCEPTUAL DESIGN AND ALGORITHMIC FORMULATIONS **54**

4.1. Image Pyramid and Single Branch Decision Tree 54
4.1.1 Feature Extraction using Image Pyramid and CS 54
4.1.2. Single-branch Decision Tree for multi-class SVM 57
4.2. Arithmetic Mean Difference with Dynamic Cascaded Classification 59
4.2.1. Feature Extraction with Arithmetic Mean Difference 59
4.2.2 Dynamic Cascaded Classifier 60
4.3. Monogenic Local Gabor Binary Patterns (mono-LGBP^{riu2}) 64
4.3.1 Feature Extraction 65
4.3.2 Training and Classification 67

5. EXPERIMENTAL RESULTS AND DISCUSSION 68

5.1. Results for Image Pyramid and Single Branch Decision Tree 68
5.1.1. Leave-One-Pose-Out cross-validation on JAFFE and TFEID. 69
5.1.2. n-fold cross-validation on CK 69
5.1.3. Discussion 72
5.1.4. Comparison 73
5.2. Results for Arithmetic Mean Difference with Dynamic Cascaded Classification 75
5.2.1. Experiments on JAFFE Database. 76
5.2.2. Experiments on CK Database. 77
5.2.3. Experiments on MMI Database 78
5.2.4. Expression Specific Performance Results 79
5.2.5. Results Comparison. 81
5.2.6. Discussions 82
5.2.7. DCC Dynamic Responses 83
5.3. Results for Monogenic Local Gabor Binary Patterns (mono-LGBP^{riu2}) 88
5.3.1. Results Discussion 89

6. CONCLUSION AND FUTURE WORK 92

6.1. CONCLUSION 92
6.2. FUTURE WORK 93

REFERENCES 94

APPENDIX 100

CARRICULUM VITAE 102

LIST OF FIGURES

Figure 3. 1 The main oral muscles of facial expression. (Oliver, 2018) 24
Figure 3. 2 Sample of FACS coding of a Fear Expression (Oliver, 2018) 27
Figure 3. 3 AUs Combination for Emotions in Lower Face 28
Figure 3. 4 AUs Combination for Emotions in Upper Face 29
Figure 3. 5 The seven universal Facial expression of emotions 31
Figure 3. 6 Appearance-based approach to FER 32
Figure 3. 8 LXP Operator 37
Figure 3. 9 Pyramid decomposition schematic 39
Figure 3. 10 four-level Pyramids decomposition using Gaussian Filter 40
Figure 3. 11 K-Nearest Neighbor classifier 43
Figure 3. 12 Decision Directed Acyclic Graph for 4 claçsses problem 47
Figure 4. 1 . Proposed Feature Extraction Flowchart 57
Figure 4. 2 Proposed SBDT. 59
Figure 4. 3 Proposed DCC schematic diagram 61
Figure 4. 4 Proposed approach Flowchart using Monogenic 67

Figure 5. 1.Cross section of JAFFE Database. 76
Figure 5. 2 Cross section of preprocessed CK Database. 78
Figure 5. 3 DCC ROC Responses for PI LOPO (JAFFE) feature Length=100 84
Figure 5. 4 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 84
Figure 5. 5 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600 85
Figure 5. 6 DCC ROC Responses for PD LOPO (JAFFE) feature Length=100 85
Figure 5. 7 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 86
Figure 5. 8 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600 86
Figure 5. 9 DCC ROC Responses for PI LOPO (CK) feature Length=100 87
Figure 5. 10 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 87
Figure 5. 11 DCC ROC Responses for PI LOPO (CK) feature Length=600 88
Figure 5. 12 Performance comparison in JAFFE with varying Gabor filter scale, v. 90

LIST OF TABLES

Table 3. 1 Facial Action Coding Systems 25
Table 3. 2 Emotions and associated Aus 26
Table 3. 3 Binary classifiers combination 48

Figure 4. 1 . Proposed Feature Extraction Flowchart 57
Figure 4. 2 Proposed SBDT. 59
Figure 4. 3 Proposed DCC schematic diagram 61
Figure 4. 4 Proposed approach Flowchart using Monogenic 67

Figure 5. 1.Cross section of JAFFE Database. 76
Figure 5. 2 Cross section of preprocessed CK Database. 78
Figure 5. 3 DCC ROC Responses for PI LOPO (JAFFE) feature Length=100 84
Figure 5. 4 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 84
Figure 5. 5 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600 85
Figure 5. 6 DCC ROC Responses for PD LOPO (JAFFE) feature Length=100 85
Figure 5. 7 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 86
Figure 5. 8 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600 86
Figure 5. 9 DCC ROC Responses for PI LOPO (CK) feature Length=100 87
Figure 5. 10 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300 87
Figure 5. 11 DCC ROC Responses for PI LOPO (CK) feature Length=600 88
Figure 5. 12 Performance comparison between different algorithms. 90

LIST OF SYMBOLS/ABBREVIATIONS

SYMBOLS	EXPLANATION
AFA	Automatic Face Analysis System
ANN	Artificial Neural networks
AU	Action Units
CS	Compressive Sensing
CT	Curvelet Transforms
CWT	Cosine Wavelets Transforms
DCC	Dynamic Cascaded Classifier
DCT	Discrete Cosine Transforms
DDAG	Decision Directed Acyclic Graph
DWT	Discrete Wavelets Transform
ECOC	Error Correcting Output Code
EMFACS	Emotion Facial Action Coding System
FACS	Facial Action Coding System
FER	Facial Expression Recognition
HCI	Human machine Interface
HVC	Human Visual Cortex
JAFFE	Japanese Female Facial Expression
GWT	Gabor Wavelets Transforms
KNN	K Nearest Neighbor
LBP	Local Binary Pattern
LDP	Local Discriminate Pattern
LGBP	Local Gabor Binary Pattern
LXP	Local XOR Pattern
PCA	Principal Component Analysis
PSO	Particle Swamp Optimization
RBF	Radial Basis Function
SBDT	Single Branch Decision Tree
SCR	Sparse Representative Classifier
SVM	Support Vector Machine

1. INTRODUCTION

1.1.Overview

The Quest for information and intelligence gathering constitutes an integral part of modern-day technological demands. This has momentously grown ever since the major breakthrough in digital technology. In recent history biometric systems for identity establishment have attracted a lot of interests from numerous researchers across the globe. The motive may not be unconnected to tremendous need for such systems in many human applications. Some of the most visible and common use for identity and behavioral pattern recognition include face and facial expressions recognition systems. Facial Expressions Recognition (FER) in particular, have recently found its way into many viable applications in security, Human-Machine Interaction (HMI), medical diagnosis, psychology, entertainments and a host of other applications. The attendant hike in interest for FER in recent history, could be attributable to a pressing need for information technology and intelligence gathering as world leaps to an era of data technology.

The essence and evolution of FER has been identified way back in times. Researches in Human interaction and social psychology has indicated that facial expressions and human emotions constituted a fast and convenient ways and modality of communications amongst humans (Fasel & Juergen, 2003). In particular, facial expressions provide a very strong insight into individuals state of mind and level of interest. Automated analysis of expressed emotions brings an affective and more coherent into understanding HMI, (Min & Feng, 2013), (Wenfei et.al 2012). Incorporating additional affective dimension into HMI can significantly improve the users' experiences in the process. It also enhances both the tightness and effectiveness of the interaction. Moreover, the overall accessibility of the system as a tool for research in behavioral science and medicine and Pattern Recognition (PR) thereby providing a more explicit approach toward classification of facial expression.

In comparison, facial expressions recognition tends to be much trickier and difficult the than the face recognition in many ways (Min & Feng, 2013). In spite of the fact that face and facial recognition have reasonable overlap and similarity in respect to challenges (i.e. Image scaling, rotation, background illumination et al.), facial expression have additional, peculiar and localized challenges related human race, culture, variation in the intensity of a *stimuli* under the same emotion type. In addition,

the overlap in emotions brings about an intrinsic correlation between universal basic expressions such as Anger and Sad which may become indistinguishable in some instances due to excessive overlap and resemblance (Min & Feng, 2013).

Contextually, FER algorithmic framework remains similar to face recognition of the most counterparts' algorithms in Pattern Recognition. The general flow of the algorithm includes preprocessing, feature extraction, classification and lastly decision making using some established standard procedures (Wenfei, et.al 2012). Preprocessing provide an opportunity to attempt to correct some of the inherent noise and data error associated to the image acquisition environmental hardware distortion of the original captured image. FER feature extraction techniques are broadly identified as either Component-based (holistic) or Feature-based (local) approach. In Component-based approach the overall face image is of interest while in the Feature-based approach localized points (or Action Units) (Wenfei, et. al, 2012) on the face such as eye, nose, mouth, are of interest for extraction of geometrical measurements and localized information to make up the feature representation (Min & Feng, 2013).

In the hierarchy of FER algorithms framework, feature extraction and classifier algorithms have received enormous attention by researcher as vital stages upon which a great deal of performance of the algorithm could be realized with respect to both robustness and accuracy. This is not to downplay the contributions of the other stages which may be valuable as well and contribute to overall performance. Some of the commonly used feature extraction algorithms includes; Discrete Cosine Transform (DCT), Principal Component Analysis (PCA) or "Eigen faces", Local Discriminate Analysis (LDA) or "Fisher faces", Local Binary Pattern (LBP), Independent Component Analysis (ICA), Discrete Wavelets Transform (DWT), Curvelets Transform (CT), Gabor Wavelets Transforms (GWT), Complex Wavelets Transform (CWT) and the host of others (Shiqing, et. al, 2012). Some of feature extractors have a multi-resolution approach whereby sample data are analyzed by being broken down into smaller sub-bands of frequencies also known as *wavelets*. GWT for instance falls into that category and they have enjoyed comparative successful applications such as pattern recognition and biometrics (Min & Feng, 2013). For instance, authors in (Turk & Pentland, 1991) used PCA in their submission whereby Principal Eigen's vectors were extracted from the face image using PCA which were eventually used for classification. Due to low performance of the conventional feature extraction algorithms in FER, researchers upon deploy multi classification and fusion techniques at both features,

score and decision level of the both feature extractor and classifier algorithms. One group of feature extractors with better performance in FER is the multi-resolution feature extractors like GWT. Gabor Transform has hugely been successful in many face related applications and still remains an excellent feature extractor for face related patterns (Shishir & Ganesh, 2008). Its success can be attributed to its higher capacity (i.e. many number of tunable parameters) and response which draws similarity to receptive fields of a biological cells in the mammalian primary visual cortex of the eyes (Fasel & Juergen, 2003). Authors in (Yimo & Zhengguang, 2008) implemented variants of GWT feature extractors to obtain more robust features for FER problems. However, GWT and other multi-resolution algorithms upon demand intensive computation and higher memory usage (Baochang & Shiguang, 2007). Due to these setbacks, most of the times a dimensionality reduction algorithm is required to trim down the size of the feature vectors. This becomes a hindrance in real-time applications where speed and simplicity are of great interest.

Moreover, the classification phases have always played a vital role in the accurate estimation and prediction in pattern recognition problems which by a large gives the overall performance of a system. Classifiers that use distance as a metric (Distance metrics) such as Euclidean (δ_{l_2}) , Manhattan (δ_{l_1}) and Cosine (δ_{cos}) have been frequently in many classification problems, thanks to their simplicity and less expensive computational cost (Baochang & Shiguang, 2007). These basically require no training and can be very handy in a less complicated problems. However, for complex classification exercises, such may not be effective and as such a much more robust state-of-the-art learning-based classifiers (e.g. Artificial Neural Network, Support Vector Machine (SVM),) may have to be deployed to get a better result. Despite their ability to handle non-linear classification problems, learning-based classifiers are not without shortcomings either. This may include convergence instability (i.e. inability to learn the patterns), generalization error, training error, optimum parameter selection and a host of other challenges (Chapelle, et. al, 2002). Moreover, most of these classifiers were initially designed for binary classification and the extension of their original goal to multi-classification problems has been one of the active fields of research in pattern recognition problems. Some of the highlights of this thesis contribution includes, but not limited to, followings:

- Proposed an automatic method for optimum parameter selection for Radius Basis Function (RBF) kernel in support vector machines. The proposed

technique could be extended to other kernels in learning-based classifiers but the scope discussed in this paper is limited to RBF kernel in support vector machines.

- Proposed a new multi-classification approaches for general pattern recognition multi classification problems.
- Combined the two methods above to proposed a new classifier with an improved performance in comparison to the current state-of-the arts classifiers used in multi classification applications.

1.2.Dearth and Resurgence of FER

The idea that emotions are linked discretely with facial expressions has its genesis in the early works of Darwin (Darwin, 1872), and subsequently by those who have followed the teachings leading to refinements and elaboration of his evolutionist claims (Ekman, 1993). Darwin argued that facial expressions can be said to be residual actions from a much more complete behavioral responses which happen in combination with other responses in our bodies such as postural, vocal, gestural, skeletal muscle movements, and other physiological responses. He further argued thus, in human attack response, anger manifests through furrowing the brow and tightening of the lips with teeth; disgust manifests through a nose wrinkle, open mouth, and tongue protrusion similar to vomiting response. Therefore, Facial expressions, are described as a coordinated response involving multiple response systems. Darwin considered that all human, irrespective of culture or race, have in them a unique ability to exactly express some emotions on their faces. To prove his points Darwin wrote "The Expression of the Emotions in Man and Animals" as a rebuttal to the view held by Sir Charles Bell, who was the leading facial anatomist of that period and coincidently a teacher of Darwin's, about how God created human with unique facial muscles to uniquely express human emotions. With significant innovations in human anatomy and photography in that period, Darwin painstakingly studied the associated facial muscle actions in emotion, and then concluded that the muscle actions are universal, and have visible precursors both in non-human primates and several other mammals.

Early attempts to substantiate Darwin's view of emotions were futile (Ekman & Friesen, 1976). This let open the adaptation of contracting view in psychology– that facial expressions were culture-specific, much like language. Darwin's arguments were revisited once again by Tomkins (Tompkins 1962), who was of the view that emotion

was the basis of human motivation, and that the center of emotion was right in the face. By this thought Tomkins, found a perfect collaboration in opinion with Paul Ekman, who was credited to experimenting in the inversibility studies of emotions.

As a direct consequence of inconclusiveness in research, the study of facial expression suffered major setbacks in modern psychology until the 1960s. These are mainly due to some reasons namely: (1) Researches in those times hinted that face provides no correct information of emotions, and these findings were believed and held in psychology mostly; (2) The height behaviorism at early days blatantly rejects study of any "unobservable" such as emotion. This attitude becomes a stumbling block to many researchers whom may have wanted to pursue this path; (3) The unavailability of measurement tool for and estimation.

Research findings on emotions in early 20th century was unfavorable and all pointed to a flawed conclusion of lack of correlation or inability of the face to revel accurate information about emotion. In way of skepticism about findings in the 20th century Landis (Landis, 1924) captured photographs of subjects under numerous way that could stimulate some variants of emotional inhibition such as fear, disgust and excitement, he then presented the photographs to a set of observers for their views. The observers' judged that the emotion captured the photographs bear correlation to the eliciting or posed situations. Based on that Landis drew conclusion that the face could not in any sensible way provide accurate information about the state of the people's emotional.

Landis method of experiment was later heavily criticized by many researchers have as methodologically flawed in all respect (Coleman, 1949; Davis, 1934; Frois-Wittman, 1930). The modified replications of Landis's research in contrast, produced findings opposite Landin's. A more organized studies later on, on the perception of facial expressions of emotions revealed that humans could consistently recognize emotion in the face (Goodenough &Tinker, 1931; Woodworth, 1938). Nonetheless, these new findings struggle for acceptance in psychology, largely due to the old view of distorted experiments by Bruner and Tagiuri (1954). Early research on the face grapple with unavailability of measurement tool for quantification. There were attempts theearnest of the 20th o century, to take a direct measurement from a face (e.g., Frois-Wittman, 1930; Landis, 1924), with both failures and few degrees of success. However, mostly researchers depend heavily on evaluations from "judges" or by ratings of subjects' expressions, from still photos, films, or live observation.

Another major factor that devastated research in the emotion recognition was the "Zeitgeist of behaviorism". This attitude in early days blatantly rejected and dissuade the study of "unobservable" such as emotion. As a consequence, to early flawed findings which led to belief in psychological that the face did not provide accurate information about human inner states, it became a complete waste of time and resource to study facial expressions. In addition, whose who attempted to give the field a second-chance stood a risk of rejection and stigmatization by the mainstream who held an opposing view.

In rebuttal to perceived behavioral disdain towards facial expression, a series of activities in the '60s and '70s resuscitated the dying field of facial expression back into comity of psychological researches. In 1962, Silvan Tomkins summited a "theory of effect" which elaborated the central role of a face as a citadel of emotion. Similarly, in 1964, Tomkins and Mc-Carter had a quiet a convincing publication on systematic study of facial expression judgments, where they proved that observers' judgement methodology repeatedly identified facial poses as an indication of certain emotions. As a buildup on momentum, Ekman (Sorenson, et. al, 1969; Friesen & Ekman, 1971) and Izard (1971) conducted a seminal and well-known cross-cultural experiment to recognize emotions from facial expressions in both literate and preliterate categories. This brought to the fore that possibility of collecting of facial expression which could be enticing and rewarding as a research field.

The popular cross-cultural experiments in the late '60s and early '70s essentially revealed that recognition of emotions in the facial behavior of other people is very much practicable. It further highlighted that peoples from preliterate and diverse cultures agreed that a major set of prototypical facial expressions were linked to some certain emotions. Despite all of that there was no spontaneous facial expression studies in a large and widespread passion as would have been anticipated. However, this may be attributed to unavailability of measurement tools.

1.3.Objectives

The main objective of the thesis is to provide an entirely new approach to Facial Expression Recognition with an improvement to the existing state-of-the-art approaches. Some of the things hoped to be accomplished include the followings:

1. Propose a new feature extraction approach which is efficient, robust and fast compare to the existing feature extraction approaches.
2. Propose a classification approach which is robust, competitive and can be applicable not only to Facial Expression Problems but to other Pattern Recognition classification problems.
3. Propose a complete approach for Facial Expression Recognition with better performance in terms of accuracy (recognition rate) compared to the existing approaches.
4. Provide Facial Expression Recognition approach which will be adaptive to Human races and able to detect Facial expressions even when the expressions are weak in their manifestation.
5. Provide a Facial Expression Recognition approach which is comprehensive, efficient, and robust.
6. Provide a Facial Expression Recognition approach which is implementable in real-time.
7. Provide a Facial Expression Recognition approach which is reproducible by other researchers.

1.4.Methodology

Some of the critical methodologies to be adopted in the course of the dissertation to achieve the set objectives include the followings:

1. Facial image pre-processing techniques would be investigated and an efficient preprocessing approach would be proposed which will work in tandem with the subsequent stages in the overall algorithm flow.
2. Robust features would be extracted from the preprocessed facial images. In doing so, a feature extraction approach would be proposed which would essentially have the attributes such as robustness, consistency, less computations compared to the already established feature extractors in the literature.
3. Classification problems would be investigated and come up with some classifier algorithms having leverages over the existing classifier algorithms with respect to ability to learn feature vectors pattern and good generalization judgement.
4. Cross-validation of the overall research to ascertain its validity, limitation and advantages over the current existing research works in the same field. This

exercise would be done using the most standard and popular databases used in this field.

Standard experimental procedures will be used and all experimental constants will also be described and published to make the work comprehensive and reproducible by fellow researchers in accordance with the standard for global best practice in research

The thesis has been grouped into six chapters as follows: Chapter one comprises of the introduction to FER and historical background to the evolution of the field. Chapter two presented a walkthrough into the literature relevant to the word being undertaking. In chapter three theoretic background of some key concepts and algorithms used in the actualization new proposals being made were discussed. Chapter four presents the conceptual design and mathematical formulations of the contributions being proposed in the thesis. In Chapter five, results from experiments were presented compared with others and discussed. In the end, conclusion and suggestion for future work will be drawn in chapter six.

2. LITERATURE REVIEW

The resurgence of FER started to gather momentum in late 60's and early 70's after investigations rebutted the early perception about FER that existed in those days in which people's emotions and are believed to be are unconnected. This was made possible by the new emerging technologies and advancement in understandings measurement capabilities of facial expression in that time. The early researches in FER after its rebirth were hinged on anatomical studies of the facial muscular systems. These anatomical investigations provided a correlation between human emotions and motion of the facial muscles.

In what appears to be a rebirth of facial expression recognition, Ekman and Friesen (Ekman, 1972; Friesen, 1972) successfully studied responses to stressful films in American and Japanese college students and collected facial behavior. They investigated the effects of viewing conditions of the film in its social context and spontaneous expressions of emotion in display. The set up for both American and Japanese and males' students was to be subjected to viewing in two different scenarios namely; (i) firstly the film view is done alone and (ii) secondly in the company of an investigator of same ethnic identity as the investigator.

The Ekman experiment was a big breakthrough being the first of its kind which empirically studied what was known then as *display rules*. The which rules from the cultural point are viewed as norms or learned rules specific to cultural values which defined suitability to express an emotion and to whom an individual feels comfortable to express his feelings to another individual. The neurocentral theory of emotion proposed by Ekman and Friesen's in 1971, opinioned that display rules may be responsible for the noticeable cultural differences in emotion expression while not denying the existence of universality of emotions. Base on the earlier experiment,

Ekman and Friesen collected measurements from the American and Japanese students' facial behavior. They used an early observational coding system known as Facial Affect Scoring Technique (FAST) proposed by Ekman, et al, in 1971 (Ekman, Friesen, & Tomkins, Facial Affect Scoring Technique (FAST) first validity study, 1971). Their results unveiled credible evidences for universality of emotion and display rules operation. It reveals that when the two subjects view the films alone, they all exhibited similar expressions of negative emotions (e.g., anger, disgust and sadness). Whereas when viewing in the mix of an investigator, however, cultural differences surfaces.

In the 1970s, study of spontaneous facial behavior got a boot with application of electromyography whereby and activities or motions in muscles were examined in respect to a different type of psychological variables, such as rating of moods, psychiatric diagnosis and processes of social cognition (Schwartz, et. al, 1976, Cacioppo & Petty, 1979). Sequel to this development, observational coding algorithms evolved, with the most popular of them being the Facial Action Coding System (FACS) developed by Ekman & Friesen in (1979) and Maximally discriminative facial movement coding system (MAX) by Izard in 1978. The FACS system provides some descriptive schemes for virtually every visual facial movement that can be distinguished which are referred as *Facial Action Coding System* (*FACS)*. In FACS system 46 *AU*s that responsible for observable facial expression changes. The combination of AUs results in a large set of possible facial expressions. For instance, to express a smile emotion is viewed as the combination of upper lip raiser (*AU*10) and pulling lip corners (*AU*12 + 13) and/or mouth opening (*AU*25 + 27) with some degree of furrow deepening (*AU*11). However, the above AUs combination for smile may differ from one smile to another depending on the type. In spite of FACS shortcomings, this system has been the

most accurate and widely adopted method for measuring human facial motion from both human and machine perspective.

Mase (1991) and (Essa & Pentland, 1997) attempted to describe patterns of optical flow which matched several AUs, without attempting to recognize them. Bartlett *et al*. (1999) recorded some of the most extensively demanding experimental reports of upper and lower face AU recognition. In the experiments, both image sequences free of head motion were used.

Automatic recognition of action units (AU) using FACS system is a difficult problem, and relatively not much work has been reported. One of the intractable problem with AUs is lack of quantitative definition which by a large, implies that AUs can appear in complex combinations. Over times with improvement in Pattern Recognition and classifier algorithms that can effectively learn even a complex pattern and be able to make accurate predictions on the unseen patterns, databases for FER were collected and made available to researchers for algorithms developments in this field.

Ying-li. *et al.* (Ying-li, et. al, 2001), developed an Automatic Face analysis System (AFA) for facial expressions analysis taking into cognizance both permanent facial feature like (brows, eyes, mouth) and features that are temporal and transient in nature (feature that depends on furrows). The used a nearly frontal-view sequences of face image. The AFA was designed to be sensitive to a fine-grained change in facial expression and encodes those changes into action units (AUs) of the Facial Action Coding system (FACS). In contrast to earlier FACS system the proposed a multi-scale face and facial component models that can both track and model the various facial features therein. The feature that are being tracked include lips, eyes, brows, cheeks, and furrow and in the process of being tracked, a comprehensive parametric description

of the facial features was obtained. Using the extracted parameters as input, the system is able to recognize whether a group of action units (six upper face AUs, neutral expression and 10 lower face AUs) occur alone or in combinations. The system has been tested for generality using independent facial image database and FACS-coded ground-truth created by different researchers.

Kanade and others, (Kanade, et. al, 2000) elaborated the problem space for facial expression analysis, this included eliciting conditions, description levels, inter-expression transitions, validity and reliability of test and training data, personal differences in people, complexity of the scene, head orientation, image properties, and non-verbal behavior. Thereafter they collected AU-Coded Face Expression Image Database, encompassing 2105 digital image sequences of 182 adult subjects from different ethnicity. In the database, a multiple token of most primary FACS action units were performed at the time the research was published. They claimed the database to be the most elaborative test-bed yet for comparative studies of facial expression problems which is subsequently been adopted by many researchers.

Lucey P. *et al* (2010), worked to improve the original work on Facial Expression recognition which led to the birth of Cohn-Canade (CK) database. Three major drawbacks were identified in the previous work. Firstly, emotion labels are not well established, whereas AU codes are well established and validated. This is because they refer to the difficulties in evaluation of levels of emotions which always represent what was requested instead of what actually was performed. Secondly, lack of a standard metric for performance measurement and evaluation of new algorithms also hampered the evaluation procedures. Thirdly, there was no standardization in the collection of facial databases. Sequel to this, the Cohn-Canade (CK) database was extensively used in AU and emotion detection (in spite of latter levels were not being validated).

Similarly, no benchmark upon comparison can be drawn for new algorithm, and also meta-analyses becomes much more difficult by adopting random subsets of the original database. the Extended Cohn-Kanade (CK+) database was later on developed to address the shortfalls of the former (CK) database. The number of sequences got boost to 22% and the subjects' size was increased by 27%. For each expression sequence, the target was fully FACS coded and emotion labels got validated.

Michael and others (Michael, et al, 1998) proposed coding Facial expression images using multi-resolution at multiple orientation sets of Gabor filters that are aligned and ordered in topographical pattern with the face. Results from the experiment with the space derived from such representations was compared with one obtained from semantic ratings of the facial images based on human observations. They indicated in the experimental outcomes the possibility to construct a facial expression classifier using Gabor coding of the facial images at the input stages. They further highlighted that Gabor representation exhibited an enormous degree of plausibility psychological, which is a vital design feature that could be very handy for HMI systems. In an attempt to implement their methods, they built a database of facial expression which is later known as Japanese Female Facial Expression Database (JAFFE).

Zhang & others (Zhang, et. al, 2012) proposed facial expression recognition approach based on Multi-Scale Local Gabor Binary Pattern (MB-LGBP) feature extraction technique with multi-level classification. the MB-LGBP operator was designed to encode both local and global informative features presence in the sample. The proposed feature extractor fundamentally was utilizing Gabor wavelet transforms aimed at viewing the facial image in different scales of the Gabor filter and finally different extracted features are fused together to form a reinforced single feature vector which represents a sample. To classify the data, they proposed a two-level classification

method. In the first level (coarse), two of the expression samples that have first two highest decision confidence were taken from 7 universal basic expression classes using the MB-LGBP features. While in the other level (fine), one of the two candidate classes were verified as final expression class using slightly more delicate 2D MB-LGBP features.

Rao & Koolagudi, (2015) worked to extract acoustic and facial features from video stream for the soul aim of identifying emotions from streams of video images of human faces. They considerably and carefully explored variations in grayscale intensity of the pixels with eyes and mouth regions to help analyze and emotion-specific knowledge from an expression. The extracted features were later correlated with acoustic features presented in the form spectral and prosodic information using auto associations. They used Auto associative neural network models capture the emotion-specific information from acoustic and facial features. They discovered that a great deal of improvement could be realized by combination of the evidences of models developed from acoustic and facial features.

Min & Feng (2013) proposed a multi resolution approach to feature extraction in FER using Curvelet transform. To optimize the performance of the Support Vector Machine, they proposed way for parameter selection SVM using biologically inspired Particle Swarm Optimization (PSO) (Min & Feng, 2013). The curvelet transforms was applied to extract facial expression feature substitutivity. However, they realized that the amount of curvelet coefficients got in the first stage was very large and therefore presents a huge setback to be classified, therefore, adopted a method for dimensionality reduction by only retaining between 5 or 10% of the least-valued coefficients and discarding the remaining chunk of the extracted features. Finally, PSO algorithm was

deployed to search for the optimum parameters of SVM that will improve the accuracy of the classification.

Gu W and others (Gu, et. al, 2012) proposed an approach for facial expression recognition based on radial encoding of a local Gabor features. They first subject the facial images sample to local, multi-scale Gabor-filter operations, and then the results from the Gabor decomposition were encoded based on radial grids, mimicking the topographical map structures of the human visual cortex (HVC). The encoded features are passed to local classifiers to generate global features, which represent facial expressions. The authors claimed that the approach is robust with regards to missing information or corruption in data, hence the accuracy of the approach was reported to be high.

Zhang, and others (Zhang , et. al, 2012) utilized a method which both Gabor wavelet Transform and Sparse Representation Classification (SRC). In order to evaluate the performance of the SRC in the realm of facial expression, the firstly extracted Gabor wavelets features from the facial image. They used three different representation in classification, using three distinct classifiers representation viz: K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), experiment was conducted on the popular JAFFE facial expression database and an improvement has been reported by the authors.

Chao and others (Chao, et. al, 2015) proposed an approach to improve facial feature extraction, known as expression-specific local binary pattern (es-LBP). The In addition, in attempt to improve the correlation between expression classes and facial features, they proposed a class-regularized locality preserving projection (cr-LPP) approach, with main objective of minimization of the class independence and at the same time preserving the local feature similarity through dimensionality reduction. The

look at the intrinsic correlations there is among the basics facial expression classes (e.g. happiness, sad *et al*) and incorporated the concept of class regulization in the training of the Support Vector Machine classifier used during the training. They reported the method to have succeeded in maximizing inter-class independence thereby suppressing the effects of overlap an improving the recognition accuracy.

In their submission (Zavaschi et al, 2011), proposed a method for efficient facial expression recognition using an ensemble of classifiers with two different feature extraction algorithms. The concept was based on producing enough evidences and then choose the best from the pool of evidences at hand. They created chains of base classifiers using Gabor Wavelet features and Local Binary Pattern (LBP) (Zavaschi, Koerich, & Oliveira, 2011) as the two-different set of feature extractors. Then genetic algorithm with multi-objective function was deployed for searching the best ensemble using as objective functions the accuracy and size of the ensemble. An improvement in the recognition accuracy was also reported.

The prevalent use of multi-resolution algorithms in many applications such as Discrete Wavelets transform (DWT) and Gabor Wavelets Transform (GWT) in the field of FER indicated the edge these feature extractor poses over other others non-multi resolution extractors such as Principal Component Analysis (PCA), Local Binary Pattern (LBP), and Local Discriminant Analysis (LDA) (Min & Feng, 2013). Authors in (Ying & Fang, 2008) deployed Curvelets Transform (CT), a variant of multi-resolution, scales and orientation to form Curvelet products which are wrapped up around their origin. Curvelets coefficients were extracted from the products based on inverse CT which are subsequently adopted as feature representation. Despite claim of performance improvements, however, extensive computations were needed to arrive at that. similarly, due to the complexity in FER most of researches focus more on the

multi-resolution feature extractors like Gabor wavelet Transform (GWT) to encode features in FER for efficient representation. This due to the fact that efficient, unique and robust encoding of the facial expression feature are needed to make the exercise a success and effectively mitigate the effects of class overlap in FER. In generality, multi-resolution algorithms come with additional cost of complexity and presents a real challenge in terms of implementation. Moreover, GWT and other multi-resolution algorithms upon demand intensive computation and higher memory usage (Baochang & Shiguang, 2007). Due to these setbacks, most of the times dimensionality reduction algorithms are required to trim down the size of the feature vectors. This becomes a hindrance in real-time applications where speed and simplicity are of great interest.

On one hand, the classification stage in FER algorithms also have been a vital component for overall performance improvement or otherwise of most of the pattern recognition problems. Classifiers based on distance metric such as Euclidean (δ_{l_2}), Manhattan (δ_{l_1}) and Cosine (δ_{cos}) have been applied in many applications and research works credit to their simplicity and less computation cost (Baochang & Shiguang, 2007). However, for complex classification exercises, such may not be effective and as such a much more robust state-of-the-art learning-based classifiers (e.g. Neural Network, Support Vector Machine (SVM)) may have to be deployed to get a more reasonable result. Despite their ability to handle non-linear classification problems, learning-based classifiers are not without shortcomings either. This may include convergence instability (i.e. inability to learn the patterns), generalization error, training error, optimum parameter selection and a host of other challenges (Chapelle, et al, 2002). Moreover, most of these classifiers were initially designed for binary classification and the extension of their original goal to multi-classification problems has been one of the active fields of research in pattern recognition problems.

The generalization error is being considered as one performance enhancer for such classifiers and is linked to both errors on the training examples and the classifier complexity (Belhumeur, 1997; Vapnik, 1995). Trade-off exists between higher capacity classifiers (i.e. with large number of tunable parameters) and low capacity classifiers (i.e. within sufficient number of adjustable parameters). Low capacity classifiers might not be able to learn the task at all but when they do they exhibit good generalization due to their low complexity. On the other hand, higher capacity classifiers can learn any classification pattern according to the learning rules without error, but they generally tend to exhibit poor generalization. However, a good generalization performance can be achieved with higher capacity classifiers when the capacity of the classification function is matched to the size of the training set (Belhumeur, 1997; Vapnik,1997). A smart way to improve such classifiers generalization capacity is by optimizing kernel parameters with cross-validation data. Authors in (Dietterich & Bakiri, 1995) proposed a method for automatically optimizing multiple kernel parameters as a way of improving classifiers generalization accuracy using standard steepest decent algorithms whereas (Min & Feng, 2013) applied a heuristic optimization approach based on Particle Swarm Optimization (PSO). Despite the presence of number of proposed parameter optimization in the literature, researchers usually resorted to naïve search techniques or use parameters values based on experience to avoid excessive computational overheads especially in huge databases.

In this thesis, a number of noble approaches a new to Facial Expression Recognition with improved contributions at feature extraction and classification levels. The contributions encompass three different feature extraction algorithms and two new classifier algorithms which are referred to Dynamic Cascaded Classifier (DCC) and Single Branch Decision Tree (SBDT) classifier. The approaches provide a holistic and

comprehensive front towards addressing most of the concerns and shortcomings of work the research work highlighted in the literature.

The first feature extractor proposed got its inspiration from multi-resolution pyramid processing and compressive sensing theory. The second set of proposed feature extractor utilizes the concept of rotation invariant LBP commonly used in texture classification. The Gabor magnitude channels were encoded with monogenic LBP which, within the limit of this study, is referred to as mono-LGBP with LBP. The third feature extraction process utilizes the compressive sensing technique with statistical analysis of the extracted compressed facial signal to represent a more robust feature representation for each individual facial expression class.

In the classification stage, a dynamic classifier referred as Dynamic Cascaded Classifier (DCC) within the scope of this work, has been proposed. DCC leverages on the proposed feature extraction approach to produce an adaptive classification approach with kernel matching techniques which lead to low training error and improve generality of the classification process. In summary, the major contribution of this thesis includes: (a) new feature extraction techniques (b) proposes an optimized technique for automatic parameter selection in multilevel classification and (c) Dynamic Cascaded Classifying techniques. Moreover, another multiclass Classifier has been proposed and referred as SBDT. It conceptually bases on the support vector machine models but the training and evaluation are done via decision tree. These approaches have a combination of multi-resolution traits, simplicity and greater performance.

3. THEORETICAL BACKGROUND

As a prelude to better understanding of the underlining work being presented in the thesis, a comprehensive background on the relevant literature from which the work is being inspired become absolutely pertinent. The theoretical background highlights the basis and formulations of both theories and algorithm through their evolutions. This help to provide some discernible outlines and the extent of the thesis. It's understood that a history for the evolution of facial expression recognition has interwoven in many fields like psychology, sociology, human anatomy and pattern recognition where enormous volumes of research were produced to measure, quantify and predicts emotions, motives and visual communication being via facial expressions by humans. In view of this vast coverage and reflections of the essence of the topic, anatomical insights were also looked at. The anatomical understandings of the facial feature motions became the foundational blocks which helped the resurgence of this field by providing researchers with a tool upon which the biased perception of facial expression connections with human behavior were debunked.

3.1. Anatomy of Facial Expression Muscles

Facial expression muscles are rightly situated in a tissue known as **subcutaneous**. These tissues originated from the bone or fascia and getting through the skin. When they contract, they pull on the surface of the skin and their effects are exerted. They are basically the only muscles group that got inserted into the skin (Oliver, 2018).

The secondary pharyngeal arch is the has been the common embryonic origin for these group of muscles. The muscles group move from the arch, alongside with their nerve supply. Therefore, all these muscles of facial expression are innervated by the facial nerve.

In general, facial expression muscles can categorized into three mayor groups thus; **nasal group**, **orbital group** and **oral group**. Figure 3.1 shows the main oral muscles for facial expression recognition.

3.1.1. Orbital Muscles Group

This have essentially two (Orbicularis and Corrugator) different types of facial expression muscles under it that are connected to the eye socket. They are in charge of the movements of the eyelids and provide an essential protection the cornea from

external harm. They also got innervated by the facial nerve and further described below (Oliver, 2018).

3.1.1.1. Orbicularis Oculi

The orbicularis muscle encircles the eye socket extending into the surrounding of the eyelid. Based on its function it can be further sub-categorized into two; the outer **orbital** orbicularis muscle and inner **palpebral** orbicularis muscle (Oliver, 2018).

- **Functions of the muscles:** They basically responsible for the closure of the eyelid. **palpebral** orbicularis muscle closes the eyelid gently whereas the outer **orbital** orbicularis muscle closes the eyelid more forcefully.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.1.2. Corrugator Supercilii

In comparison corrugator supercilii muscle is literarily small muscle and is rightly situated, related to the orbicularis oculi muscle, in **posterior** (Oliver, 2018).

- **Placement:** It get its origin from the super-ciliary arch, extending all through a superolateral ways. It got inserted into the skin of the eyebrow.
- **Function: It main function is to bring** eyebrows closer together, thereby generating wrinkles on the bridge of the nose vertically.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.2. Nasal Muscles Group

This group of the facial muscles are responsible for displacement of the **nose**, and the skin around the nose. They are categorized into three groups; Nasalis, Procerus and depressor septic-nasi. In human literally insignificant.

3.1.2.1. Nasalis

In terms of size nasalis muscles are the largest compared to the other nasal muscles. They are further sub-categorized into two; Alar nasalis muscle and transverse nasalis muscle (Oliver, 2018).

- **Placement:** they all got their origin portions from the maxilla. The transverse nasalis muscle link to an aponeurosis across the dorsum of the nose. Whereas portion of Alar nasalis muscle link to the nasal skeleton via alar cartilage.
- **Function:** Alar nasalis muscle is responsible opening of the nares while transverse nasalis muscle is responsible for the compression of the nares.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.2.2. Procerus

Procerus muscle, of all the nasal muscles, is rated as the most **superior** of them. in respect to the other muscles of facial expression, it is situated **superficially** across (Oliver, 2018).

- **Placement:** It got its origin from the nasal bone, getting into the lower portion of the medial forehead.
- **Actions:** They are responsible for the transverse wrinkles over the nose by pulling down eyebrows when they contract.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.2.3. Depressor Septi Nasi

The Depressor Septi Nasi muscle helps the Alar nasalis muscle of the nasali in the process opening the nostrils.

- **Placement:** It is situated right from the maxilla, above the medial incisor tooth, to the nasal septum.
- **Function:** It is responsible for the opening of nares by pulling the nose from the inferior portion**.**
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.3. Oral Muscles Group

In the context of facial expression, they are most vital group of expresser muscles. Their activities result in the motion of the **mouth** and **lips** portions. These motions are essentially prerequisite for activities like whistling and singing and they bring to bear the essence of vocal communication. They comprise of other muscles such as orbicularis oris muscle, buccinator muscle, and numerous small muscles (Oliver, 2018).

3.1.3.1. Orbicularis Oris Muscle

They are the gateway to the opening to the oral cavity via their fibers of the orbicularis.

- **Placement:** They originated from the maxilla and also some got their origin from the other muscles of the cheek. they got inserted into the mucous membranes of the lips and the skin.
- **Function: Responsible for the** Pursing of the lips.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.3.2. Buccinator Muscle

Buccinator muscle is situated in a region between maxilla and mandible, deeper into the other facial muscles (Oliver, 2018).

- **Placements**: They got their origin from the mandible and maxilla. The muscles fibers blend with the orbicularis oris and the skin of the lips and go through the inferomedial medial ways.
- **Function**: It prevents the accumulation of food in the region of buccinator by pulling the cheek inwards against the teeth.
- **Innervation nerve:** They got innervated by the Facial nerve.

3.1.3.3. Other Oral Muscles

The other oral muscles referred to here are that contribute to the actions or movements lips and mouth. From the anatomical perspective, they are categorized into upper and lower (Oliver, 2018):

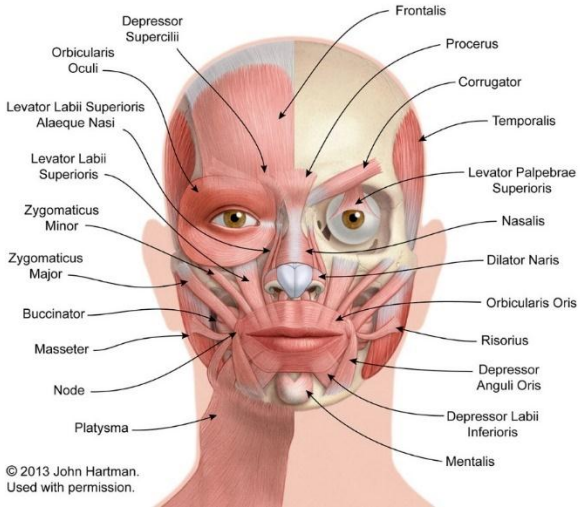


Figure 3. 1 The main oral muscles of facial expression. *(Oliver, 2018)*

3.1.4. Facial Action Coding Systems (FACS)

FACS is an encoding system which is rooted on anatomy for taking measurements of all sort observable facial behavioral pattern. The FACS system underpinned 44 Action Units (AUs) responsible for independently motion of the all facial muscles and musculature corresponding to all possible movements on human face. It relies on the functional anatomy for its identifications rather than structural approach. It organizes muscles or their combinations into AUs base on their functions. For example, considering frontalis muscle as a unit muscle, the inner portion and the outer portion of this single muscle can act independent of one another. FACS is able to differentiate and encode this variation between these two different actions (AUs) as AU1 and AU2, inspite of their common origin and categorization of this muscle by the anatomists as a single entity of muscle. Similarly, in human for instance, corrugator supercilli muscle, depressor supercilli muscle, being anatomically classified separately, act dependently of one another to draws the brows down and together. In anatomy perspective the action is caused by three separate muscles but FACS is able to identified it as a single FACS, AU4.

FACS system in theory is capable of encoding all the facial behavioral pattern exhibited, not confined only to emotion related actions. The Coding schemes encompasses in delicate identification of facial muscles innervation during expression using the observable change in appearance in the moments a particular expression happens. It also comprises of the intensity rating, laterality, and timing of the various

AUs. Knowing the fact that each of the 44 AUs can trigger independently of one another, and that each AU can be coded for rated intensity, laterality, and timing features, FACS system is very detailing and elaborative, as such makes it complex and effort demanding. In Table 3.1 AU codes are listed which Darwin considered universal. Subsequently, research proved the expressions to be universal (Ekman & Friesen, 1976). Figure 3.2 depicts a sample of FACS coding for a fear expression.

Table 3. 1 Facial Action Coding Systems

AU Number	FAC Name	Muscular Basis
1	Inner Brow Raiser	Pars Medialis, Frontalis.
2	Outer Brow Raiser	Pars Lateralis, Frontalis.
4	Brow Lowerer	Depressor Supercilli; Corrugator Depressor Glabellae;
5	Upper Lid Raiser	Levator Palpebrae Superioris.
6	Cheek Raiser	Pars Orbitalis, Orbicularis Oculi.
7	Lid Tightener	Pars Palebralis, Orbicularis Oculi.
9	Nose Wrinkler	Alaeque Nasi , Levator Labii Superioris.
10	Upper Lid Raiser	Caput Infraorbitalis , Levator Labii Superioris.
11	Nasolabial Fold Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Cheek Puffer	Caninus
14	Dimpler	Buccinnator
15	Lip Corner Depressor	Triangularis
16	Lower Lip Depressor	Depressor Labii
17	Chin Raiser	Mentalis

18	Lip Puckerer	Incisive Labii Inferioris, Incisivii Labii Superioris.
20	Lip Stretcher	Risorius
22	Lip Funneler	Orbicularis Oris
23	Lip Tightner	Orbicularis Oris
24	Lip Pressor	Orbicularis Oris
25	Lips Part	Relaxation of Mentalis or Orbicularis Oris , Depressor Labii.
26	Jaw Drop	Masetter; Temporal and Internal Pterygoid Relaxed
27	Mouth Stretch	Pterygoids; Digastric
28	Lip Suck	Orbicularis Oris

Table 3. 2 Emotions and associated AUs

Emotion	Darwin's Description	FACS AUs in Darwin's Description	FACS AUs in research Associated with this Emotion in
Anger	Nostrils raised, mouth compressed, furrowed brow, eyes wide open, head erect, (chest expanded, arms rigid by sides, stamp ground, body sways backwards/forwards, tremble]	4, 5, 24, 38	4, 5 and/or 7, 22, 23, 24
Contempt	lip protrusion, nose wrinkle, partial closure of eyelids, turn away eyes, upper lip raised, (snort, body expiration, expiration)	9, 10, 22, 41, 61 or 62	Unilateral 12, unilateral 14
Disgust	lower lip turned down, upper lip raised, expiration, mouth open, spitting, blowing out protruding lips, clear throat sound, lower lip, tongue protruded	10, 16, 22, 25 or 26	9 and/or 10, (25 or 26)
Fear	eyes open, mouth open, lips retracted, eye brows raised, (crouch, pale, perspiration,	1, 2, 5, 20	1, 2, 4, 5, 7, 20, (25 or 26)

	hair stands on end, muscles shiver, yawn, tremble		
Happiness	eyes sparkle, skin under eyes wrinkled, mouth drawn back at corners	6, 12	6, 12 a
Joy	zygomatic, orbicularis, upper lip raised, naso labial fold formed (muscle tremble, purposeless movements, laughter, clapping hands, jumping, dancing about, stamping, chuckle/giggle)	6, 7, 12	6, 12
Sadness	corner mouth depressed, inner corner eyebrows raised (low spirits	1, 15	1, (4), 15, (17)
Surprise	eyebrows raised, mouth open, eyes open, lips protruded, (expiration, blowing/hissing, open hands high above head, palms toward person with straightened fingers, arms backwards)	1, 2, 5, 25 or 26	1, 2, 5, 25 or 26

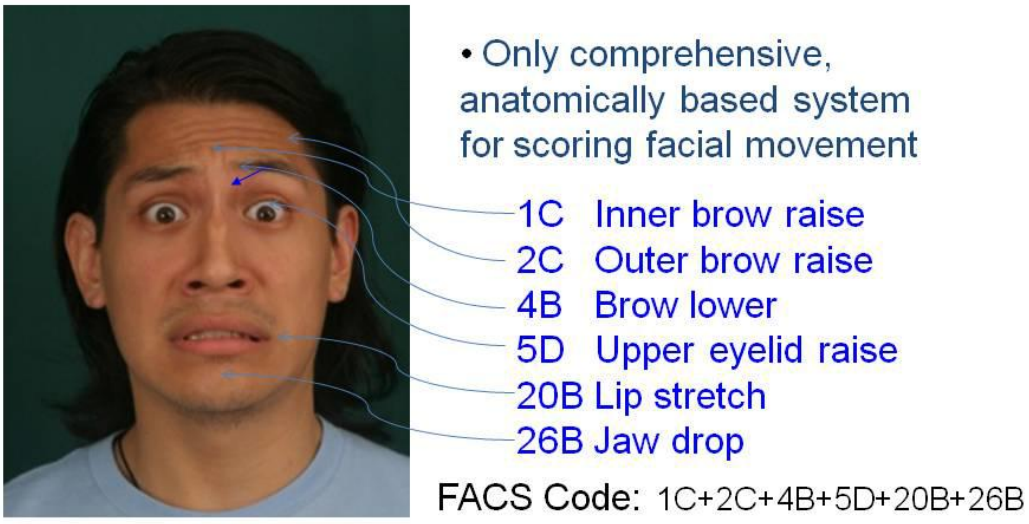


Figure 3. 2 Sample of FACS coding of a Fear Expression *(Oliver, 2018)*

Out of the 44 FACS AUs which has been identified, 30 AUs are identified to be associated with the contractions of specific facial muscles anatomically. Twelve are connected to upper face while 18 are connected to lower face. AUs occurs in a singular form or as a combination. The combinational occurrence of the AUs may lead to an *Additive* effect whereby the combination does not affect the appearance of the constituents AUs. Whereas in *non-additive* scenario, the appearance of the constituent

AUs does change. Despite the fact that number of AUs are relatively small, a sizeable number of AUs combinations have been observed. Figure 3.3 and 3.4 depicts the combination of AUs in both upper and lower face respectively (Ying-li, Kanade, & Cohn, 2001).

NEUTRAL	AU 9	AU 10	AU 12	AU 20
Lips relaxed and closed.	The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent.	Lip corners are pulled obliquely.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU15	AU 17	AU 25	AU 26	AU 27
The corners of the lips are pulled down.	The chin boss is pushed upwards.	Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched open and the mandible pulled downwards.
AU 23+24	AU 9+17	AU9+25	AU9+17+23+24	AU10+17
Lips tightened, narrowed, and pressed together.				
AU 10+25	AU 10+15+17	AU 12+25	AU12+26	AU 15+17
AU 17+23+24	AU 20+25			

Figure 3. 3 AUs Combination for Emotions in Lower Face
(Ying-li, et al, 2001).

NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 3. 4 AUs Combination for Emotions in Upper Face
(Ying-li, et al, 2001).

A quite number of FACS variants emerged over times. For example, Ekman and Friesen (Ekman & Friesen, EMFACS, 1982) proposed a short version of FACS system known as Emotion FACS (EMFACS), which is limited and confines only AU that are relevant to emotion theoretically or empirically. EMFACS system therefore is less complex than the original FACS and is specific to identification of a subset of AUs or AU combinations, reducing the overall labor intensity and improving efficiency.

3.2. Holistic-based Approach to FER

In recent times, collection of a standard facial expression database surfaced. This is aided by the establishment of universality facial expression connected to emotion signaling. Most of these facial expression databases are multi-cultural, multi-racial and were collected under different conditions using the well-established FACS system. Example of facial database collected using FACS is the popular Cohn-Kanade (CK) and its extended version which were subsequently grouped into the 7-basic universal facial expression of emotions (Disgust, Contempt, Fear, Surprise, Happy, Anger, Sad). Moreover, other facial expression databases were collected through rigorous experiments supported by institutions involving numerous voluntarily participants. Samples of the basic emotions expression obtained from these individuals under varying

conditions are analyzed and categorized by experts to form a viable database. Example of a database collected in this form is the popular facial expression database. The availability of these standard facial expression databases couple with advancement in machine learning algorithms bred the ground for another approach to automatic emotions recognition which is holistic or appearance-based. The Holistic approach or Appearance based views the facial image as single input and try to extract the most unique information about the sample that differentiate it from other samples that fall into a different category with it. The representation of that unique features, which is called Feature vector, is used to train a learning algorithm using many samples from the database, to iterate and learn the pattern therein in the samples and be able to make prediction when a never-seen-before sample of facial expression is presented to the learning algorithm.

It is important to note that learning process is mostly knowledge based and help to provide fast and much simpler approach to facial expression recognition compared to complex but more universal systems like FACS. The holistic-based facial expression recognitions are now well established with a quite huge number of approaches being proposed but nonetheless a general framework and algorithmic flow for these systems is remains unchanged. The pattern for these algorithms begins with preprocessing of the facial expression sample image to mitigate effect of noise and poor contrast in the sample as one of the inherent problems with image acquisition. The proposed sample is used to extract feature vector which more or less a representation of the preprocessed sample in another dimensional space. The feature vectors are then used to train a learning algorithm (classifier) to learn the pattern of the feature vectors od each sample. Finally, based on the prior knowledge of the classes which the learned patterns belong to a decision can be made to classify a new sample. Figure 3.5 depicted the seven-universal facial expression of emotions Whereas the general flow chart of the Appearance-based approach to facial expression recognition is shown in Figure 3.6. These emotions classes represent the emotions classes found in most of the facial expression databases.

Figure 3. 5 The seven universal Facial expression of emotions
(Oliver, 2018)

3.3.Feature Extraction Algorithms

One of the key stages in holistic-based approach to facial expression recognition is the feature extraction stage. Poor representation of facial expression sample image in its vector space in-adversely makes the learning of the patterns extremely difficult in the classifier algorithm. In other words, some classifiers may fail to learn at all due to incoherent and non-distinctive pattern in the feature vectors of the input samples. Having understood this effect and also complexity of facial expression data, complex and robust feature extractions algorithm were investigated and applied by many researchers in this field. The feature extractors must have efficient and rich projections of the sample data to be able to elaborate and distinguish even the least sensitive emotion variation in the sample data. Sequel to this, multi-resolution feature extractors have been seen by many researcher as viable candidates capable of projection of the sample data into a rich vector space at different resolutions and scales. The sections under this chapter discuss some of the frequently used feature extraction algorithms in FER as related to the scope of this work.

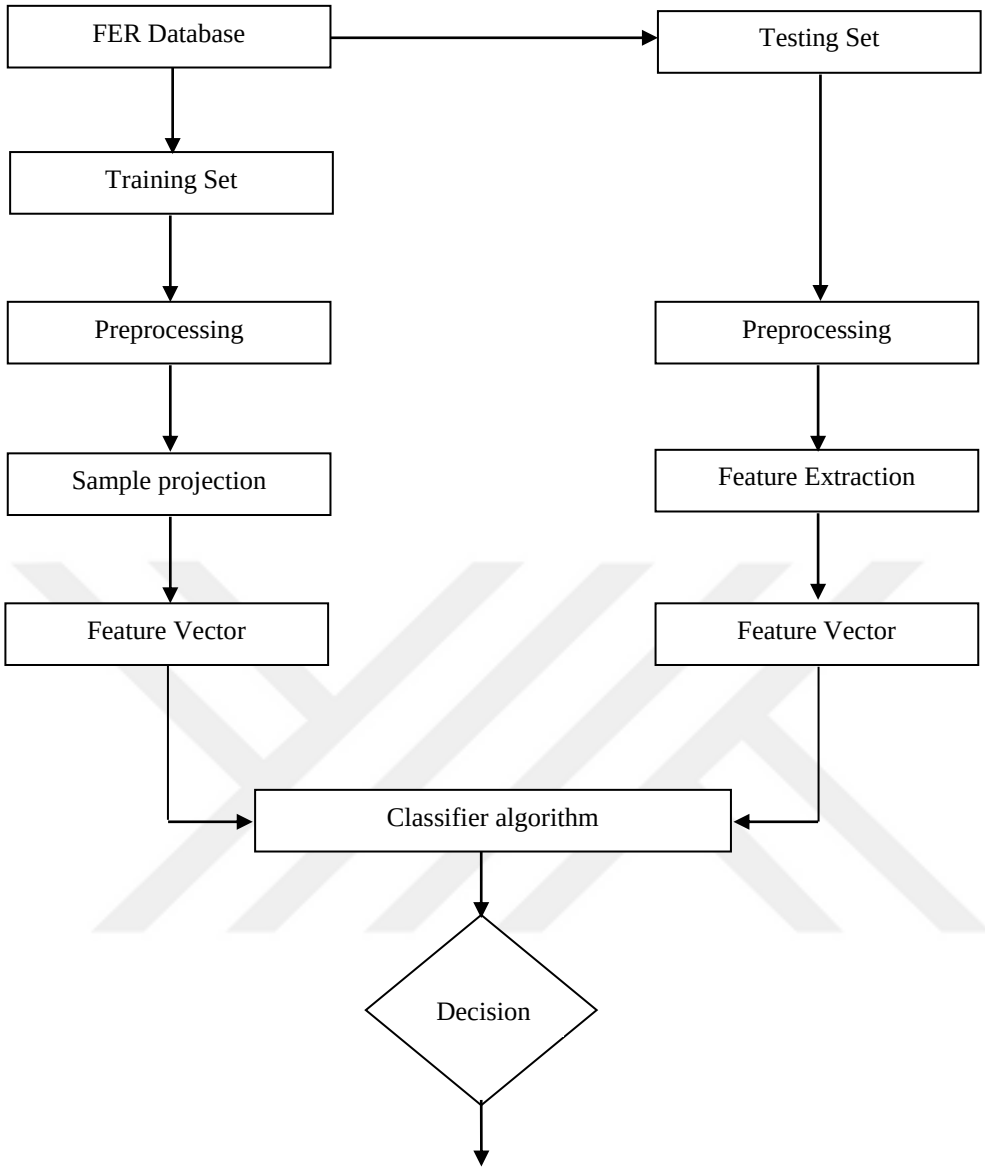


Figure 3. 6 Appearance-based approach to FER

3.3.1. Local Binary Pattern (LBP)

Local Binary Pattern (LBP) algorithm first appeared in the literature through the work of Ojala and others (Ojala, et al, 2002). For pixels f_p (P = 0, . . . ,7) of an image where f_p represent gray intensity of a neighboring pixel and P as its position, LBP operate by sliding through all the pixels in the image each time moving it center to a pixel with an intensity value f_c and computing a binary number around 3×3 neighborhood of the center pixel. Each of the center pixel within the window is computed using the function $S(f_p - f_c)$ as represented by Equation 3.1.

$$S(f_p - f_c) = \begin{cases} 1 & f_p \geq f_c \\ 0 & otherwise \end{cases} \quad (3.1)$$

To compute the new pixel magnitude for the center pixel f_c , binomial factor 2^p is assigned for each $S(f_p - f_c)$ as such the LBP is computed as shown by Equation 3.2:

$$LBP(f_c) = \sum_{p=0}^{7} S(f_p - f_c)\, 2^p \quad (3.2)$$

Over times LBPs have proved to be powerful and very effective in digital image processing. It has been applied in many applications such as motion detection, visual inspection, and texture analysis. One of the very useful attributes of LBP features is simplicity in terms of computation and robustness against monotonic illumination changes. The LBP operator has the strength to detect a variety of different textures (e.g., line end, spot, edge, corner), being accumulated into a histogram typically over a region to encode and extract localized texture information.

3.3.1.1 Uniform and non-uniform LBP

The original LBP considers a center pixel with a 3×3 neighborhoods which is limited to only 8 neighbors' pixel. This concept was extended by considering of different sizes of neighborhood, not limited to 8-neighbors, to encode dominant features at different scales. For $LBP(P, R)$ denotes an operator with P neighborhood and sampling points that are spaced equally on a circle of radius R. It was also observed that a small portion of the 2^p patterns accounted for the majority of the texture of images. These patterns, referred to uniform patterns, consist of at most two binary bitwise transitions from 1 to 0 or 0 to 1 for a circular binary string. For instance, 01100000 and 11011111 are uniform patterns. The above binary patterns, can be handy for representation of textures such as spots, edges and corners. In general, uniform patterns consist of $(P-1)P+2$ total binary patterns. The portion $(P-1)P$ represents rotational patterns, including edges and two other patterns that are non-rotational, spots and flat areas. For the estimation of LBP histogram, every uniform pattern has a separate bin in the histogram, and all non-uniform patterns are categorized under a single bin. Base on the uniform patterns for an 8-bit gray value, for a single cell, the size of the feature vector decreases from 59 to 256. In an 8-bit pixel value there are 58

uniform binary patterns thus; “0”, “1”, “2”, “3”, “4”, “6”, “7”, “8”, “12”, “14”, “15”, “16”, “24”, “28”, “30”, “31”, “32”, “48”, “56”, “60”, “62”, “63”, “64”, “96”, “112”, “120”, “124”, “126”, “127”, “128”, “129”, “131”, “135”, “143”, “159”, “191”, “192”, “193”, “195”, “199”, “207”, “223”, “224”, “225”, “227”, “231”, “239”, “240”, “241”, “243”, “247”, “248”, “249”, “251”, “252”, “253”, “254” and “255”. LBP Patterns whose bitwise transition *U(x) > 2*, are known as non-uniform LBP patterns:

$$LBP_{P,R}^{u2} = \begin{cases} z \;\; if \;\; U\left(LBP_{p,R}\right) \leq 2 \;\; LBP_{p,R} = I_Z, \;\; I_Z \in I \\ \qquad where \; |I| = (P-1)P + 1 \\ (P-1)P + 2 \;\; otherwise \end{cases} \tag{3.3}$$

Where the term $U\left(LBP_{p,R}\right)$, is computed using Equation 3.4.

$$U\left(LBP_{p,R}\right) = \left\|S(f_{p-1} - f_c) - S(f_0 - f_c)\right\| + \sum_{p=0}^{p} \left\|S(f_p - f_c) - S(f_{p-1} - f_c)\right\| \tag{3.4}$$

Equation 3.3 denotes a uniform pattern LBP ($LBP_{P,R}^{u2}$) which is only true when the terms U(x)=2. When the term U(x) is less than 2, it is represented by I_Z otherwise the patterns are referred to as non-uniform LBP Pattern with values equals to $(P-1)P+2$. The term I_Z which is an indexing function consists of $(P-1)P+2$ indices that are assigned to every uniform pattern. For LBP uniform patterns with 8-neighborhood (P = 8), its histogram decreases from 256 bins to 59 bins (58 bins belong to uniform patterns and last 1 bin belongs to non-uniform patterns) without sacrifice to performances (Ojala, et al, 2002).

3.3.1.2 Rotation invariant uniform LBP

The $LBP_{P,R}^{riu2}$ is a decedent of the original LBP operator. Its main objective is to address some of the shortcomings of the original LBP by rendering it insensitive to image rotation with further reduction in the overall length of the feature vector in its histogram form. In this way an improved performance and robustness and adequate representation of the patterns is excepted. Rotation invariant LBP operates by right shifting a circular bitwise on the P-bit number *x*, for *i* times as indicated by Equation 3.5.

$$LBP_{P,R}^{ri} = min_{i=0}^{P-1}\{ ROR(LBP_{P,R}, i)\} \quad (3.5)$$

Where $ROR(x, i)$ stands for the bit shift operator. Rotation invariant LBP decreases the size of the feature length. For instance, for P = 8, $LBP_{P,R}^{ri}$ consist of 36 unique rotational invariant patterns. rotation invariant uniform pattern $LBP_{P,R}^{riu2}$ is generated from the rotation invariant patterns $LBP_{P,R}^{ri}$, by simply using the idea of LBP. This further trims down bins size from 36 to 9.

3.3.1.3 Magnitude LBP

LBP operator has also been used for gradient magnitude estimation image and the operator is referred to Magnitude LBP (LBP^{gm}). To generate an image representing gradient magnitude, gradients along both x and y directions are computed. To realize this a first order derivative operator will suffice (e.g. Sobel). The derivative operator s computes the gradient each pixel intensity as it slides over. The output of this operation indicates smoothness or sudden variation in the intensity from one point to another. The operation consists of sliding a mask of two 3×3 sizes and then apply a convolution operation with the original image similar to the operation of popular gradient operators like Sobel as represented by Equations 3.6 and 3.7.

$$I_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} *\mathrm{I} \text{ and } I_y = \begin{bmatrix} +1 & 0 & +1 \\ +2 & 0 & +2 \\ +1 & 0 & +1 \end{bmatrix} *\mathrm{I} \quad (3.6)$$

$$I_{gm} = \sqrt{I_x^2 + I_y^2} \quad (3.7)$$

Where * stands for convolution operation. *I* represent the input image, where I_x and I_y are the outputs of the containing both the vertical and horizontal derivatives approximations. From the generated gradient approximations, it is possible to derive a gradient magnitude image. I_{gm} is the gradient magnitude.

LBP^{u2} is operated on I_{gm} to get LBP^{gm} features maps. This method is basically dependent on derivative LBP which extracts the local variation within the image. This feature has been reported to have been applied in many face related researches.

Usually the extracted features above are combined together by concatenating them into one feature vector Histogram (HG) of LBP^{xxx}, consisting of an *n* number of sub-blocks as formulated Equation 3.8:

$$HG(LBP^{xxx}) = (H_0, H_1 \dots \dots H_{n-1}) \tag{3.8}$$

For the r^{th} sub-block of LBP^{xxx} its histogram with a total number of bins U, is computed as follows (Equation 3.9):

$$H_r = (h_{r,0}, h_{r,1} \dots \dots h_{r,n-1}) \tag{3.9}$$

Where *h* is defined as in Equation 3.10:

$$h_i = \sum_{xy} I\{LBP^{xxx}(x, y = i)\}, \quad i = 0,1,2 \dots u - 1 \tag{3.10}$$

where *i* is the i^{th} bin of histogram *h* and *hi* represent the number of patterns contained with LBP^{xxx} pattern *i*.

$$f(x) = \begin{cases} 1, & if\ A\ is\ true \\ 0, & otherwise \end{cases} \tag{3.11}$$

3.3.1.4 Multi-scale LBP

The concept of multi resolution could be realized in LBP by changing the values of P and R. The multi-scale LBP, LBP^{ms} has an edge over the standard LBP in application such as face recognitions (Arof & Deravi, 1998) and facial expression recognitions with frontal view samples (Maenpa, et al, 2000). The first multi-scale parameter for LBP^{ms} is *P* which stands for neighborhood size. With a big size *P*, the histogram features size increase which in turn scales up the computational complexity of process. Smaller value of *P* could be detrimental by decreasing performance due to loss of vital information. The other parameter *R*, which is represents the size of multi-scale operators. There is also a tradeoff with this parameter where small value leads to insufficient information especially for application such as facial expression recognition, where as a bigger value of the operator can affect the accuracy of the algorithm by reducing the number of uniform patterns.

3.3.1.5 Local Binary XOR Operator

This operator bears resemblance with LBP and only differs from the LBP by using XOR operator to instead hard decision operator used by LBP. Within the neighborhood the operator performs XOR operation between the center pixel and its neighboring pixels. To realize this XOR operation the pixel magnitudes must be binarized first (Zhang, et al, 2007). For example, the output of a convolution operation between a sample facial frontal view image with Gabor mask can be thresholded with a hard decision value to get the binary image. Figure 3.7 depicts LXP is operation on binarized image. To compute of $I(z)$ using LXP, the neighborhood pixels are XOR-ed with the center pixel $I(z)$ and then the generated code is converted to an integer value.

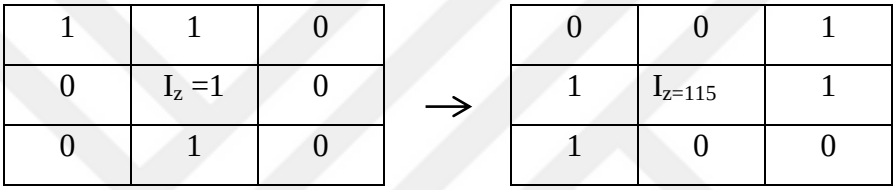


Figure 3. 7 LXP Operator

For example, LXP code for the center pixel $I(z)$ on the right starting from top-left corner and rotating clockwise is '00111011' or 115 as an integer.

3.3.2. Gabor Wavelets Transforms

Gabor wavelets are obtained from a set of filters as a when Gaussian function is modulated with a sinusoidal wave. The convolution operation of Gabor filter, $\psi_{\theta,v}(z)$ with an image, $I(z)$ can be denoted by $G_{\theta,v}(z)$ in Equation 3.12.

$$G_{\theta,v}(z) = I(z) * \psi_{\theta,v}(z), \tag{3.12}$$

where, $z = (x, y)$ and is the 2-Dimensional pixel's indexing along x and y planes. The term '$*$' represents 2-Dimentional convolution operator. The terms θ and v represent orientation and scales parameters of the filters, respectively. This Gabor kernel is mathematically represented in Equation 3.13:

$$\psi_{\theta,v}(z) = \frac{\|k_{\theta,v}\|}{\sigma^2} e^{\left(-\frac{\|k_{\theta,v}\|^2 \|z\|^2}{2\sigma^2}\right)} \left[e^{-ik_{\theta,v}z} - e^{-\sigma^2/2}\right] \quad (3.13)$$

The operator $\|.\|$ is a norm and σ represents the standard deviation of the distribution whereas vector $k_{\theta,v}$ is described by Equation 3.14.

$$k_{\theta,v} = k_v e^{-i\phi_\theta} \quad (3.14)$$

where $k_v = k_{max}/f^v$ and $\phi_\theta = \pi\theta/8$; k_{max} stands for maximum frequency, ϕ_θ the kernel's orientation and f stands for kernels spacing in the frequency domain (Eleyan, et al, 2008; Liu & Wechsler, 2003; Cootes, et al, 1995).

3.3.3. Image Pyramid Processing

Image pyramid processing is relatively a simple in its concept and complexity wise is one of the most formidable and earliest way of multi-resolution of an image. The inspiration for pyramid processing, and by extension most of multi-resolution algorithms, is to be able to exploit other dimensional properties of an image that may appear salient in some other resolutions or dimension. In this way features that are hard to extract in one resolution may be conspicuously distinct in other resolution and then can easily be extracted (Gonzales & Woods, 2008). For example, objects with low contrast tends to exhibit more details information at higher resolutions whereas objects with higher contrast reveal more detail at lower resolution (Gonzales & Woods, 2008). In a natural where there is coexistence of both objects with low and high contrast, the chances that a more robust representation of that image by extracting its features at different resolutions will definitely be enhanced.

Image pyramid can be viewed as a set of decreasing resolution images placed one over the other in a shape of a pyramid. From the bottom base of the pyramid both resolutions and sizes are large and decrease as one moves to the top of the pyramid. Every pyramid level is derivative of its bottom pyramid level and is computed by convolving approximation filter or (low-pass filter) with the bottom pyramid level image and the results are down sampled along rows and columns to a factor of 2. The algorithms have a reconstruction capability where original image can be reconstructed

from its pyramid by using level residual and interpolation filters (Gonzales & Woods, 2008). Figure 3.8 and 3.9 depicts decomposition process into image pyramid levels.

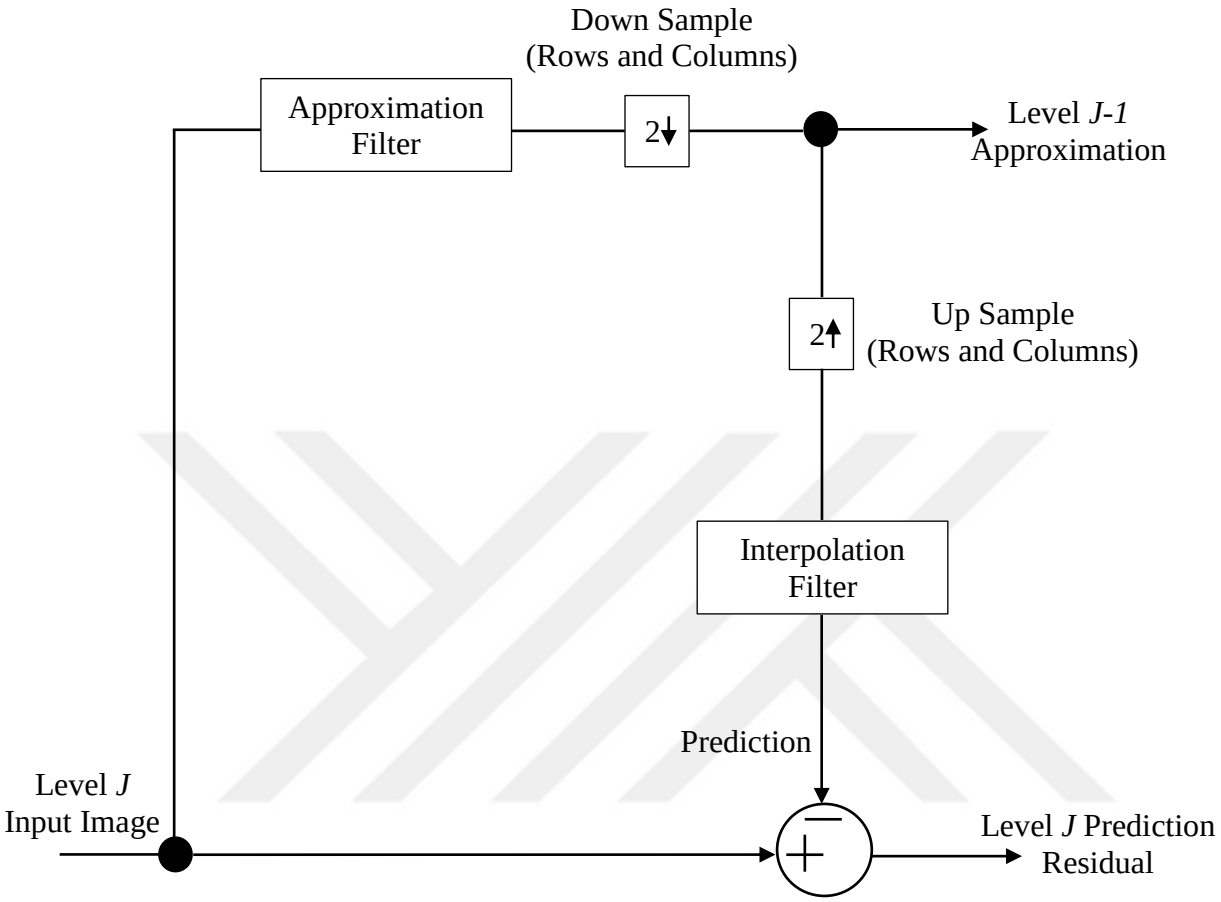


Figure 3. 8 Pyramid decomposition schematic

A variety of a low pass filters can be used for the pyramid decomposition. For instance, 2D Gaussian low pass filter can suffice. For Gaussian filter, its weights can be constructed by Equation 3.15. The filter has a mean distribution at the center point of the kernel i.e. $k(x,y) = k(0,0)$. Where x and y represent the horizontal and vertical coordinates of the kernel respectively.

$$k(x,y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}+\frac{(y-\mu)^2}{2\sigma^2}\right]} \quad (3.15)$$

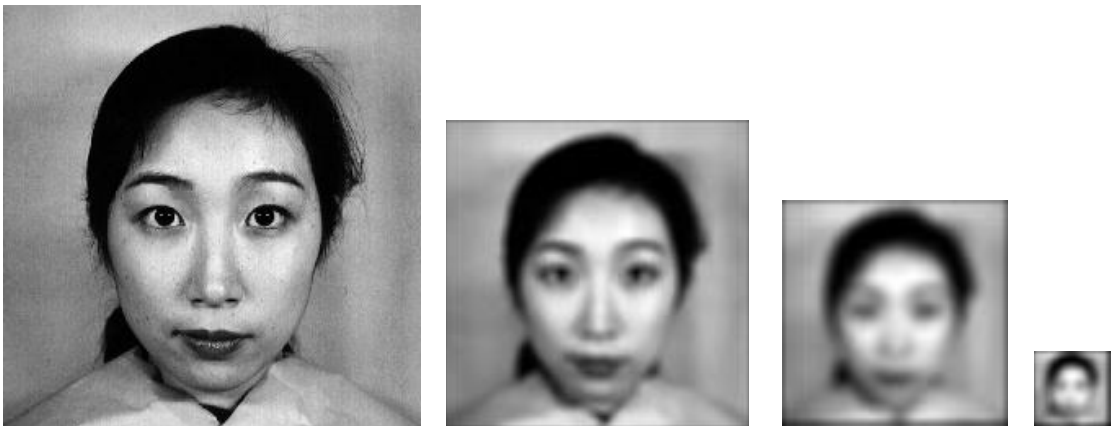


Figure 3. 9 four-level Pyramids decomposition using Gaussian Filter

3.3.4. Compressive sensing Theory

After the landmark Nyquist-Shannon sampling theory, an evolution into the digital and information technology era had its major boosts. The theory provided a limit for efficient reconstruction of an analog signal sampled at regularly spaced intervals (i.e. period). It states that if a signal is sampled by a frequency at least twice its bandwidth, then the signal can be perfectly reconstructed from its samples (Shannon C. , 1949). Sooner enough, researchers realized sampling at Nyquist-*rate* results into a large number of samples resulting to large data samples, inefficient use of communication channels and huge processing of data. Inspired by the information theory (Shannon C. E., 1948), number of signal encoding algorithms evolved over times in an attempt to reduce the signal size samples without loss of information. For instance, in JPEG compression, images to be compressed are transformed into the Discrete Cosine Transforms (DCT) domain and then most of the of representation with negligible amplitudes value of DCT coefficients are thrown away (Shannon C. E., 1948). Compressive sensing (CS) is one of the many data compression algorithms exhibiting approach to data compression more radically than most of its counterparts (Candès, 2006).

Generally, in data compression algorithms, a N length signal x in R^N space is transformed into a linear sum of its basis functions $\psi_i\ (i = 1,2, \ldots \ldots . N)$ and scaling vector, s as in Equation 3.15. The signal x considered K-sparse if at most, it contains K-nonzero coefficients and $(N - K)$ zero coefficients in s. The case of interest in compressive sensing is when, $K \ll N$ (Baraniuk, 2007; Eleyan, et al, 2014). CS theory addressed three inefficiencies with the conventional compression techniques. Firstly, for a large number of samples N the compression techniques must be applied to all samples

even though desired K is small. Secondly, the set of all N transform coefficients $\{s_i\}$, must be computed despite the fact that only $(N-K)$ samples will be retained. Thirdly, the positions of the bigger coefficients must be encoded, thereby adding up overheads (Candès, 2006).

$$x = \sum_{i=1}^{N} s_i \psi_i \tag{3.16}$$

CS considers a holistic linear measurements process that compute $M < N$ inner products between x. The measurement y of x can be obtained from a $M \times N$ measurement matrix $\Phi = [\phi_1, \phi_2, \phi_3, \ldots\ldots\ldots \phi_N]$, with column vectors ϕ_i $(i = 1, 2, \ldots\ldots M)$. using vector form of Equation 3.16, a new measurement vector y is computed.

$$y = \Phi x = \Phi \Psi s = \Theta s \tag{3.17}$$

Where $\Theta = \Phi\Psi$ and co-efficient vector s can be obtained $s = \Theta^T y$. The measurement matrix has the property that enables the reconstruction of the signal x of length N from $M < N$ measurements satisfying Equation 3.18 for any arbitrary vector v with 3K-sparse. This above condition is referred to as restricted isometry property (RIP) (Candès, 2006).

$$1 - \epsilon \leq \frac{||\Theta v||_2}{||v||_2} \leq 1 + \epsilon \tag{3.18}$$

v is any vector that shares same non-zero entries as s and for some $\epsilon > 0$. The authors in (Candès, 2006) provided a proof of the measurement matrix that satisfy (3.18). They indicated that such matrix can be generated with high probability. For example, Gaussian probability density function generated from independent identically distributed (i.i.d) random variables with mean zero and variance 1/N satisfies RIP.

3.4.Classification Algorithms

Classifier algorithms play vital role in the overall process of facial expression recognition. It is the backbone upon which patterns for a particular expression class is learned for future prediction. The fundamentally provided a framework and a platform upon which the classification process becomes a success. These stages are tasked with

the responsivity for decision making and predictions based on the prior knowledge of a pattern in data, which have been provided during training. These algorithms can be based on supervised or unsupervised learning processes or simply a sort of simple classifier algorithm that does not need training but simply computes some score metrics. Some of these classifier algorithms include; Artificial Neural Network, Support Vector machine, K-Nearest Neighbor, Decision trees, and host of many others.

3.4.1. K-Nearest Neighbor

K Nearest Neighbors classifier, upon known as k-NN with k denoting the neighborhood size, is a non-parametric approach used for regression and data classification. In both cases, the input to the algorithm consists of the k number of closest training samples to the test vector to be classified. The output of the algorithm depends on its application whether for regression or classification. In classification, majority vote rule is use to decide a class to which an object belongs, implying that object is assigned among its k nearest neighbors that is most common. In a situation where k = 1, object is assigned to the class of its nearest neighbor. For regression, the output is the property value for the object which is equivalent to the mean of the values of its k-nearest neighbors.

k-NN can be described as a sort of instance-based learning, or lazy learning, in which the function is only locally approximated and all computations are deferred until classification. The k-NN classification is considered as one of simplest of machine learning classification algorithms. A common practice in, classification and regression, is that nearer neighbor are given more priority in terms of weight than the distance neighbors. For instance, a common weighting scheme allocates each neighbor a weight proportional to its closeness or otherwise to the test data. To compute the distance between a test vector and the training set vectors distance matric algorithms such as Euclidean distance (l_2-*norm*), Manhattan Distance (l_1-*norm*), etc. are frequently used. Figure 3.10, shows a diagram of how K-NN classifier or associate a new test data to a class. Equations (3.19) and (3.20) show how the distance between two *n*-dimensional vectors two vectors x and y, can be computed using Euclidean distance $\delta_{\ell_2}(x, y)$, and Manhattan Distance $\delta_{\ell_1}(x, y)$ metrics, respectively.

$$\delta_{\ell_1}(x, y) = |x - y| \tag{3.19}$$

$$\delta_{\ell_2}(x, y) = \|x - y\|^2 \tag{3.20}$$

For instance, in Figure 3.10, two distinct classes are represented by the red triangle and blue square while the test data is represented by the green circle whose class is to be determined. Initially distance matric algorithm is used to compute the position of all the training sets (both class 1 and 2) relative to the unknown test data (green circle), then the k neighbors are selected to determine the class to which test data belongs. If k=3 (i.e. 3 nearest neighbors are considered), the test data will belong to red triangle class because it has majority of its 3 neighbors coming from that class. However, if k=5 (i.e. 5 nearest neighbors are considered), the test data will be assigned to blue square class.

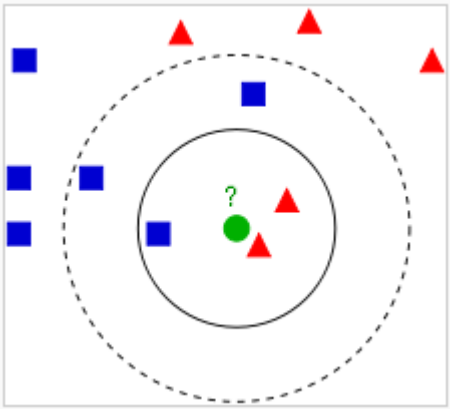


Figure 3. 10 K-Nearest Neighbor classifier

3.4.2. Support Vector Machine (SVM)

In machine learning and pattern recognition SVM is frequently use for regression and classifications. Support Vector Machine belongs to the supervised learning algorithm. In classification problems where the aim is to separate a given dataset a number of classes, SVM fundamentally in an $n - dimensional$ space, it creates $(n - 1)$ hyper-planes which are used to classify the datasets. In the initial stage of SVM, the algorithm was only capable of linear classifications using maximum-margin hyper-plane concept. In 1992, (Boser, Guyon, & Vapnik, 1992) proposed a concept of kernel tricks to modify the previous version of SVM that uses maximum margin hyper-planes for data classification. The kernel tricks make it now possible for SVM to train and

classify non-linear data. The dot *product* used in Maximum Margin SVM was replaced by a kernel function which can match non-liner dataset. For a vector x with existing in an $n-dimensional$ space whose class falls under A or B, if y is a level or output representing the class which x is assigned, then SVM will have a sequence $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots (x_n, y_n)$ as its input. In a binary classification the class level y will have only two classes and hence all the class level sequence $(y_1), (y_2), \ldots\ldots\ldots (, y_n)$ must have one of the two levels as in Equation 3.21 (Boser, et al, 1992).

$$\begin{cases} y_k = 1, & if\ x_k \in class\ A \\ y_k = -1, & if\ x_k \in class\ B \end{cases} \tag{3.21}$$

The decision function $D(x)$ is determined in accordance with the rule in (3.22).

$$\begin{cases} x \in A, & if\ D(x) > 0 \\ x \in B, & otherwise \end{cases} \tag{3.22}$$

The authors in (Boser, et al, 1992) proved that the decision function $D(x)$ can train both linear and non-linear dataset and it does not depend on the linearity properties of the dataset but only needed to have linear parameters. The decision function now becomes the kernel function and can transform a non-linear dataset into a linear $n-dimensional$ space in conformity with basic operational principles of SVM. The decision function has two forms; direct form in Equation 3.23 and the dual form in Equation 3.24 space (Boser, et al, 1992).

$$D(x) = \sum_{k=1}^{n} \omega_i \varphi_i(x) + b \tag{3.23}$$

$$D(x) = \sum_{k=1}^{n} \alpha_k k(x_k, x) + b \tag{3.24}$$

Where α , ω and *b* are the parameters that need to be tuned in the training phase. Functions $\varphi_i(x)$ direct kernel function and $k(x_k, x)$ is the dual kernel function. Numerous kernel functions have been used and found fitting in the maximum margin SVM satisfying the training rules (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002). Some of the kernel functions use in SVM are; polynomial, perceptron, RBF and arc-

tangent functions. With direct space kernels function $\varphi_i(x)$, and assuming that a margin M between class boundary A and B exists, the training rules will aim at maximizing the hyperplane margin such that for a function output, M (Equation 3.25):

$$\frac{y_k D(x_k)}{\|\omega\|} \geq M \tag{3.25}$$

The generalization error for such classifier is linked to the error on both error on the training data classifier and complexity (Boser, et al, 1992; Vapnik, 1991). Trade-off exists between higher capacity classifiers (i.e. with large number of tunable parameters) and low capacity classifiers (i.e. within sufficient number of adjustable parameters). Low capacity classifiers might not be able to learn the task at all but when they do they exhibit good generalization due to their low complexity. On the other hand, higher capacity classifiers can learn any classification pattern according to the learning rules without error, but they generally tend to exhibit poor generalization. However, a good generalization performance can be achieved with higher capacity classifiers by matching the capacity of the classification function with the size of the training data (Boser, et al, 1992; Vapnik , 1991).

3.4.2.1.Multi-class Support Vector Machine

In the early days of machine learning, most of the classification algorithms were used for binary classification e.g. binary SVM and perceptron learning (Lin & Hsu, 2002). However, in real-life multi-classification problems are very common in most of the nowadays applications. Over times, researchers have proposed number of multi-classification algorithms by either combining binary classifiers or using an optimization approach to solve non-binary classification. Research in this field is still ongoing. For instance, SVM multiclass learning approach is mostly executed in two ways. The first approach constructs a predetermined number of binary SVM and then they are combined prediction of the multiclass classification. Popular amongst these algorithms are*: Decision Directed Acyclic Graph (DDAG)*, *one-versus-one (1-v-1)*, *one-versus-rest (1-v-r)* and *Error Correcting Output Code (ECOC)* (Lin & Hsu, 2002). The second approach considers the whole dataset as one optimization problem and try to solve the problem in one single step e.g. “Crammer and Singer” approach (Lin & Hsu, 2002). However, this method is very difficult and computationally expensive to solve due to its

complexity and is hardly used in practical SVM multi-classification problems (Lin & Hsu, 2002).

I. *one-versus-rest.*

In *one-versus-rest* as it's popularly referred to, This approach construct *K* binary SVM classifier models for *K* number of class level. It then assigns a positive class level to the training vectors in a particular class and assigns a negative class level to all other training vectors that does not belong to that class. During prediction, a testing vector x is tested in all the *K* binary SVMs and the output of their decision function $D(x)$ (3) is computed. The classifier model which produces the largest function output is assumed to be the class which the input vector x belongs (Vapnik V. , 1998).

II. one-versus-one

Here, binary SVM models are constructed equivalent to the combination $K!/(K-2)!\,2!$. Hence $N = K(K-1)/2$ binary classifier models are required to be constructed (Knerr, et al, 1990). In the training phase, N classifier models are trained which corresponds to all possible combinations of the number of classes K, taking two at a time. During the training, in each combination of two classes, one class is assigned a positive label and the other negative label. In the prediction phase, the original authors (Knerr, et al, 1990) proposed combining the classifiers with an 'AND' gate to decide the winner. However, maximum voting approach proposed in (Friedman, 1996) is commonly used to decide the winning class. In maximum voting approach, all the *K* classes start with zero votes and an input vector x is tested against all the N binary SVM models. If for instance output of N^{th} binary SVM matches K^{th} class label which was assigned to K^{th} class members during training, the vote of K^{th} class is increased by 1. At the end, class that receives maximum votes is declared the winner. This classification method introduces an overhead of number of binary classifiers required to train compared to *one-versus-rest* though the time required to train each model is much less.

III. Decision Directed Acyclic Graph (DDAG)

DDAG training algorithm has the same approach with *one-versus-one* during training but only differ in the prediction phase. Instead of adopting maximum voting approach as described (Friedman, 1996), directed acyclic graph was proposed in (John, et al, 2000) which have different approach and shorter way in determining the winner. The author described *DAG* as decision tree whereby the N binary classifiers constitutes

the number of nodes on the tree and only K nodes are required to be traversed to determine the winner. Each node in DDAG is called a *decision node* and the path that leads to the solution is referred as *evaluation path*. Consider a classification containing 4 classes in Figure 3.11, the first *decision node* is called the *root node* and it contains the list of all the classes arrange serially, in increasing order. Each *decision node* is a binary classier constructed from the first and last class on its list. During prediction, an input vector x, initially begins with the root node which according to the rule is made from class 1 and 4 (i.e. first and last class on its list). The decision of from where to exit the node depends on the $D(x)$ output. If for instance the decision output favors class 1, the node is exited from the right and class 4 is eliminated from the contest and the next decision node will have its list with class 4 missing. The algorithm proceeds with elimination until only one class remains. Therefore, for a problem with K classes, only $K-1$ *decision nodes* will be evaluated to find the winner (John, et al, 2000) .

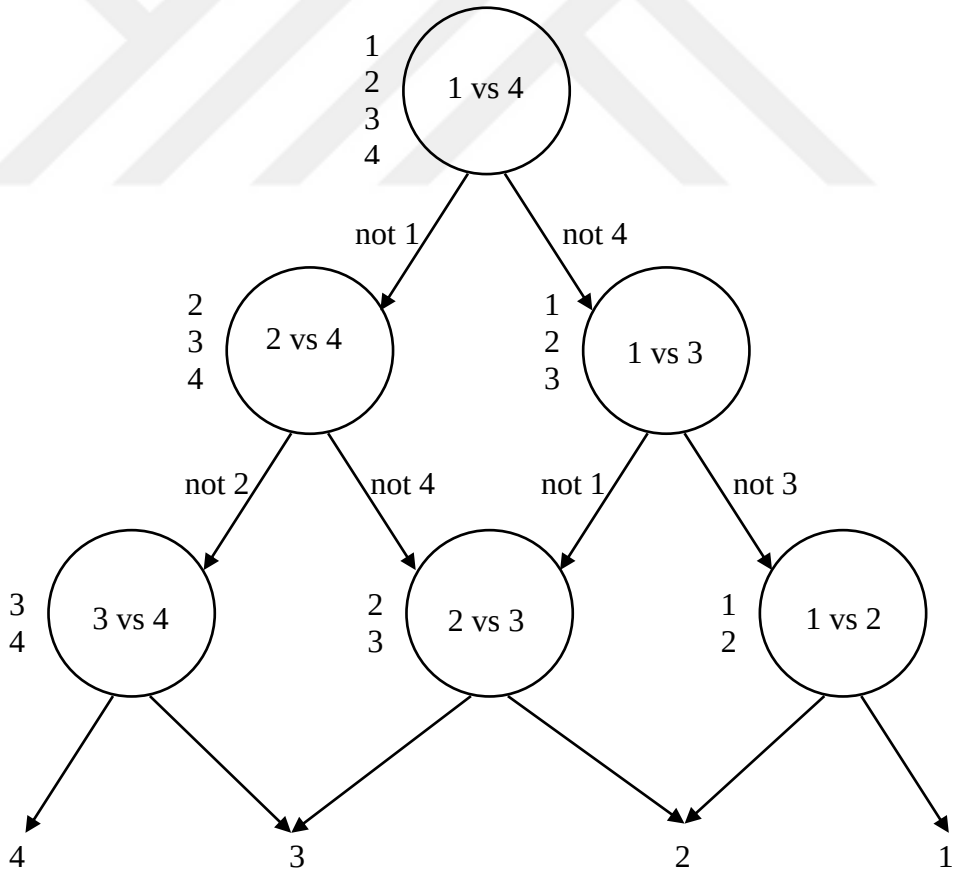


Figure 3. 11 Decision Directed Acyclic Graph for 4 claçsses problem
(John, et al, 2000)

IV. Error Correcting Output Code (ECOC)

ECOC like its counterparts above has two phases: the training phase and prediction phase. In *ECOC* a coding scheme is deployed to encode all the binary classifier models created with the available number of classes. For example, a *coding matrix*, $\mathbf{M} \in \{-1, 0, 1\}^{Q \times S}$ is encoded such that the row of the matrix represents a unique "codeword" for each class. The matrix M, has Q row equivalent to total number of classes and S columns corresponding to the total number of binary classifier models to be trained. The authors (Dietterich & Bakiri, 1995) of ECOC restricted the eateries of the *coding matrix* to only binary numbers. During the prediction, a new "codeword" of length S is generated when test vector x is passed through S dichotomies. Hamming distance is use to decode the "codeword" between each row of the coding matrix. The row with least hamming distance is assumed to be the class to which vector x belongs. Table 3.3 shows coding scheme of 4-class problem. The dichotomies are created based on *one-versus-one* training approach. Entry 1 indicates that the corresponding class members have been used in the binary classifier with positive class level 1, whereas as -1 indicates class members have been used with negative class level. The entry 0 indicates that the corresponding class is not being used for training in that dichotomy. Note that the original *ECOC* coding (John, Nello, & John, 2000) includes only binary entries. Table 1 is only used to explain the basic coding scheme of the *ECOC* concept as many variants of original *ECOC* have been proposed by other researchers (Zhanyi, et al, 2010).

Table 3. 3 Binary classifiers combination

	Dichotomies or binary classifier combination					
Classes	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
1	1	1	1	0	0	0
2	-1	0	0	1	1	0
3	0	-1	0	-1	0	1
4	0	0	-1	0	-1	-1

3.4.2.2. SVM kernel parameter Selection

Formulations and empirical results from many researchers have indicated that the key to good generalization performance to the SVM and by extension learning algorithms, is the proper selection of kernel for a given classification task at hand. Moreover, the kernel generalization performance is influenced by its parameters (Chapelle, et al, 2002). Selection of proper kernel parameters for a given classification task becomes very difficult and tasking especially for kernels with many parameters (Chapelle, et al, 2002). This difficulty limits the use of many kernels in practice (Chapelle, et al, 2002). Parameter tuning is an optimization problem which gears at finding kernel parameters that will minimize the classifier's generalization error. One of the strategies used is the exhaustive search over all possible values within the parameter space (unbounded) that will result in a least generalization error. The strategy is naïve since it requires the algorithm to be run on every possible value of the parameter. Authors in (Chapelle, et al, 2002) proposed an automatic approach for choosing multiple parameters in SVM. They used the standard gradient decent approach to estimates the parameters and claimed that the approach can optimize up to 100 parameters with reasonable performance in terms of the generalization error.

3.4.2.3. Radial Basis Function

D. S. Broomhead and D. Lowe (Broomhead & Lowe, 1988) were the first to use RBF as a kernel function in Radial Basis Neural Networks. They observed that the kernel function is analogous to curve-fitting in a higher dimensional space of the most feed-forward networks. Their work revealed that learning was similar to generating a best fit surface to a finite set of data points (training data) in higher dimensional space. They also likened generalization as simply act of interpolation of the test data on the fitted surface Lowe (Broomhead & Lowe, 1988; Lowe, 1989). RBF can be any real-valued function, $\phi(x, c)$, whereby its value only depends on the distance from an origin $(c = 0)$, or alternatively depends on the distance from some other point c, called a *center*, such that (Equation 3.26):

$$\phi(x, c) = \phi(\|x - c\|) \tag{3.26}$$

The $\ell_2 - norm$ operator $\|.\|$, is usually Euclidean distance, though other distance functions can also be used. Example of popularly used RBF includes; multi-quadratic, Gaussian, Inverse quadratic and inverse multi-quadratic functions (Lin & Hsu, 2002). Equation 3.27 represents a RBF kernel using Gaussian function.

$$\phi(x, c) = e^{-(\frac{\|x-c\|^2}{2\sigma^2})} \tag{3.27}$$

Where sigma σ, is the standard deviation of the kernel which is adjustable.

3.4.3. Artificial Neural Networks (ANN)

The study of Artificial Neural Network (ANN) has been inspired completely by the fact that biological learning system, particularly humans, are built of very complex webs of interconnected neurons. These densely interconnected network of neurons in human brain are capable of solving a very complex problem with precision in a short period of time. The human brain is estimated to contain a densely inter-connected network of approximately 10^{11} neurons. Every neuron is connected, on average, to 10^4 other neurons. The fastest neuron switching is 10^{-3} seconds, quite slow compared to digital computers which is 10^{-10} seconds. Yet biological neural network makes surprisingly complex decision extremely fast which no computer can. The above facts made many researchers to suspect that the information processing capabilities of biological neural systems must follow from highly parallel processes operating on representations that are distributed over many neurons. ANN systems are being inspired by this fascinating ability of biological neurons of parallel computation based on distributed representations. While ANN attempts to model biological neural network, there are many complexities of biological networks that cannot be modelled in ANN.

Artificial Neural networks are mathematical entities that are modeled after existing biological neurons found in the brain. The model is aim at achieving some of features that exist in biological neural networks such as distributed parallel computing, learning capabilities, adaptiveness, fault tolerance, input-output mapping, decision making, non-linearity and so on. ANN are built from a densely interconnected set of simple units (layers), where each unit takes a number of real-valued inputs and produces a single real-valued output.

In general, ANN networks structure consists of Inputs, hidden Layers, Outputs and their Interconnections. The inputs to the NN represents a real-valued vector of a system or a process being investigated. All the elements of the input vector are interconnected to one another via Weight function. Each layer or group of Neurons consist of a summing function, Bias and Activation Function that computes its output. At each neuron within a layer, the Summing function adds the weighed inputs with the computed value of the Bias of that neuron and the result is fed to the activation function to compute the final output of that neuron. The output of all the neurons (which is single real-valued) in the same layer form an input vector to the next layer or becomes the network output if the layer in consideration is the last layer of the network.

3.5. Normalization Schemes

Normalization techniques in most cases have received little or no attention from researchers and in most cases are basically deployed for convenience without minding the underlining effects they can exert on the general statistical distribution data (Ribaric & Fratric, 2006; Nandakumar, et al, 2005). For example, in an application where features from different sources are needed to be fused for reinforcement, normalization may be typically use eliminate disparity in distribution range, but in actual sense its effects do not stops at just bringing data to same range but it does a kind of transformation which may completely alter the distribution. Therefore, careful consideration of a normalization algorithms becomes paramount going by the effect it can have on the overall distribution of the data. To understands the effects of normalization schemes, some commonly used normalization algorithms are examined.

3.5.1. Z-Score Normalization

Z-score normalization is very common in many data normalization applications. It uses two major statistical properties of a data i.e. standard deviation and arithmetic mean. It produces a good normalization means for dataset which are inherently Gaussian in their distribution. One of its shortcomings is lack of robustness attributable to its dependence on the statistical mean and standard deviation which are both susceptible to noise (i.e. outliers) (Ribaric & Fratric, 2006). For a data , Z-score normalization for an input data S_k produces output S'_k, using Equation 3.28.

$$S'_k = \frac{S_k - \mu}{\sigma} \quad , \tag{3.28}$$

where σ and μ are the statistical standard deviation and mean of the distribution respectively.

3.5.2. MIN-MAX Normalization

It has attributes of simplicity amongst the normalization techniques. Its major approach is to bring a dataset into a predetermined interval [0, 1]. This normalization scheme also suffers lack of robustness as the z-score because a noise (outlier) right outside of the normal distribution may takes the minimum or maximum value of the distribution which may result in undesirable results. Min-Max is defined in Equation 3.29.

$$S'_k = \frac{S_k - min}{max - min}, \tag{3.29}$$

The term max represents the maximum entry value in the distribution and min is the minimum entry value.

3.5.3. Median-MAD Normalization

Median-MAD stands for median and median absolute deviation, is considered as one of the normalization schemes which is less sensitive to noise (outliers) and data entry on the fringes of the statistical distribution and hence can been seen as robust. One of its setbacks is that for a non-Gaussian distribution, median and MAD are poorly estimates the location and the scales parameters (Ribaric & Fratric, 2006; imren, ve diğerleri, 2014; Nandakumar, Anil, & Arun, 2005). This effect leads to distortion in the original data a distribution and to transform the data entries into an uncommon numerical range (Ribaric & Fratric, 2006). median-MAD operation is represented by Equation 3.30.

$$S'_k = \frac{S_k - median}{MAD}, \tag{3.30}$$

The median stands for the statistical median of the distribution and MAD represents Median Absolute Deviation from the *median* class*;* $MAD = median(|S_k - median|)$.

3.5.4. Tangent-hyperbolic (Tanh) Normalization

This normalization technique one of the most efficient type of schemes. It is robust to (noise) outliners has been with success used in variety of normalization schemes (Nandakumar, Anil, & Arun, 2005). The tanh operator has both attributes of robustness and efficiency. Tanh normalization is represented by Equation 3.31.

$$S'_k = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{S_k - \mu_{GH}}{\sigma_{GH}}\right)\right) + 1\right\} \quad , \tag{3.31}$$

where σ_{GH} and μ_{GH} represent the standard deviation and mean estimates, respectively.

So many other types of the normalization schemes have been reported in literature and each may have a particular preference to an application, for instance Decimal Scaling normalization is found effective in logarithmic and Euclidean scales (Ashir & Eleyan, 2017).

4. DESIGN AND ALGORITHMIC FORMULATIONS

This chapter dissects through the proposed novel approaches to FER. It outlines the conceptualization of the mathematical models and algorithmic flows in a more elaborative way. Two keys areas were vigorously scrutinized (feature extraction and Classifiers) and valuable contributions were proposed to address some of the loopholes in these areas for overall improvement in accuracy, simplicity and speed of the Facial Expression Recognition.

4.1. Image Pyramid and Single Branch Decision Tree

This proposed approach consists of a feature extraction technique based on image pyramid processing (Figure 4.1) and a data classification technique for multilevel classification problems in FER.

4.1.1 Feature Extraction using Image Pyramid and CS

Here the feature extraction technique proposed is aimed to combining the properties multi-resolution algorithms with robustness, simplicity and data compression and reconstruction properties of compressive sensing approach. This is realized in steps outlined below:

1. *Preprocessing*: databases of facial expression like any other image database is liable to distortion and noise in the process of image acquisition. Apart from the inherent noise in the facial expression data, most of the sample images do consist of background and non-facial information. For fidelity and correctness, this non-facial information is cropped up from the rest of the images to ensure that only correct representation is used. In this work, those images were manually separated (cropped up) from the non-facial part appropriately resized.
2. *Pyramid feature extraction*: For a 2-dimensional sample image *I*, of size $M \times N$. The sample *I*, I is initially converted to a column vector x of size $Q \times 1$, and then normalized in the interval [0 1]. To compute a pyramid feature vector y_J with a predetermined pyramid level *J*, the base of the feature vector of the pyramid can be estimated using Equation 4.1

$$y_J = x. \tag{4.1}$$

The remaining pyramid feature vectors at the succeeding levels $J-1, J-2, \ldots\ldots 1$ (i.e. $y_{J-1}, y_{J-2}, \ldots\ldots y_1$) can be computed as using Equation 4.2. Firstly, a level matrix $G_{J-i}(m,n)$ is computed for $i = 1,2, \ldots . J-1$ such that:

$$G_{J-i}(m,n) = Y_{J-k} * k(m,n), \tag{4.2}$$

The term Y_{J-k}, is a 2-dimensional matrix of feature vector y_{J-k}. Where $k = i-1$ and $*$ is a convolution operator. For level $J-i$, its feature vector matrix at computed using Equation 4.3.

$$Y_{J-i} = f_{2\downarrow}(G_{J-i}(m,n)) \tag{4.3}$$

The term $f_{2\downarrow}$, represents down sampling operation for rows *m* and columns *n* of the level matrix $G_{J-i}(m,n)$. The down sampling operation is given as $f_{2\downarrow}(n) = f(2n)$.

Final extracted pyramid feature y_p , is formed by concatenating the *J* number of pyramid features as in Equation 4.4.

$$y_p = [y_J, y_{J-1}, y_{J-2} \ldots\ldots y_1] \quad . \tag{4.4}$$

3. Scaling and Outliers Suppression:

Outliers in data and artifacts are very much a common possibility during computation such as convolution operation, to mitigate this effect and others like improper scaling and dimensionality curse in the feature vector we proposed arithmetic mean difference. The noise-suppressed extracted feature vector y'_p ,is computed by taking the arithmetic mean difference of a vector y_p as contained in Equation 4.5.

$$y'_{p_i} = y_{p_i} - \frac{1}{Q}\sum_{i=1}^{Q} y_{p_i} \qquad (i = 1,2, \ldots\ldots\ldots\ldots, Q) \tag{4.5}$$

4. *CS feature extraction and dimensionality reduction:*

The noise-suppressed pyramid feature y'_p is subjected to additional reinforcement and reduced dimensionally wise using Compressive Sensing techniques as in Equation 4.6, to extract the final feature vector f_v.

$$f_v = \phi . y'_p \ . \tag{4.6}$$

Where ϕ is the measurement matrix which must satisfy a condition for the reconstruction of y'_p from $D \ll P$ measurements for f_v, of size $D \times 1$, and y'_p of size $P \times 1$. For also arbitrary vector v of 3K-sparse restricted isometry property (RIP) in Equation 4.7 (Candès, 2006) must be satisfied.

$$1 - \epsilon \leq \frac{||\Theta v||_2}{||v||_2} \leq 1 + \epsilon \tag{4.7}$$

The term v can be any vector that shares the same non-zero entries as s and for some $\epsilon > 0$.

The author in (Candès, 2006) provided a proof of the measurement matrix that satisfy (3.17). They indicated that such matrix can be generated with high probability. For example, Gaussian probability density function generated from independent identically distributed (i.i.d) random variables with mean zero and variance 1/N satisfies RIP

For the selection of measurement matrix ϕ, a Gaussian probability function with random variables, zero mean and variance 1/P were used having satisfied (4.6). For a ϕ of size $D \times P$, the length D can be freely set without constrain as long as it enables reconstruction of y'_p from f_v. whereas length P is always equals to the length of a feature vector y'_p . Moreover, once the measurement matrix ϕ is generated it is saved and used throughout the entire exercise so that the need to another matrix does not arise. The ability to set freely the size of of ϕ becomes handy and a feature for dimensionality reduction.

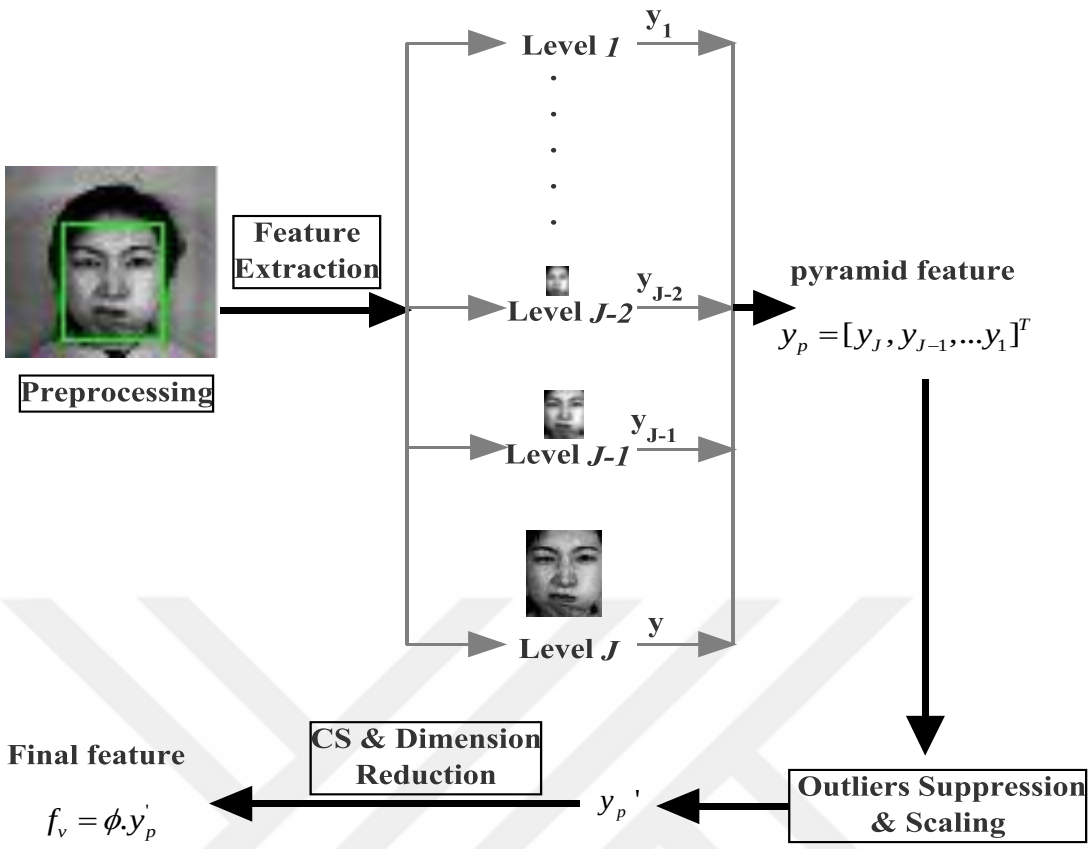


Figure 4. 1 . Proposed Feature Extraction Flowchart
(Ashir & Eleyan, 2017)

4.1.2. Single-branch Decision Tree for multi-class SVM

This is a multiclass classifier with bears resemblance with binary SVM classifier that utilizes *one-versus-rest* classification approach in the training phase.

Consider a set of data with number of classes equivalent to an integer *K*, the proposed multiclassification here train (construct) *K* number of binary SVM models. The *K* binary SVM models are generated by using members one dataset under the same class level as one class and all the remaining datasets outside that class level as tagged under the same class. The selected class level is given a positive class level of 1 and the others are given a negative -1 class level. In this way the multiclass problem is reduced to binary classification. This process of creation binary SVM models is repeated until every distinct class is used against the rest of the classes to generate *K binary SVM models.* In the prediction phase, for a test vector x is evaluated throughout all the trained binary SVM models each time recording the output of the decision function $D(x)$ of each model. The winner is determined via single branch decision tree

(SBDT) (Fig. 4.2) also proposed. SBDT consist of a single branch with *decision nodes* equal to the number of the constructed binary SVM. Each *decision node* represents a unique binary SVM with two leaves on the right and left sides corresponding to the responses of the *node* when test data r x is presented to it. The test Vector x will travel from the *root* of the tree to end node via the *evaluating path*. At every decision node, the output of the trained SVM on that node $D(x)$ is determined, if $D(x) > 0$ the corresponding class on the left side of that node have a scores 1 whereas if $D(x) \leq 0$ the class will have a score 0. When the end of the evaluation is reached, classes with scores 1 are advanced for further evaluation whereas those with 0 scores are eliminated and dropped. the winners are investigated for further thus;

a. Single winner: This occurs when it is only one binary SVM models that have a score of 1. Hence under this scenario the test vector x classified under same class as the original members of that binary SVM model.
b. Multiple winners: This occurs when more than one binary SVM models have scores 1. This is interpreted as a case false positive where more than one binary models claim ownership of test vector x. Under this situation the dispute is resolved by the use of another classifier (Euclidean distance) to make the final decision.
c. No winner: This occurs when none of the binary SVM models have scores of 1. This is a case false rejection where even the rightful class to which test vector x belongs reject its claim. This also portends the failure of the classifier and another classifier (Euclidean distance) to make the final decision.

Euclidean distance measure $\delta_{\ell_2}(x, y)$, for an *n*-dimensional vectors y and x, is computed using Equation 4.8.

$$\delta_{\ell_2}(x, y) = \|x - y\|^2 \tag{4.8}$$

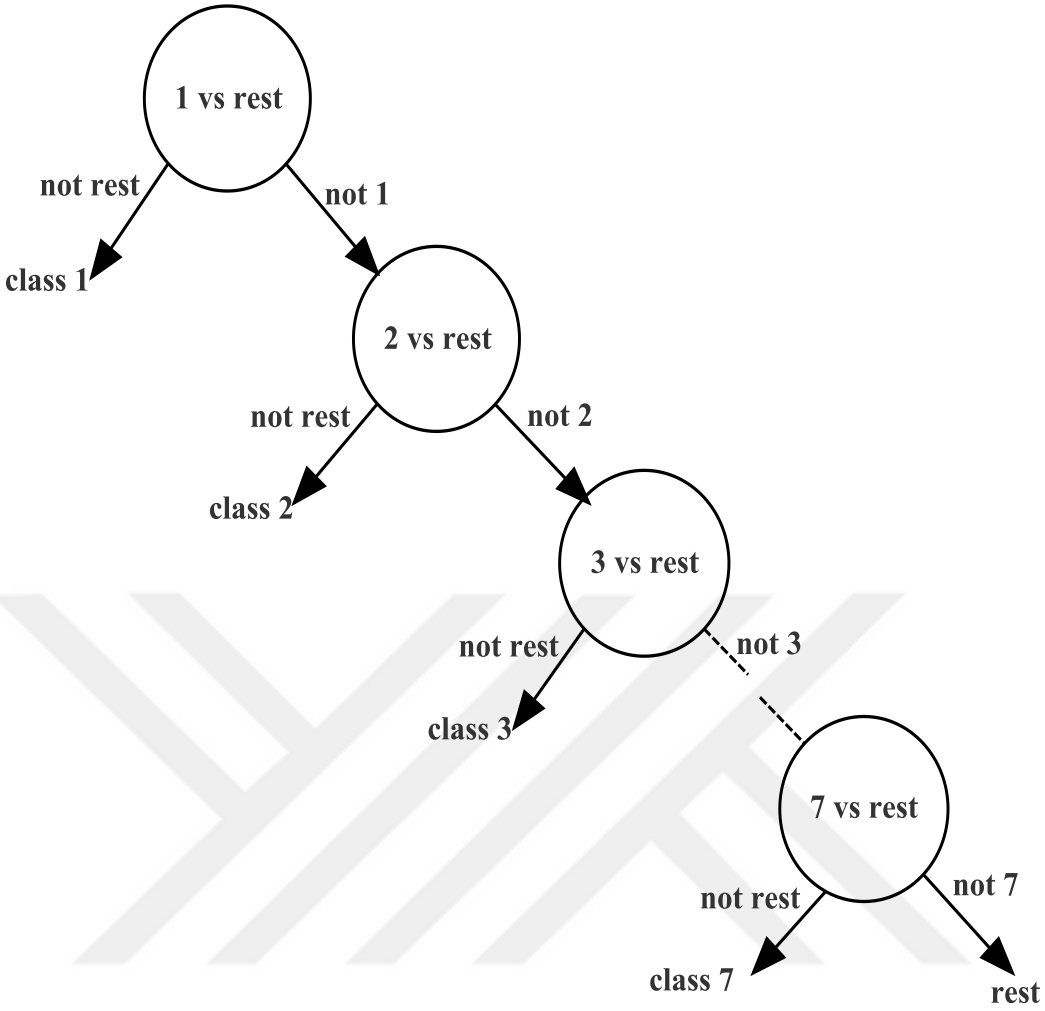


Figure 4. 2 Proposed SBDT.
(Ashir & Eleyan, 2017)

4.2. Arithmetic Mean Difference with Dynamic Cascaded Classification

This noble approach provides an entirely different outlook to confronting facial expression recognition from the one proposed earlier. Fundamentally, it gears towards achieving good performance by minimization of generalization error in the training and classification stage of the process. In an attempt to achieve that both the feature extraction algorithm and classifiers are carefully designed to work in tandem and reinforcing one another.

4.2.1. Feature Extraction with Arithmetic Mean Difference

One of the important attributes of feature extraction algorithm is its ability to represent the original row information in another form (features) while retaining the original data uniqueness. In the proposed approach, arithmetic mean difference (AMD)

and CS were proposed as the feature extractor. Initially a 2D image sample I of size $M \times N$ is vectorized and then normalized within the range [0 1] to generate a column vector x of size $Q \times 1$, where $Q = M \times N$. The AMD vector x^{amd} of the normalized image vector x is computed as follows (Equation 4.9):

$$x_i^{amd} = x_i - \frac{1}{Q}\sum_{i=1}^{Q} x_i \qquad (i = 1,2, \ldots\ldots\ldots, Q) \tag{4.9}$$

The final feature f_v with measurement matrix Φis given

$$f_v = \Phi . x^{amd}. \tag{4.10}$$

4.2.2 Dynamic Cascaded Classifier

The objective of DCC is to form a classifier capable of learning any training patterns with minimum error and with a good generalization capacity. The tradeoffs between training error and generalization error of the learning-based classifiers have been receiving lots of attentions from researchers. As discussed in the preceding section, higher capacity classifiers can learn any training pattern without error but exhibits poor generalization due to their complexity (Boser, et al, 1992). However, it has been shown that, a good generalization performance can be achieved when the capacity of the classification function is matched to the size of the training set or by minimizing the complexity of the classifier (Boser, Guyon, & Vapnik, 1992). Therefore, the proposed DCC addressed the issue of minimizing training error using CS by matching the training sets with the classifier's kernel. The second problem of generalization is addressed by proposing an automatic way of determining the classifier optimum kernel's parameter. Proper tuning of classifiers kernel parameter has been linked to the factor resulting in low generalization error (Vapnik V. , 1991). Based on this, DCC has been designed as a 3-stage classifier built around binary SVMs with RBF as a kernel function. Figure 4.3 is the schematic of the proposed DCC with training data x_{tr} and testing data x_{ts} as inputs.

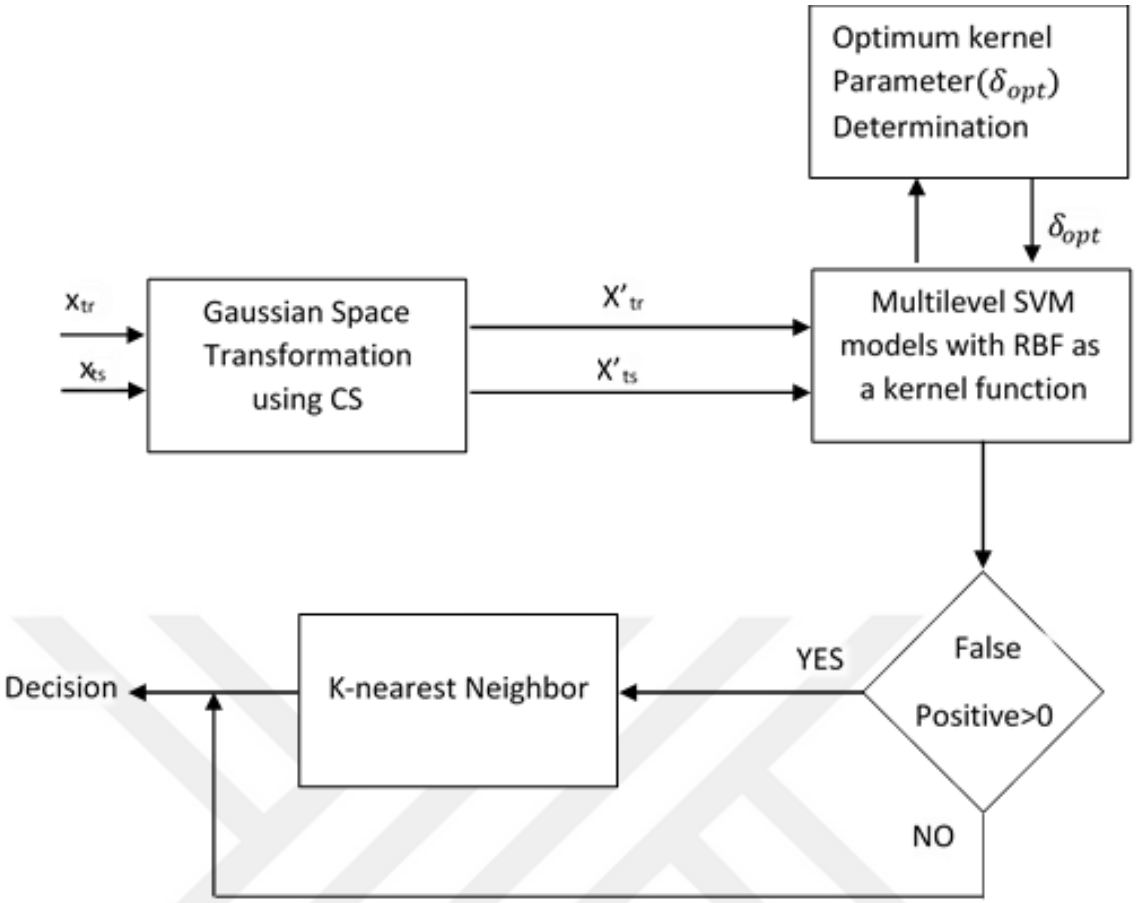


Figure 4. 3 Proposed DCC schematic diagram

4.2.2.1. Stage One

The first stage in the cascade is a prerequisite to the succeeding stages. The aim here is to transform the training vectors into a random variable Gaussian space so that the training sets are matched with the Gaussian RBF kernel. CS with random variables Gaussian measurement matrix will ensures this goal is achieved. Since such kernel basically performs a task synonymous to curve fitting (Lowe, 1989) best fit with least mean square error would be expected in this way. The Gaussian transformation of the training patterns can be realized based on (3.17) as described in CS theory. In accordance with the theory, the measurements preserved the class uniqueness of each input vector x. In addition, in this way a dimensionality reduction is also realized since the measurements can be set to predefined length. For training vector x the Gaussian transformation according to CS will produce a vector x' such that:

$$x' = \Phi x \tag{4.11}$$

4.2.2.2. Stage Two

This is the strongest link in the cascade where maximum margin binary SVM classifier models are created and trained with RBF as kernel function. The numbers of models created are equals to the number of classes in the training sets as in *one-versus-rest* multiclass training. Sequel to this training, an optimum value of RBF sigma (σ_{opt}) is determined from the datasets which have not been used for training. The two key processes in the link are described in the following subsections.

4.2.2.3. Proposed Automatic RBF Kernel Parameter Selection

An important aspect of tackling the recurring of generalization error in many classifiers for an optimum performance is appropriate selection of parameters of the classifier's (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002). The main objective of this part is to provide an automatic means determining an optimal kernel parameter in order to minimally mitigate the effects of generalization error. Since RBF is used as the classifier's kernel, the parameter to be optimized is the mean of the function (σ). In this context, an optimization process using the receiver operating characteristic (ROC) responses of the classifier was proposed. Firstly, a cross-validation with the portion of the datasets that have not been used in the training is used to get the responses of the K constructed binary SVM models.

Assuming K binary SVM models are created based on training dataset, where K is equals to the total number of expression classes in the dataset. Measurement matrixes i.e. False Accept Rate (FAR) and False Reject Rate (FRR) are generated using the cross-validation dataset. For each cross-validation vector x, K scores are recorded being the outputs of K binary SVM decision function $D(x)$. Since each cross-validation vector x belongs to only one 1 binary SVM model, there are $P(p = 1)$ number of genuine claim and $N(n = k - 1)$ number of imposters claims to be made and hence FAR and FRR can be computed for each cross-validation vector x, using Equations 4.12 and 4.13 respectively.

$$FAR = \frac{1}{N}\sum_{i=1}^{N} 1 * (\delta(I_i, T_i) > \gamma) \quad (4.12)$$

$$FRR = \frac{1}{P}\sum_{i=1}^{P} 1 * (\delta(I_i, T_i) < \gamma) \quad (4.13)$$

Where $\delta(I_i, T_i)$is the measurement matrix between the Imput vector I and Template vector T and γ is the hard decision threshold. The matrix $\delta(I_i, T_i)$ is equivalent to the SVM decision function $D(x)$ and the threshold γ is equal to zero as used in binary SVM. Both expressions $(\delta(I_i, T_i) > \gamma)$ and $(\delta(I_i, T_i) < \gamma)$ are assigned logical 1 if true, otherwise are assigned logical 0.

An ideal situation is where both FRR and FAR are equivalent to zero hence the cumulative objective function F, for M cross-validation vectors, is defined by Equation 4.14.

$$F = min \sum_{j=1}^{M} \left\{ \frac{1}{N} \sum_{i=1}^{N} 1 * (\delta(I_i, T_i) > \gamma) + \frac{1}{P} \sum_{i=1}^{P} 1 * (\delta(I_i, T_i) < \gamma) \right\} \quad (4.14)$$

The RBF kernel parameter, *sigma* σ is to be optimized and automatically selected, is arbitrarily initialized and then subsequently updated using Equation 4.15.

$$\sigma_{new} = \sigma_{old} + \Delta\sigma \quad (4.15)$$

Where $\Delta\sigma$ is the incremental step of σ and is determined based on the present and preceding output of the objective function F to ensure fast convergence based on Equation 4.15.

$$\Delta\sigma = 0.1 * (F_{old} - F_{new}) \quad (4.15)$$

σ_{new} is returned as the optimal value optimum sigma (σ_{opt}), in the case of convergence or maximum number of iterations is reached.

Here variant of multiclass SVM proposed in (Ashir & Eleyan, 2017) is adopted. The classification approach bears a resemblance with method used in one-*versus*-*rest* technique in training phase but differs in the prediction the winner is determined via

single branch decision tree (SBDT) (Fig. 4.2) also proposed. SBDT consist of a single branch with *decision nodes* equaling the number of the constructed binary SVM. Each *decision node* represents a unique binary SVM with two leaves on the right and left sides corresponding to the output responses of the *node* when test data x is presented to it. The test data x will travel from the *root* of the tree to end node via the *evaluating path*. At every decision node, the responses of the trained SVM on that node $D(x)$ is determined, if $D(x) > 0$ the corresponding class on the left side of that node have a scores 1 whereas if $D(x) \leq 0$ the class will have a score 0 (Ashir & Eleyan, 2017).

4.2.2.3. Stage Three

Conceptually, the third stage deals with false positives and false rejection responses of the second stage classifier if any exist. Sequel to this, a simple classifier but with different orientation (i.e. dissimilar approach) to the binary SVMs used in second stage will suffice. Because a feature may still exist of a testing vector x, associating it with its rightful class which the second stage is unable to exploit. To keep the approach simple having gone far through the cascades, k-nearest neighbor based on $\delta_{\ell_2} - norm$ (Euclidean) distance measure has been proposed to be used in this stage. Though other classifiers may work as well. For an *n*-dimensional training vector x_{tr} and testing vector x_{ts}, their Euclidean distance is measured using Equation 4.17.

$$\delta_{\ell_2}(x_{tr}, x_{ts}) = \|x_{tr} - x_{ts}\|^2 \tag{4.17}$$

4.3. Monogenic Local Gabor Binary Patterns (mono-LGBP^{riu2})

This proposed mono-LGBP^{riu2} encapsulated all the vital steps that make difference for good facial expression algorithms. The first major step is the extraction of a multi-resolution feature vector using Gabor Wavelet transform. The feature is extracted at different scales magnitude and orientation of the 2-dimentional Gabor filters. At every combination of orientation and scales of the Gabor filter convolution, the output features are subjected to monogenic LBP operation using Equations 4.19, 4.21 and 4.22. These monogenic LBP features ($LBP_{P,R}^{riu2}, \varphi_c, and\ S_c$) are alter recombined into a single monogenic 3-D texton feature vector whisch is used as the feature vector and refererd as mono-LGBP within the confines of this work.

To investigate further some more avenues for better representation of the feature vector, normalization schemes are applied to the mono-LGBP feature vector to investigate any improvement in performance before classification.

At the classification level, KNN was chosen with $l_{2\text{-}norm}$ as the measures for distance metric. The mono-LGBP feature responses a uniquely when a different normalization scheme is applied to it, these responses are evaluated for classification separately using KNN. Depending on the outcome of the classification with different normalization of mono-LGBP feature vector, two best performing results are fused based on simple *sum rule* at the score level of the classifier. See Figure 4.4.

4.3.1 Feature Extraction

The motivation here is to extract a powerful feature that best represents each individual facial expression class. To achieved this, both multi-scale capability of Gabor wavelets and rotation invariant feature of LBP were exploited. Now consider a gray facial expression sample image I, of size $M \times N$, a Gobor wavelets features are extracted using the 2-Dimensional, convolution operation with the sample image extracted in Equation 4.18.

$$G_{\theta,v}(z) = I * \psi_{\theta,v}(z), \tag{4.18}$$

Where $G_{\theta,v}(z)$ is the results of convolution operation between the input data and the Gabor filter with orientation θ, and scale v.

The next stage is to extract monogenic features from the Gabor Wavelets features. This thought is inspired by monogenic signal processing theory. The main objective is two extracts as many different features as possible from a given vector and these features can recombine to give a more robust representation than using only on feature. Three distinct features are of interest in monogenic-LBP namely; phase information, surface type information, convention LBP features which valuable feature in many applications like texture classification (Zhanyi, Weiran, Jiani, & Jun, 2010). In accordance with the theory, these (phase, φ_c, rotation invariant uniform patterns, LBP^{riu2} and the monogenic curvature tensor Sc using higher order Riesz transforms.) features are collapsed into a single monogenic 3-D texton feature (Zhang *et al*., 2010).

Equations (4.19) -(4.23) represents the formulation for the extraction of the three features.

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{p} s(g_p - f_c) \;\; if \; U(LBP_{p,r}) \leq 2, \\ P+1, \qquad otherwise \end{cases}, \tag{4.19}$$

where;

$$U(LBP_{p,r}) = \left|s(g_{p-1} - g_c) - s(g_o - g_c)\right| + \sum_{p=1}^{P-1}\left|s(g_p - g_c) - s(g_{p-1} - g_c)\right| \tag{4.20}$$

The superscript "riu2" in $LBP_{P,R}^{riu2}$ describe rotation invariant "uniform" patterns LBP where U less than or equals to 2; s is the sign function; g_c represents the center pixel intensity of the local neighborhood and g_p ($p = 0, \ldots, P-1$) is the values of P equally spaced pixels on a circle of radius R.

Phase φ_c, is defined as ;

$$\varphi_c = {\varphi}/{({\pi}/{M})}, \tag{4.21}$$

where $M = 5$.

The last parameter S_c is defined by:

$$S_c = \begin{cases} 0, \; \det(T_e) \leq 0 \\ 1, \; else \end{cases}, \tag{4.22}$$

where det (T_e) is the determinant of the monogenic curvature tensor.

The final feature vector is obtained by the concatenation of the three monogenic feature f_{v-mono}obtained aboved as depicted in Equation (4.20).

$$f_{v-mono} = \{LBP_{P,R}^{riu2}, \varphi_c, S_c\} \tag{4.22}$$

4.3.2 Training and Classification

During the training of the extracted feature vectors, KNN has being used with $l_{2\text{-}norm}$ as the distance measure. With powerful individual representation of the feature vector, KNN can simply do the classification to reduce additional computational cost. The equation used in (3.19) is equally applicable here for the training and classification of the facial feature vectors.

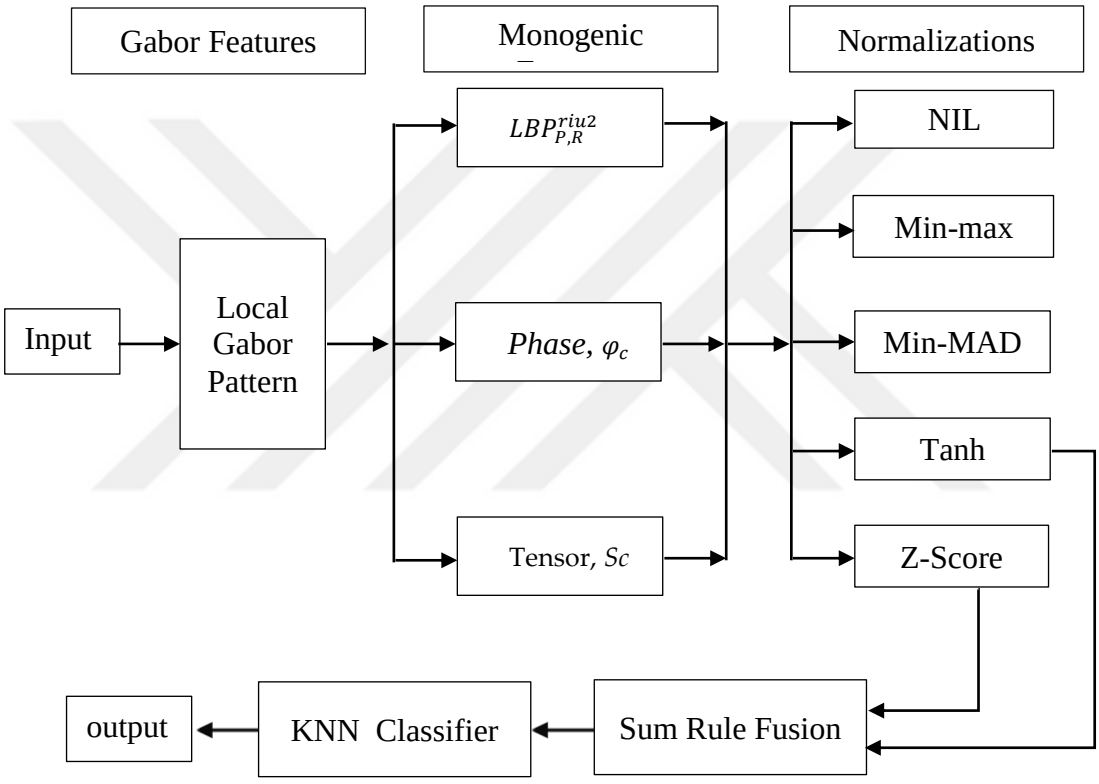


Figure 4. 4 Proposed approach Flowchart using Monogenic
(Ashir & Alaa, 2017)

5. EXPERIMENTAL RESULTS AND DISCUSSION

Under the chapter experimental results and discussion are presented for the three proposed approaches in chapter 4. Each approach has been presented and discussed under a different subsection.

5.1. Results for Image Pyramid and Single Branch Decision Tree

Experiments were conducted using three different facial expression databases and performance results were presented based on the proposed approaches. During the training phase, standard cross-validation methods were adopted throughout. The standard cross-validation methods are *Leave-One-Pose-Out* (LOPO) and *n-fold* cross-validation. With Japanese Female Facial Expression (JAFFE) (Michael, et al, 1998) and Taiwanese Facial Expression Image Database (TFEID) (Li-Fen & Yu-Shiuan, 2007) databases, LOPO cross-validation was used. In Cohn-Kanade (CK) facial expression database (Kanade, et al, 2000), *n-fold* cross validation was deployed as a result of its relatively large size. These cross-validations have become the standards in literatures and they provide meta-training, reduces data over-fitting and ensure fidelity of the experiment (Fasel & Juergen, 2003). Throughout the experiment where SVM was used, radial basis function (RBF) has been the kernel choice. The RBF kernel parameter (σ) was predetermined at different length *N* of the feature vector. Some the parameters used and their values are listed below:

- Size of the Gaussian Filter: 9×9: The filter size selection is guided by the fact that a very small sized filter window (e.g. 3×3) may be too localized concentering on limited local pixels which does not results in improvement in performance and also increases the computational cost of the operation. Large sized filter window (e.g. 16×16) has less computation but it has over smoothing effects where some localized information is not accounted for. 9×9 window is ideal for most experiments as it strikes a balance (Ashir & Eleyan, 2017).
- Mean value of Gaussian $\mu = 0$.
- Standard Deviation of the Gaussian Filter: $\sigma = 20$.
- RBF kernel (σ) parameters for six *N length*: [8 10 11 12 13 14].

5.1.1. Leave-One-Pose-Out cross-validation on JAFFE and TFEID.

JAFFE database consists of 213 images collected from 10 different subjects. Each subject has at least 3 samples and at most 4 sample images per single expression class. To evaluate the proposed method performance 210 out of 213 images in this database were used. Meanwhile TFEID database is made up of facial images collected from 20 different female models. There are 8 expressions in this database including contempt, each image with frontal view inclination of 0^o to 45^o (Li-Fen & Yu-Shiuan, 2007). In the training phase using LOPO cross-validation, samples images were categorized into their primary expressions type classes (i.e. 7 for JAFFE and 8 for TFEID). One image sample at a time is taken from each expression's group as a test sample whereas the remaining image samples are used for training. The procedure is repeated using another sample that has not been used, until all the image samples were individually and uniquely used as testing data.

5.1.2. n-fold cross-validation on CK.

In this cross-validation procedure the whole images in the database were categorized into seven facial expression classes as contained in the database. The categories are further broken into sub-classes. For *n-fold* where n =10, each category will have 10 sub-classes each having an equal number of sample images. With 10 sub-classes for each of the expression category, one sub-class is drawn randomly from each of the category of facial expression to form 10 new groups or *folds* as they are referred to, with each group containing 7 unique sub-classes of facial expression categories. In the training phase the 10-folds are trained in a similar way LOPO cross validation is carried out by picking one-fold at a time for testing and using the remaining 9-folds as training sets. The recognition accuracy is computed as the average of the 10-cross validation.

The original CK database consist of images with 640×490 size including non-facial background on the images. In the context of this work, all facial images were preprocessed to remove the unwanted background information and then resized to 256×256.

Table 5.1 and 5.2 present LOPO cross-validation results conducted on JAFFE database. Two levels (level 1& level 2) of the pyramid was used with a varying feature vector length N, and variants of SVM multiclass classification algorithms. Table 5.3 and 5.4 contains similar results on CK database based on *10-fold* cross-validation. In Table 5.5 and 5.6, LOPO results on TFEID is presented using 8 expression class (including contempt expression). Results from other literatures were compared in Table 5.7 and

5.8 with the results from the proposed method under the same scenario whereas Table 5.9-5.12 contains the confusion matrix results.

Table 5. 1 LOPO Recognition Rate (%) On JAFFE With 1 Level Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	68.11	71.91	71.91	71.91	87.63
200	76.20	77.63	77.15	77.63	88.58
300	79.06	85.25	84.30	85.25	90.49
500	75.25	90.49	90.49	90.49	87.63
700	68.11	87.15	87.63	87.15	89.06
800	67.63	87.63	87.63	87.63	88.11

Table 5. 2 LOPO Recognition Rate (%) On JAFFE With 2 Levels Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	57.63	65.72	67.63	65.72	88.11
200	66.20	76.20	76.21	76.20	88.58
300	69.06	76.20	77.15	76.20	91.44
500	74.77	86.68	86.68	86.68	90.49
700	75.72	85.72	85.72	85.72	89.06
800	74.77	85.72	85.72	85.72	88.11

Table 5. 3 10-Fold Recognition Rate (%) On CK with 1 Level Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	71.92	88.57	86.39	88.57	79.38

200	82.97	91.92	90.67	91.92	91.95
300	77.33	92.85	92.83	92.85	93.07
500	88.14	95.20	94.74	95.20	96.52
700	84.48	96.35	96.49	96.35	92.59
800	90.44	95.54	95.75	95.54	92.81

Table 5. 4 10-Fold Recognition Rate (%) On CK with 2 Levels Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	80.50	92.89	90.50	92.89	93.59
200	87.29	95.50	94.64	95.50	95.87
300	86.08	93.92	95.15	93.92	97.77
500	93.11	97.31	95.87	97.31	97.85
700	93.92	97.77	97.77	97.77	95.67
800	90.92	97.83	97.85	97.83	93.03

Table 5. 5 LOPO Recognition Rate (%) On TFEID with 1 Level Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	79.42	86.77	87.51	86.77	87.51
200	87.51	91.19	91.19	91.19	91.45
300	86.77	91.92	91.92	91.92	89.35
500	86.41	90.82	90.82	90.82	90.45
700	82.73	90.08	90.08	90.08	89.35
800	83.11	88.25	88.25	88.25	89.72

Table 5. 6 LOPO Recognıtıon Rate (%) On TFEID Wıth 2 Level Pyramid.

N	1-vs-rest	1-vs-1	DDAG	ECOC	SBDT
100	77.95	90.45	90.08	90.45	90.45
200	86.04	91.92	91.92	91.92	93.03
300	87.51	93.02	93.02	93.03	92.29
500	87.88	91.92	91.92	91.92	93.39
700	84.94	90.82	91.55	90.82	91.55
800	83.47	91.19	91.55	91.19	90.82

5.1.3. Discussion

One of the major observations worth mentioning is that the proposed approach which includes both the feature extraction and multiclass classification have been able to produce a good result unmatched by most of its counterparts being compared.

The edge which all the multiclass algorithms have over well their counterparts in terms of performance is an attestation of the validity and consistency of the proposed feature extraction approach. It can however be noted that using pyramid level greater than 2 does not further improves the performance of the algorithm hence the results were excluded.

Despite the fact that SBDT classifier matches the best performance of *ECOC* and DDAG with pyramid level 1, as the number of pyramid levels is increased to 2 it exhibits a better performance with even shorter feature vector length compared to ECOC and DDAG. Moreover, in addition to that SBDT less expensive in terms of computation compared to the others excluding ***1-vs-rest.*** The justification for this can be noticed in the training phase of SBDT (similar to ***1-vs-rest*****),** only trains SVM models equals to the number of expression classes there is in the database however other variants train SVM models equals to the combination of the size of expression classes taking two at a time. Similarly, during evaluations, SBDT finds the shortest path to

solution and no additional computations are needed except where false negative or false positive or incidence is encountered.

5.1.4. Comparison

Tables 5.7 and 5.8, comparison results were presented for the of validation of the proposed technique with others in approaches in the literature under the same set-ups. The best results generated from the proposed approach were used in the tabulation of the comparison.

Table 5. 7 Comparison with Different Approaches on Jaffe Database.

Ref.	Feature	Dimensionality	Classifier	Recognition (%)
(Wenfei, et. al. 2012)	Radial Encoded Gabor	cr_LPP	SVM	89.67
(Wei-Lu, et al, 2015)	LPQ +es_LBP-s	PCA	SVM	76.67
(Zilu & Xieyan, 2008)	LBP	GDA	SVM	53.80
(Zilu & Xieyan, 2008)	LBP	Adaboost	SVM	65.70
(Zilu & Xieyan, 2008)	LBP	MV-Boost	SVM	65.71
(Guo & Dye, 2007)	Gabor Histogram	SRC	SVM	58.70
(Huang, Wang, 2010)	LBP	LLE	SVM	62.90
(Cai & Ying, 2009)	CT	N/A	SVM	63.81
(Cai & Ying, 2009)	ASM	N/A	SVM	68.50
(Zavaschi, et al, 2011)	LBP+Gabor	N/A	Ensembles SVMs	70.00
(Zhang, et al, 2011)	Gabor Filter	N/A	Fusion SVMs	72.00
(Zhang Z. et al, 2012)	MB-LGBP	CS	Two level SVMs	74.18
	Proposed		SBDT SVM	91.43

Table 5. 8 Comparison with Different Approaches on CK Database.

References.	Subjects	Classes	Validation	Recognition (%)
(Wenfei, et. al. 2012)	97	7	10-fold	91.51
(Yeasin, et al, 2004)	76	6	5-fold	90.90
(Aleksic & Katsaggelos, 2006)	90	6	10-fold	93.66
(Li, et al, 2010)	97	6	LOPO	96.26

(Shan, et al, 2005)	90	6	10-fold	96.33
(Zhao & Pietikainen, 2007)	96	7(6)	10-fold	88.4 (92.1)
(Bartlett, et al, 2003)	90	7	LOPO	86.90
(Littlewort, et al, 2006)	90	6	–	93.80
(Tian, 2004)	97	6	10-fold	93.80
proposed	97	7	10-fold	97.84

Table 5. 9 Confusion matrix with SBDT on JAFFE database**.**

N=300 Levels=*2*	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	89.00	9.30	0.00	0.00	0.00	9.30	3.30
Disgust	5.80	**90.70**	6.70	0.00	0.00	0.00	0.00
Fear	0.30	0.00	**90.00**	3.30	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**90.70**	5.30	0.00	1.70
Normal	1.30	0.00	0.00	2.30	**94.30**	1.70	0.30
Sad	1.30	0.00	0.00	6.00	0.00	**89.00**	2.30
Surprise	1.30	0.00	3.30	0.00	0.00	0.00	**95.70**

Table 5. 10 Confusion matrix with SBDT on CK database**.**

N=500 Levels=2	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**91.0**	3.3	0.0	0.0	0.0	0.0	0.0
Disgust	3.4	**93.3**	6.7	0.0	0.0	0.0	0.0
Fear	0.0	0.0	**89.7**	1.8	0.0	0.0	3.3
Happy	0.0	0.0	0.0	**94.0**	3.3	4.0	0.0
Neutral	0.0	3.4	0.0	0.0	**96.7**	3.3	0.0
Sad	3.3	0.0	0.3	0.0	0.0	**97.7**	0.0
Surprise	2.3	0.0	3.3	4.2	0.0	0.0	**96.7**

Table 5. 11 Confusion Matrix with DCC on CK Database.

N=500 Level=2	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	97.0	3.3	0.0	0.0	0.0	1.1	0.0
Disgust	1.3	96.3	0.0	0.0	0.0	0.0	0.0
Fear	0.0	0.0	95.4	0.0	0.0	0.0	0.0
Happy	0.0	0.0	0.0	100	0.0	0.0	0.0
Normal	0.0	0.4	0.0	0.0	100	0.0	0.0
Sad	1.7	0.0	3.3	0.0	0.0	92.7	0.0
Surprise	00	0.0	1.3	4.2	0.0	0.0	96.7

5.2. Results for Arithmetic Mean Difference with Dynamic Cascaded Classification

For the evaluation of the performance of this proposed approach, two standard methods for performance evaluation in FER are adopted. These methods include expression specific performance and the general cross-validation results which give the overall performance of the algorithm. Three popular FER databases were used to evaluate the performance of the proposed approach. Extensive experiments were conducted on these FER databases which includes: Japanese Female Facial Expression (JAFFE) (Michael, Shigeru, Miyuki, & Jiro, 1998) and Kohn-Kanade (CK) (Kanade, Cohn, & Tian, 2000) and MMI facial expression database (Valstar & Pantic, 2010). On all databases, identity-independent (II) approach were adopted i.e. all the samples in the database were grouped into their constituents' expression classes (i.e. Sad, Disgust, Fear, Anger, Neural, Happy, and surprise) independently of the person which the sample belong to. This denotes that based on the seven-basic facial expression considered, there are at most 7 possible classes independent of the size of the database. The results obtained from the proposed approach were validated into categories:

(a) Comparisons with existing state-of-the-art SVM multilevel classification algorithms. In both cases, proposed RBF kernel parameter selection is used to select the kernel's optimum parameter (sigma).

(b) Comparison with results obtained by other researchers under the same databases and experimental procedures.

The results presented in Table 5.15-5.17 compared proposed DCC results with other multiclass classification approaches. The other approaches were implemented using the same proposed future extraction approach and the optimized kernel parameter computed with proposed approach for parameter selection in this paper.

5.2.1. Experiments on JAFFE Database.

JAFFE database consists of 213 images collected from 10 different subjects. Each subject has at least 3 samples and at most 4 sample images per single expression class. To evaluate the proposed method performance 210 out of 213 images in this database were used. Meanwhile TFEID database is made up of facial images collected from 20 different female models. There are 8 expressions in this database including contempt, each image with frontal view inclination of 0^o to 45^o (Li-Fen & Yu-Shiuan, 2007). In the training phase using LOPO cross-validation, samples images were categorized into their primary expressions type classes (i.e. 7 for JAFFE and 8 for TFEID). One image sample at a time is taken from each expression's group as a test sample whereas the remaining image samples were used for training. The procedure is repeated using another sample that has not been used, until all the image samples were individually and uniquely used as testing data. Figure 5.1 depicts a cross-section of the JAFFE database. The vertically aligned columns consist of the 7 PI expression classes whereas each row indicates 7 PD expression classes for one distinct subject.

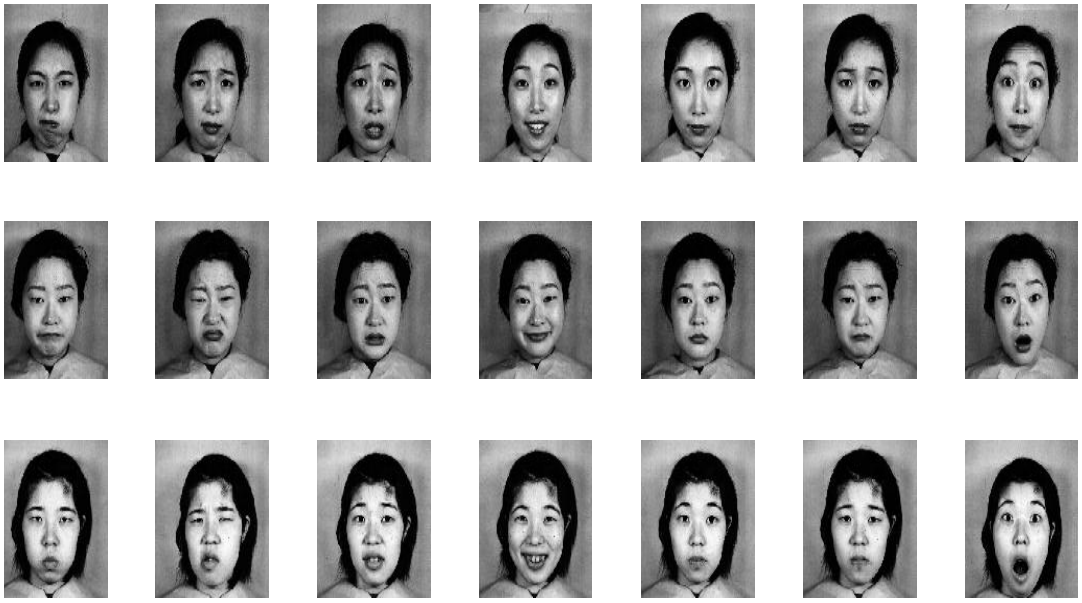


Figure 5. 1.Cross section of JAFFE Database.

5.2.2. Experiments on CK Database.

CK database contains 97 subjects with a total sum of 8795 facial images (Kanade, Cohn, & Tian, 2000). Two standard cross-validation procedures were used here (i.e. *LOPO* and *n-fold* validation). LOPO cross-validation procedure is very exhaustive and intensive computationally (Fasel & Juergen, 2003) and hence, LOPO cross validation was run only once at a fixed feature vector length. For the *n-fold* validation, 10-fold cross-validation method of training was used. In 10-fold cross-validation using II approach, all the samples facial images were firstly grouped into their constituent's classes depending on the expression content. Each of the expression group is divided into 10 sub-classes with each sub-class having equal number of samples. Thereafter, the 10-subclasses from each unique group were recombined (using random selection) with the remaining subclasses from other unique groups to form 10-fold with each fold having all the unique groups samples (i.e. unique expression classes) and the same size. In the training phase, one-fold was used as the testing set while the remaining 9 folds were used as training data. The training was repeated 10 times, each time rotating the fold as a testing data whereas the remaining 9 folds as the training set. The recognition accuracy is computed by taking the average performance from the 10 cross-validation experiments.

The original CK database consist of images with 640×490 size including non-facial background on the images. In the context of this work, all facial sample images were preprocessed to drop the unwanted background information and then resized to 256×256. Moreover, in these experiments neural samples were excluded from the non-neutral facial expression classes. Examples of pre-processed facial images from CK database are presented in Figure 5.2.

Figure 5. 2 Cross section of preprocessed CK Database.

5.2.3. Experiments on MMI Database

The MMI database was compiled from more than 20 students and research staff members' images using both genders (44% female). Their ages ranged between 19 and 62 years (Valstar & Pantic, 2010). For each session, an image sequence is captured with a neutral face at the beginning and the end. During the experiments, 121 samples sequences were selected from 28 subjects. Seven emotion classes were and LOPO cross-validation procedures. In each sequence, the first frames are chosen as the neutral while peak frames were used to a facial expression class. The cross-validation results at different feature vector length are presented in Table 5.12 to 5.14. Whereas Table 5.15 to 5.17 presents the confusion matrix using the proposed approach using 3 different databases.

Table 5. 12 LOPO Cross-validation Results on JAFFE Database Using PI Grouping 1.

Length	Sigma			*1-vs-1*			*1-vs-rest*			*DCC*		
	JF	CK	MM1	JF	CK	MM1	JF	CK	MM1	JF	CK	MM1
100	8.1	4.0	1.2	72.4	91.7	89.3	76.2	97.6	93.3	90.5	97.7	96.1
200	9.8	5.8	2.0	69.5	95.6	90.0	76.2	98.3	93.7	92.9	97.5	96.8
300	11.2	7.5	2.8	71.9	94.0	90.0	78.6	98.0	95.0	92.4	98.3	98.1
400	12.3	7.3	3.4	75.2	96.2	92.3	81.4	98.5	95.4	90.5	99.3	98.6

500	13.4	10.4	4.0	74.8	93.3	91.5	78.1	97.6	95.6	94.3	97.1	98.4
600	14.2	11.9	4.9	79.1	91.9	90.0	82.4	97.7	94.3	92.4	96.5	97.7

Table 5. 13 LOPO Cross-validation Results on JAFFE Database Using PI Grouping 2

Length	Sigma			*DDAG*			*ECOC*			*DCC*		
	JF	CK	MM1	JF	CK	MM1	JF	CK	MM1	JF	CK	MM1
100	8.1	4.0	1.2	75.7	97.0	94.0	77.6	97.7	92.1	90.5	97.7	96.1
200	9.8	5.8	2.0	76.2	98.0	94.0	78.1	98.2	94.5	92.9	97.5	96.8
300	11.2	7.5	2.8	80.0	98.0	94.8	82.4	97.2	95.0	92.4	98.3	98.1
400	12.3	7.3	3.4	80.5	98.1	96.2	80.0	98.3	96.0	90.5	99.3	98.6
500	13.4	10.4	4.0	83.3	97.9	96.3	81.9	98.1	97.3	94.3	97.1	98.4
600	14.2	11.9	4.9	85.7	97.2	96.0	82.4	96.7	95.7	92.4	96.5	97.7

Table 5. 14 LOPO Cross Validation Results on JAFFE Database Using PD Grouping.

100	31.67	57.14	84.76	79.52	84.29	88.10
200	28.33	59.52	84.29	84.76	86.67	88.57
300	41.00	67.14	87.62	86.67	85.71	88.10
400	43.33	68.57	87.14	87.14	84.29	87.62
500	44.67	71.90	87.62	88.10	84.29	90.05
600	44.00	70.62	83.81	84.76	85.24	90.94

5.2.4. Expression Specific Performance Results

Apart from the cross-validation results, one of the most important indicators on how well FER algorithm performs is its performances based on expression modes. Sequel to this fact, confusion matrices have been computed on the three databases and presented in Table 5.15 to 5.18. The confusion matrices are computed where the proposed approach has its best cross-validation results i.e. for JAFFE, the best cross validation result is 94.3% at 500 feature vector length.

Table 5. 15 Confusion Matrix with DCC on JAFFE Database**.**

Length=500	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**90.20**	5.30	0.00	0.00	0.00	7.30	0.00
Disgust	3.40	**94.70**	3.70	0.00	0.00	0.00	0.00
Fear	0.40	0.00	**93.00**	2.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**95.70**	3.30	0.00	0.00
Normal	0.00	0.00	0.00	2.30	**96.30**	1.70	0.00
Sad	3.30	0.00	0.00	0.00	0.40	**91.00**	0.00
Surprise	2.70	0.00	3.30	0.00	0.00	0.00	**100.00**

Table 5. 16 Confusion Matrix with DCC on MMI Database**.**

Length=400	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**97.00**	0.30	0.00	0.00	0.00	0.00	0.00
Disgust	1.40	**98.30**	2.70	0.00	0.00	0.00	0.00
Fear	0.00	0.00	**96.70**	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00
Neutral	0.00	1.40	0.00	0.00	**100.00**	2.30	0.00
Sad	1.30	0.00	0.30	0.00	0.00	**97.70**	0.00
Surprise	0.30	0.00	0.30	0.00	0.00	0.00	**100.00**

Table 5. 17 Confusion Matrix with DCC on CK Database**.**

Length=400	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**99.00**	1.30	0.00	0.00	0.00	1.10	0.00
Disgust	0.70	**98.30**	0.00	0.00	0.00	0.00	0.00
Fear	0.00	0.00	**99.40**	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00
Normal	0.00	0.40	0.00	0.00	**100.00**	0.00	0.00
Sad	0.30	0.00	0.30	0.00	0.00	**98.90**	0.00
Surprise	0.00	0.00	0.30	0.00	0.00	0.00	**100.00**

5.2.5. Results Comparison.

Table 5. 18 Comparison with Different Approaches on JAFFE Database.

Ref.	Feature	Dimensionality	Classifier	Rate (%)
(Wenfei, et. al. 2012)	Radial Encoded Gabor		SVM	89.67
(Wei-Lu, et al, 2015)	LPQ +es_LBP-s	cr_LPP	SVM	76.67
(Ying, et al, 2008)	LBP	PCA	SVM	53.80
(Ying, et al, 2008)	LBP	GDA	SVM	65.70
(Ying, et al, 2008)	LBP	Adaboost	SVM	65.71
(Guo & Dye, 2007)	Gabor Histogram	MV-Boost	SVM	58.70
(Huang, et al, 2010)	LBP	SRC	SVM	62.90
(Cai & Yin, 2009)	CT	LLE	SVM	63.81
(Cai & Yin, 2009)	ASM	N/A	SVM	68.50
(Zavaschi, et al, 2011)	LBP+Gabor	N/A	Ensembles SVMs	70.00
(Zhang &Fang, 2011)	Gabor Filter	N/A	Fusion SVMs	72.00
(Zhang Z. , Xu, 2012)	MB-LGBP	N/A	Two level SVMs	74.18
(Moeini, et al, 2016)	LNBPOA	- (10-fold)	SVM	95.5
(Moeini, et al, 2016)	GEM+LNBPOA	-(10-fold)	SVM	63.0
(Moeini, et al, 2016)	GEM+LNBPTA	-(10 fold)	SVM	62.4
(Moeini, et al, 2016)	LBP(no 3Dfeatures)	-(10-fold)	SVM	81.5
(Rivera , 2013)	LDN CODES	-	SVM	90.1
(Liu, & Ruan, 2012)	TR1DGPA	-	NN classifier	92.2
(Shan, Gong, 2009)	Boosted-LBP	-	SVM	81.0
(Gu, et al,2012)	Gabor Encoded		FLD	89.6
Proposed	Proposed	CS	SBDT SVM	94.3

Table 5. 19 Comparison with Different Approaches on CK Database.

Ref.	Subjects	Classes	Validation	Recognition Rate (%)
(Wenfei, et al, 2011)	97	7	10-fold	91.51
(Moeini, et al, 2016)	-	7	10-fold	97.4
(Moeini, et al, 2016)	-	7	10-fold	50.2
(Moeini, et al, 2016)	-	7	10-fold	89.7
(Rivera, et al, 2013)	96	7	10-fold	96.6
(Shan, Gong, 2009)	-	7	10-fold	88.90
(Gu, et al,2012)	96	7	10-fold	91.5
(Yeasin, et al, 2004)	76	6	5-fold	90.90
(Aleksicet al, 2006)	90	6		93.66

(Shan, et al, 2005)	97	6	10-fold	96.26
(Zhao & Pietika, 2007)	90	6	LOPO	96.33
(Zhao & Pietikain, 2007)	96	7(6)	10-fold	88.4 (92.1)
(Bartlett, et al, 2003)	90	7	10-fold	86.90
(Bartlett, et al, 2003)	90	6	LOPO	93.80
(Littlewort, et al, 2006)	97	6	–	93.80
(Tian, 2004)	96	7	LOPO	94.6
(Rudovic, et al, 2012)	96	6	10-fold	86.8
proposed	97	7	10-fold	97.84

5.2.6. Discussions

In general, a new approach has been proposed which has been effective, competitive and promising based on the experimental results obtained. The experimental results validate the consistency and good performance of this proposed approach over most of the state-of-art approaches compared with. Some novel contributions contained in this paper that could be seen as the backbone to the good performance exhibited by the proposed approach are worthy of being highlighted:

a) AMD and CS feature extraction proposed to be used along DCC has been simple and help to drastically avoid exhaustive computations as is the case in many feature extractions. The proposed feature extraction has worked excellently well alongside the proposed DCC classifier.

b) The concept of using compressive sensing to project the input vectors into random Gaussian space has been instrumental and a prerequisite to ensure that the input vectors, no matter the type, are matched with the RBF kernel of the SVM classifiers. This proposed method has the combined benefits of dimensionality reduction and generally making it possible for the SVM kernel to be able to learn any pattern with minimum error irrespective of the vectors type. By doing so, the approach makes it possible for SVM classifiers to learn even with a training pattern which conventional SVM might have hitherto fail to learn.

c) It could be inferred that, the proposed approach for RBF kernel parameter optimization has been very effective. This could be testified from the fact that, all the multiclass classification algorithms used in the experiments have good performance. Moreover, the search is not a naive one and it can be seen that based on the rules used for searching the parameter it is fast and consistent. However, this proposed automatic parameter selection is limited here to a single parameter

optimization. Multiple parameters selection using the proposed method could be computationally expensive as it may require numerous combinations of the multiple parameters to be optimized.

d) With regards to the third stage of the classifier, records from the experimental results show that it is only invoked in some rare cases. For example, in JAFFE and TFEID databases, stage three was never invoked while in CK and MMI it was invoked in only 1% and 2% of the overall test data.

e) In summary all the approaches proposed together achieved simplicity, good performance, less classifier generalization error, ability to learn pattern and consistency.

5.2.7. DCC Dynamic Responses

Figure 5.3 to 5.11 presents the receiver operating characteristics (ROC) responses of the DCC classifier at different feature vector lengths. These responses are generated from the FAR, FRR and with a changing RBF kernel's parameter based on the proposed technique for optimization of the kernel parameter. From all the three scenarios, it is worth noticing that, the optimum kernel's parameter scales up linearly with the increase feature vectors length. This means that both the search area and the execution time increase. It is unsurprising, how the FRR assumes the shape of the RBF kernel, it is attributable to the fact that all the input vectors got transformed into Gaussian space before training with RBF kernel. It is also noteworthy that the FRR of the DCC in CK (PI) and JAFFE (PI) becomes highly unstable after its best operating point and parameter chosen in that region will result in a very poor generalization performance. Generally, the DCC ROC plots suggest that the proposed approach is always capable of determining the best kernel parameter can always be found within a finite search space.

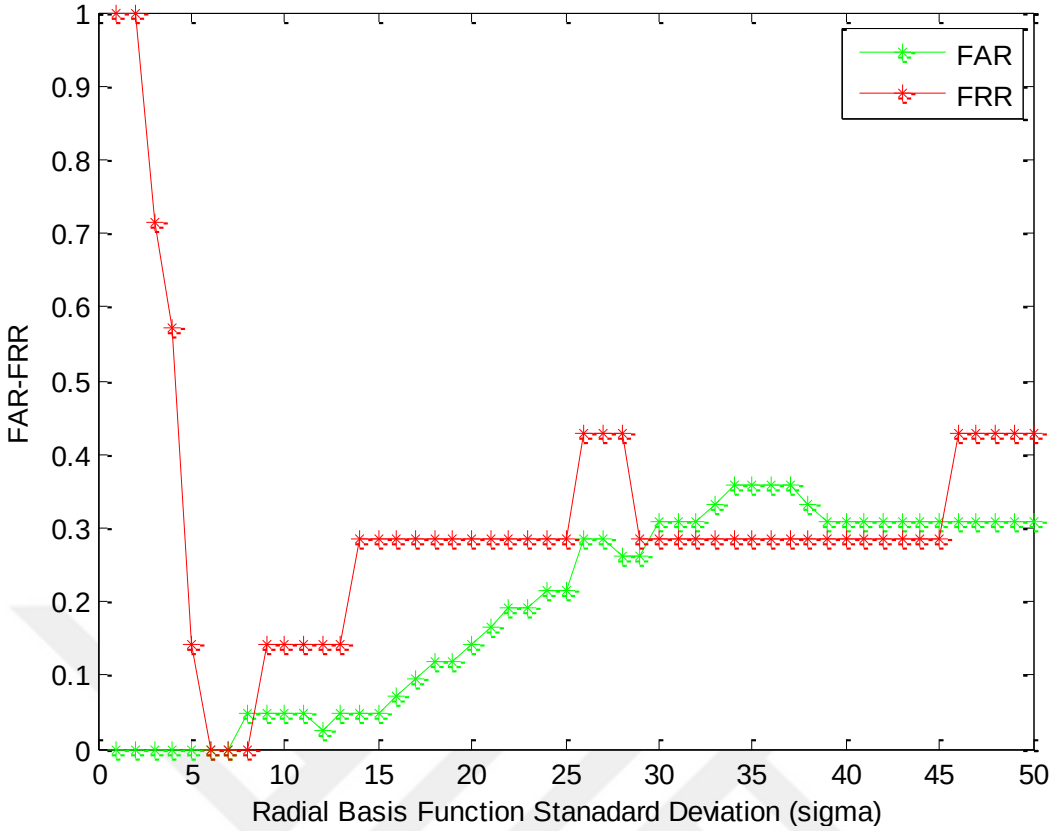


Figure 5. 3 DCC ROC Responses for PI LOPO (JAFFE) feature Length=100

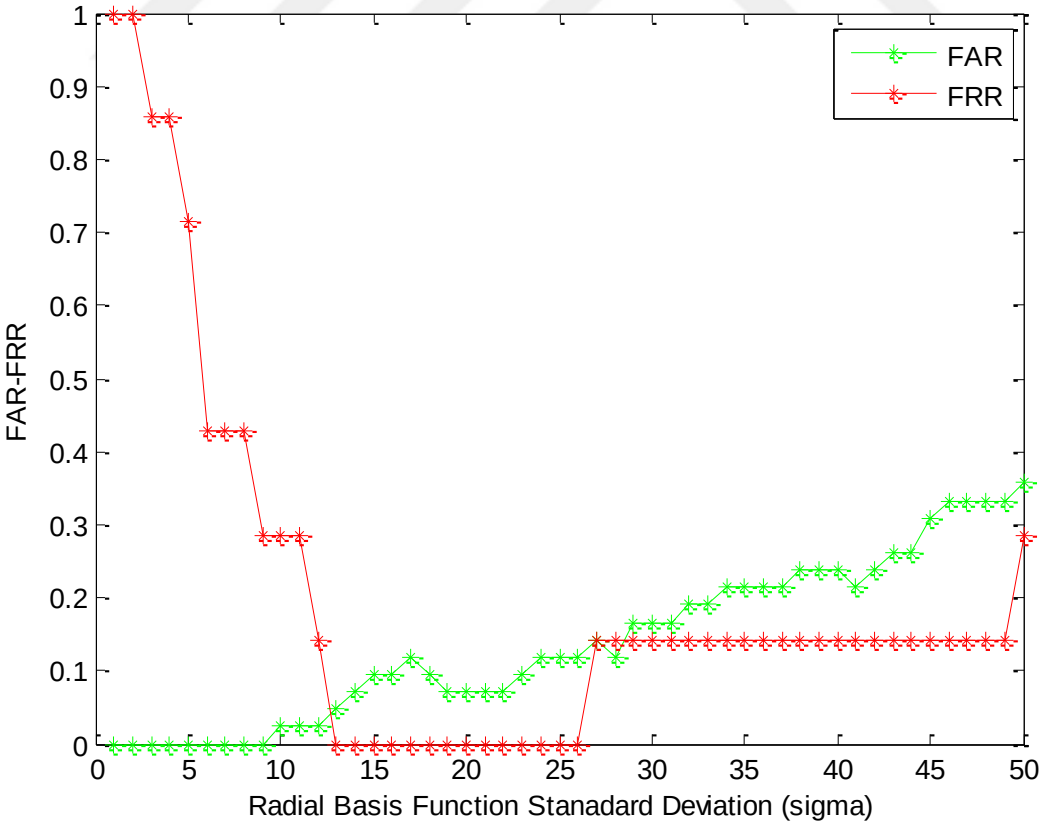


Figure 5. 4 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300

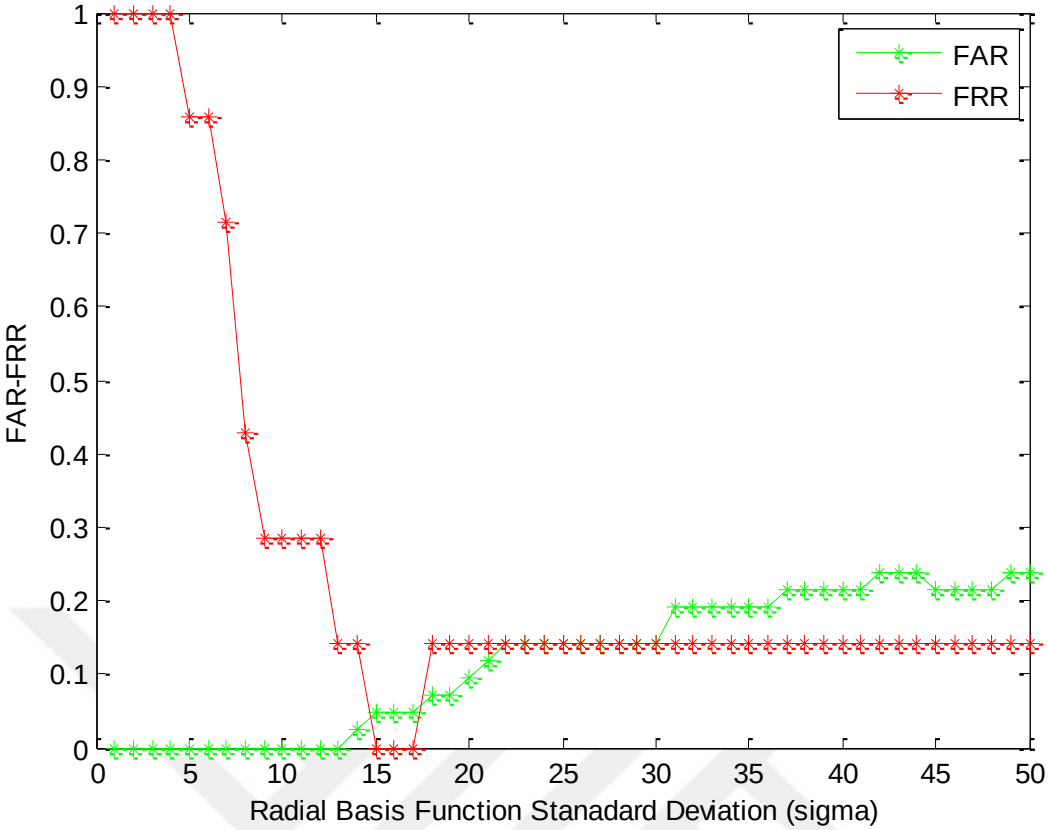


Figure 5. 5 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600

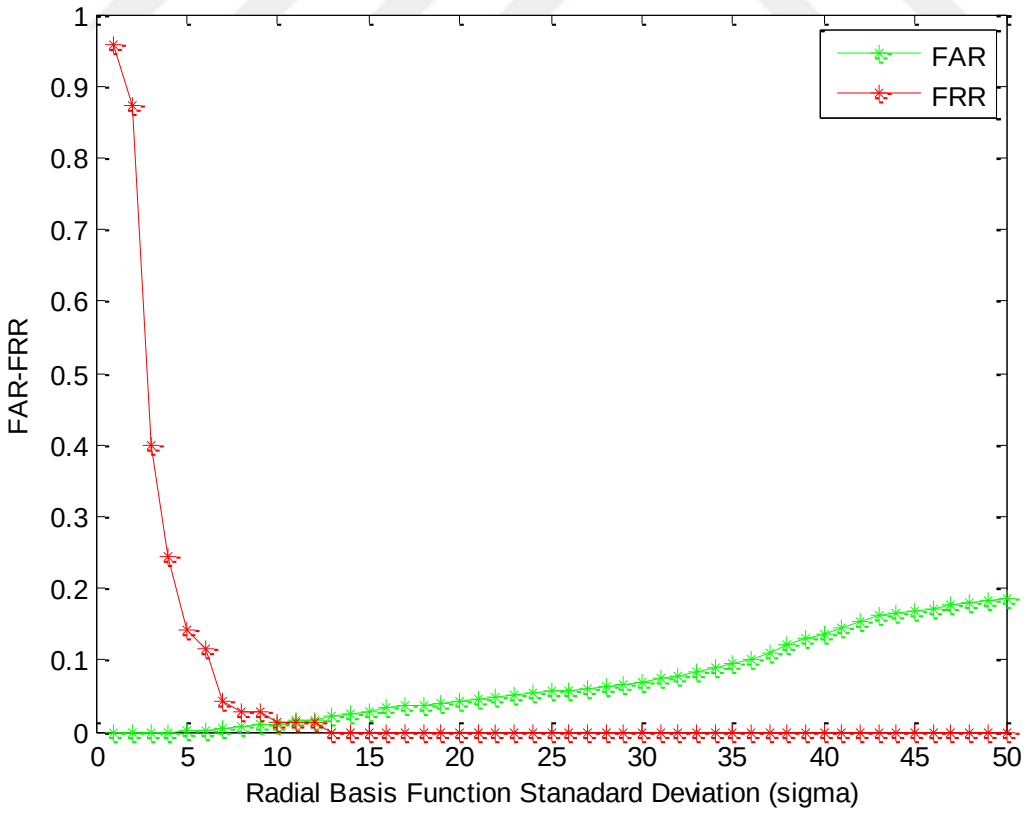


Figure 5. 6 DCC ROC Responses for PD LOPO (JAFFE) feature Length=100

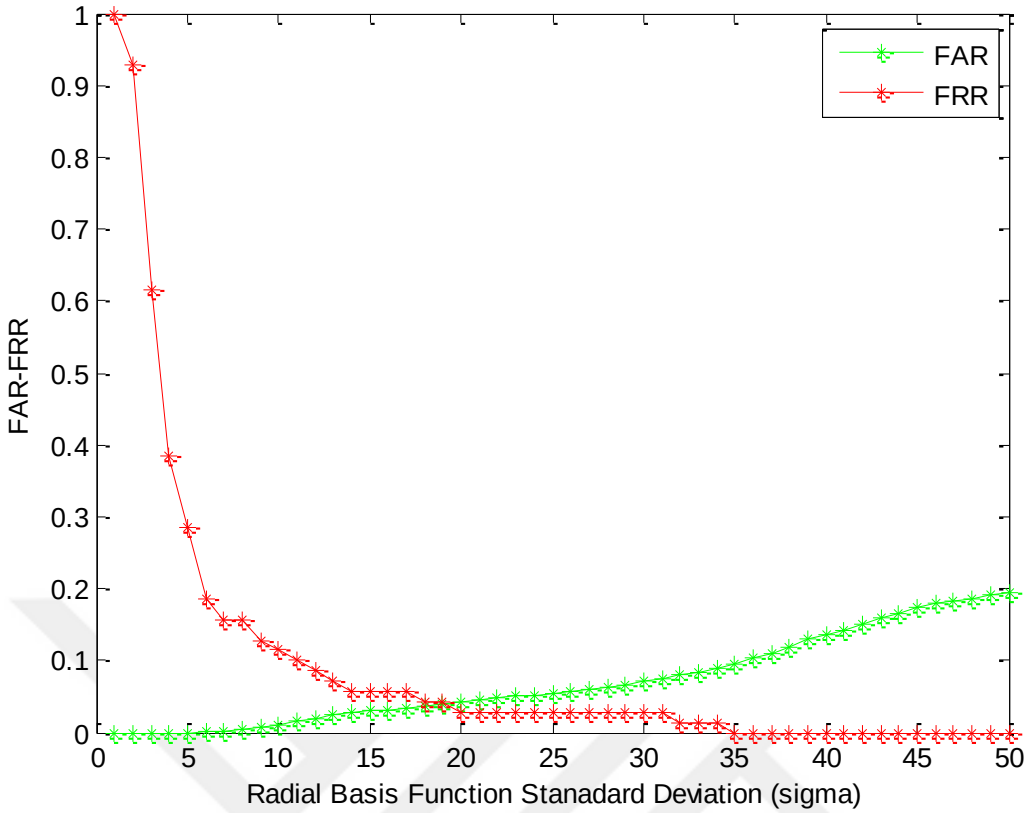


Figure 5. 7 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300

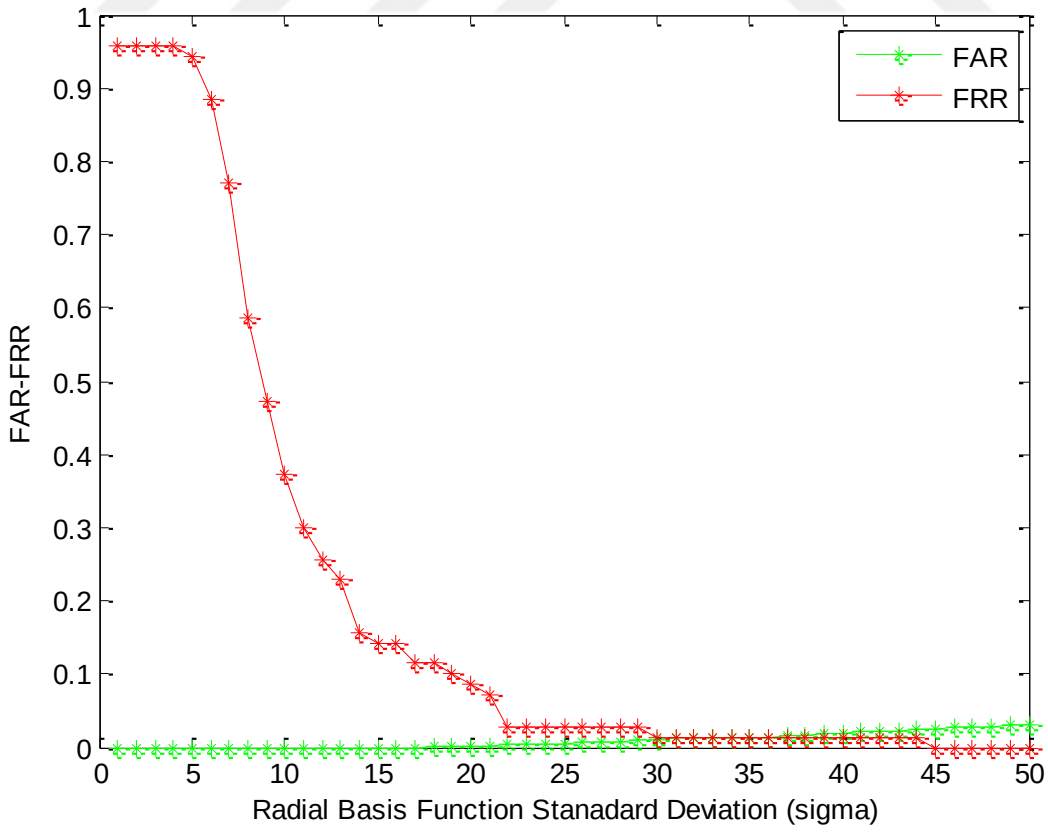


Figure 5. 8 DCC ROC Responses for PI LOPO (JAFFE) feature Length=600

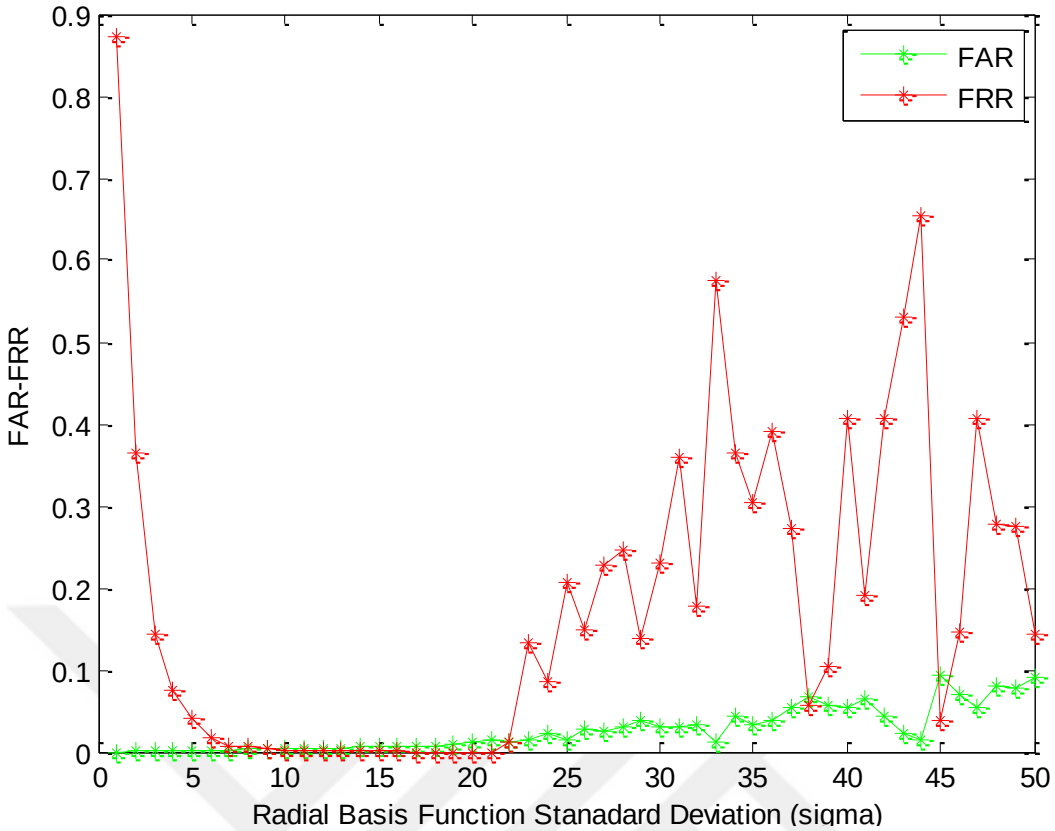


Figure 5. 9 DCC ROC Responses for PI LOPO (CK) feature Length=100

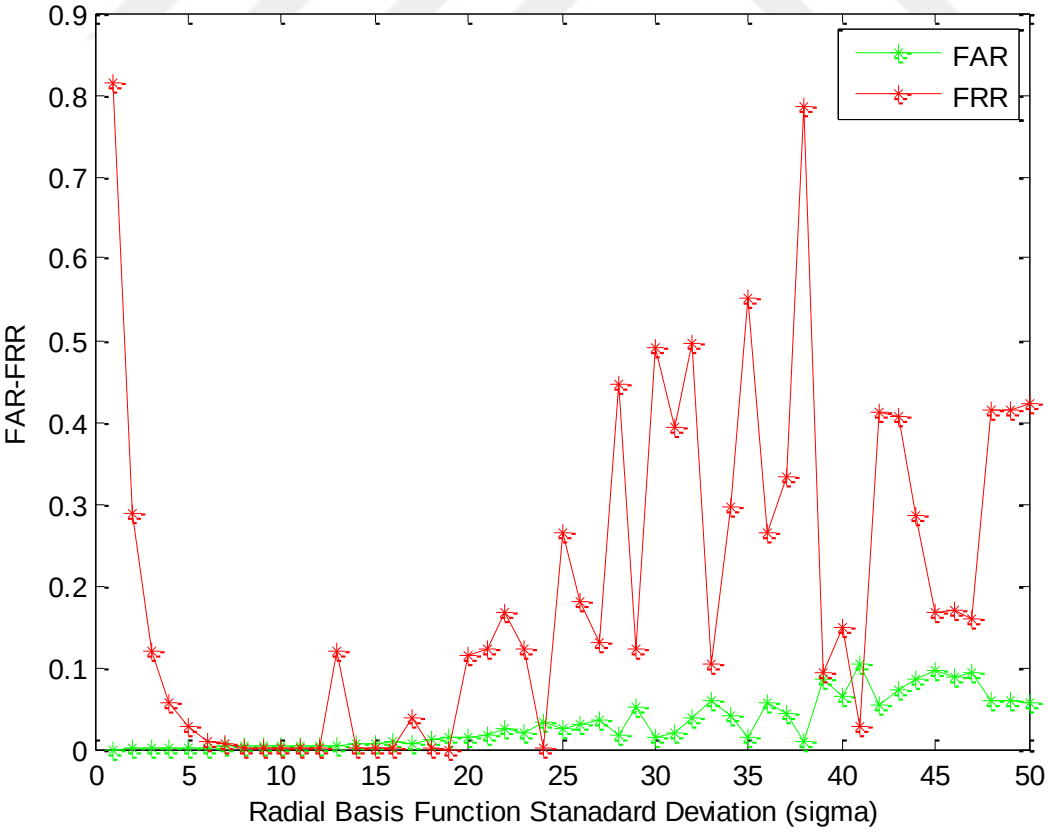


Figure 5. 10 DCC ROC Responses for PI LOPO (JAFFE) feature Length=300

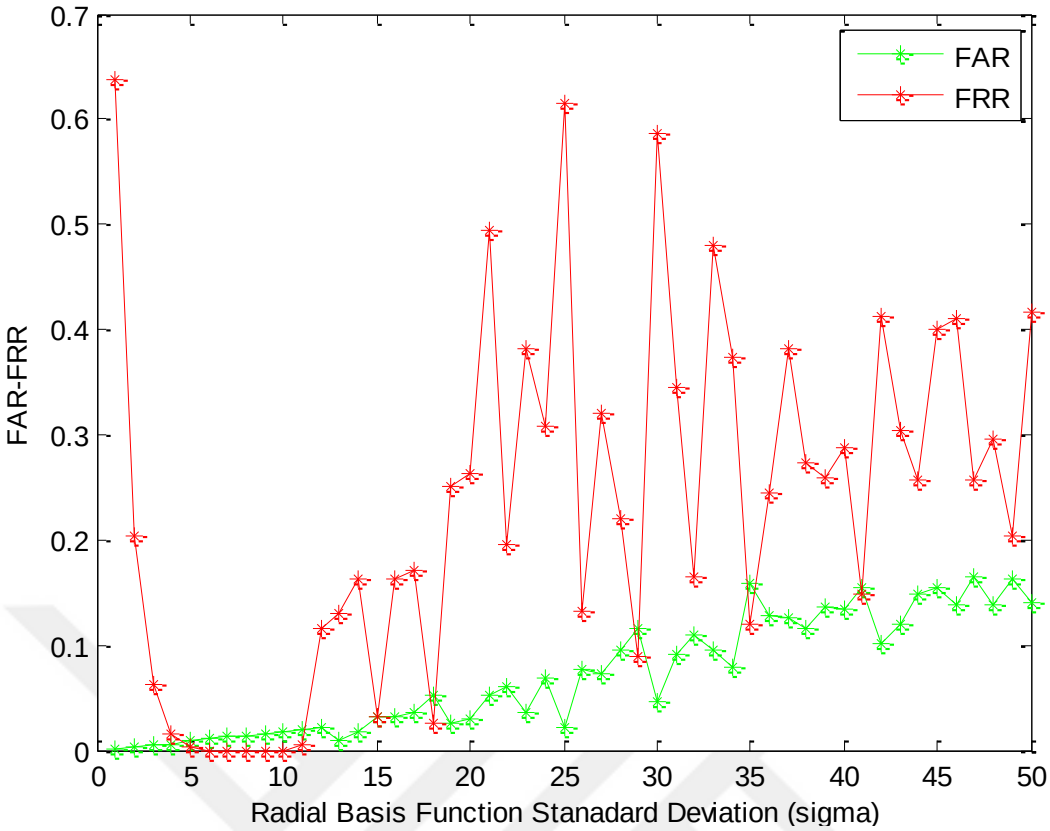


Figure 5. 11 DCC ROC Responses for PI LOPO (CK) feature Length=600

5.3. Results for Monogenic Local Gabor Binary Patterns (mono-LGBP^{riu2})

From this technique experiments were conducted using Japanese Female Facial Expression JAFFE (Michael, J. L *et. al* 1998) and Taiwanese Facial Expression Image Database TFEID (Yurtkan K. *et. al* 2014). JAFFE database is made up of 213 images collected from 10 different subjects. Each subject has at least 3 samples and at most 4 sample images per single expression class. To evaluate the proposed method performance 210 out of 213 images in this database were used. It is made up of facial images collected from 20 different female models. There are 8 expressions in this database including contempt, each image with frontal view inclination of 0^o to 45^o (Li-Fen & Yu-Shiuan, 2007). In the training phase using LOPO cross-validation, samples images were categorized into their primary expressions type (i.e. 7 for JAFFE and 8 for TFEID). One image sample at a time is taken from each expression's group as a testing sample whereas the remaining image samples were used for training. The procedure is repeated using another sample that has not been used, until all the image samples were individually and uniquely used as testing data.

Table 5.20 to 5.21 present results from the experimental using the proposed approach as compared with other approaches. Outcomes of numerous normalization algorithms and three other types of LGP algorithms (LGBP, Gabor magnitude features, and LGPP) were computed and included to give a basis for comparison. For all the variants of LGP and proposed approach same experimental procedures were used to ensure fairness. to highlight the effects of the fusion over the non-fusion techniques, results from the proposed mono_LGBP were used to implement fusion techniques.

5.3.1. Results Discussion

Worthy of noticing is that, the mono-LGBP performance improves with the increase in the scales of the Gabor filter (Figure 5.12-5.13). The fusion results using Tanh and Z-score normalization schemes recorded better performance. This may not be unconnected to the fact that in mono-LGBP, individually the normalization schemes have unique ability to extract features that are independent of one another. Therefore, fusion of these results led into an improvement in performance. Similar circumstance may not be true for the others LGP algorithms. For instance, Gabor magnitude (*Gabor mag*) produced best results with Tanh, Z-score, and min-max normalizations schemes but unfortunately, there is not much of a distinctive feature between then and hence all are pointing at similar information content as per facial expression class is concerned and applying fusion to them does not record improve in performances. The same argument is applied to LGBP and LGPP (Ashir & Eleyan, 2017).

Table 5. 20 Experimental Results at three scales ($\theta = 8,\ v = 3$) using JAFFE Database.

	Normalization Algorithms				
Feature Extractors	None	Z-score	Min-Max	M-MAD	Tanh
LGP-Magnitude	90.00	91.4 0	91.40	90.00	91.40
LGPP	77.10	82.10	80.10	77.10	85.70
LGBP	71.40	71.40	60.00	71.40	80.10
mono-LGBP	87.10	91.40	82.90	90 .00	91.40
Fused mono-LGBP		**Z-Score+Tanh = 92.83**			

Table 5. 21 Experimental Results at three scales ($\theta = 8,\ v = 3$) using TFEID Database.

	Normalization Algorithms				
Feature Extractors	None	Z-score	Min-Max	M-MAD	Tanh
LGP-Magnitude	89.20	90.50	91.40	93.30	92.40
LGPP	77.10	82.10	80.10	77.10	85.70
LGBP	71.40	71.40	60.00	71.40	80.10
mono-LGBP	91.10	96.40	92.90	94.50	97.40
Fused mono-LGBP		**Z-Score+Tanh = 97.9**			

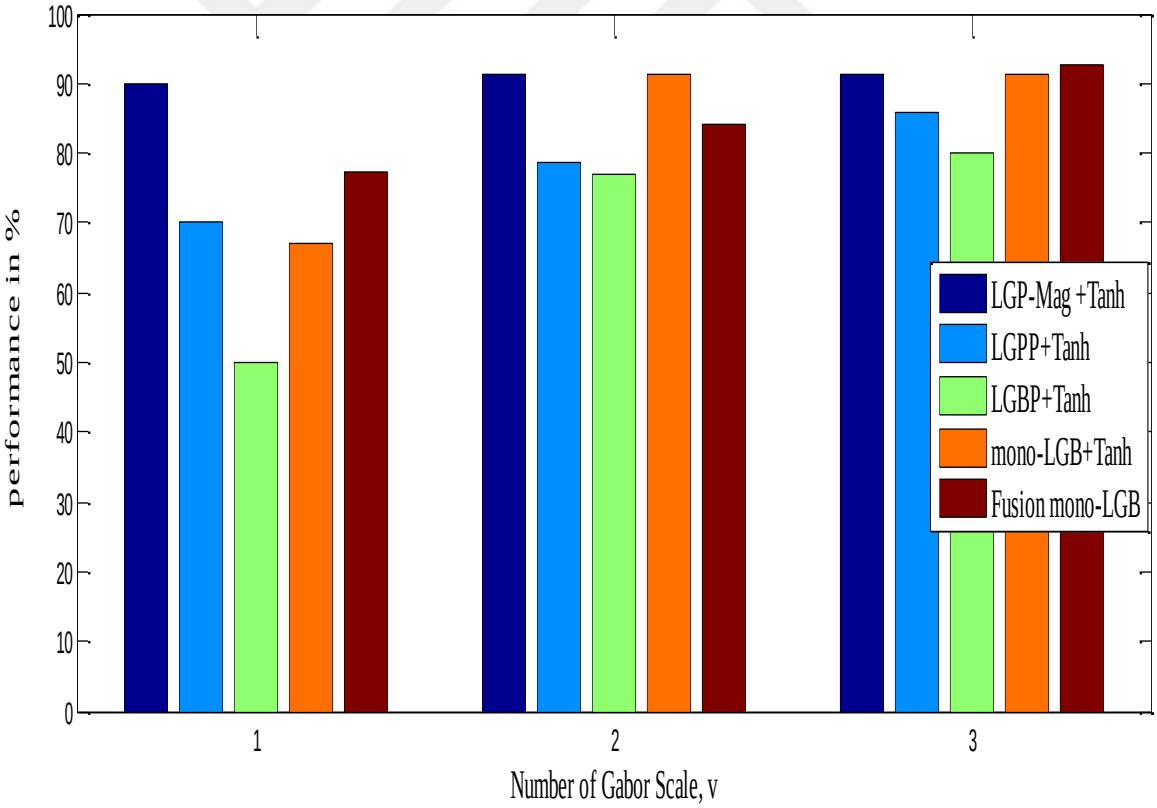


Figure 5. 12 Performance comparison in JAFFE with varying Gabor filter scale, v.

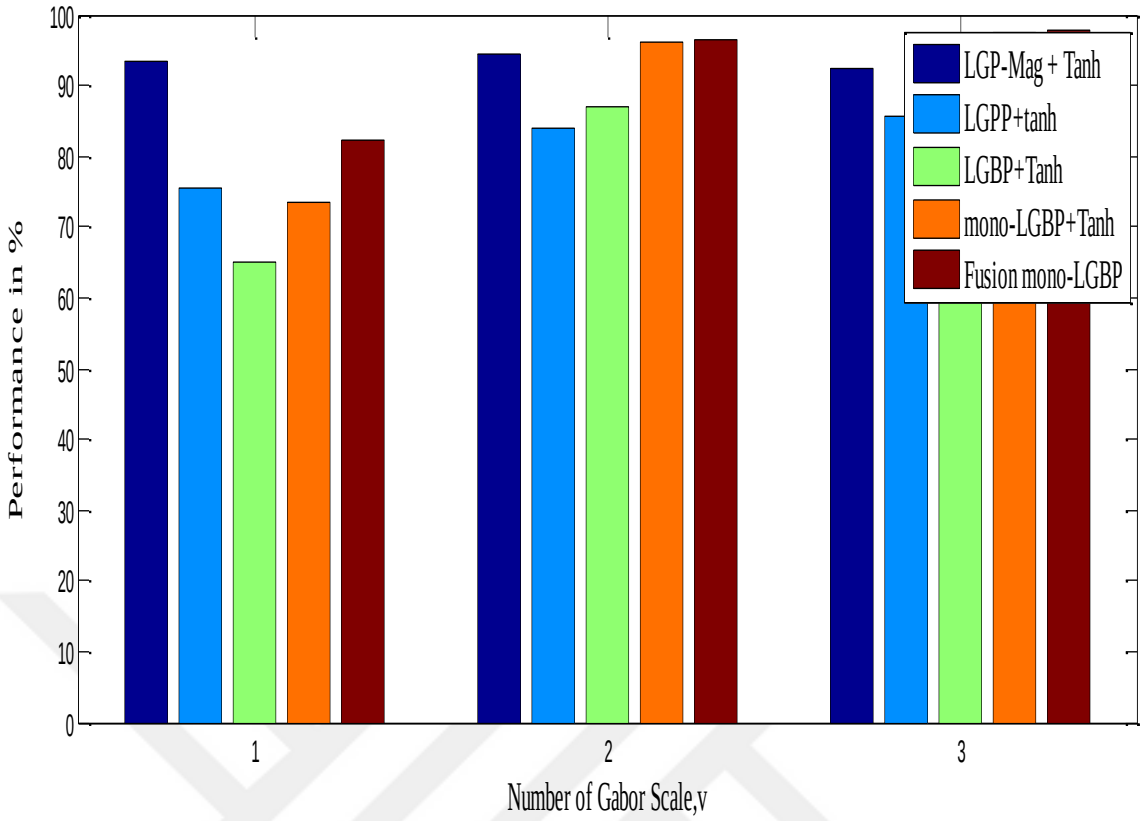


Figure 5. 13 Performance comparison in TFEID with varying Gabor filter scale, v.

6. CONCLUSION AND FUTURE WORK

6.1. CONCLUSION

A new approach for facial expression recognition using Image Pyramid and Single Branch Decision Tree has been implemented. The approach possesses both simplicity and robustness. The feature extraction technique capitalizes on both the properties of multi-resolution preprocessing algorithms and the robustness of compressive sensing algorithm with its exceptional dimension reduction capability without compromise to performance. The measurement metric used to collect compressive measurements has been strictly random variable Gaussian matrix which is easy and simple to generate. With respect to overall performance, SBDT for multi-class SVM has proven to be consistent, effective competitive, and computationally less expensive compared to the state-of-the-art multi-level classification algorithms compared with in the course of this work. Overall, this technique has leverage over most of its counterparts in the literature as a testament for consistency.

Moreover, another novel approach for facial expression recognition based on Arithmetic Mean Difference with Dynamic Cascaded Classification has been developed. This technique was conceptually designed for robust feature extraction capability and efficient classification tolerance. The combination of arithmetic mean difference and compressive sensing as feature extraction was proven valid through the experimental results. Similarly, the compressive sensing techniques used to transform feature vector into Gaussian space in the proposed DCC, which provides a platform upon which it can virtually handle any feature extraction technique with capability of learning any training pattern. From the experimental results it is also clear that automatic kernel parameter selection optimization technique converges to an optimum value. Furthermore, the single-branch decision tree proposed has also lived to expectations and has a less overheads in evaluating solution compared to its counterparts. The overall approach has been able to achieve effectiveness and simplicity in feature extraction and reduced training and generalization error of the classifier which together, results in good performance.

Lastly a new approach for facial expression recognition using Monogenic Local Gabor Binary Patterns was also developed. LGP algorithms with varying normalization schemes were used for comparison to evaluate its performance. An improved performance of approximately 92.80% on JAFFE database and 97.90% on TFEID

database were recorded. The results were compared with the state-of-the-art results on JAFFE database and TFEID in literature using KNN as a classifier. The normalization schemes used indicated that some improved performances could be achieved when suitable normalization algorithms are applied to the feature vectors. Experimental results also highlighted the effectiveness of the fusion technique used in the evaluation stage

6.2. FUTURE WORK

Researches in Facial Expression Recognition are ongoing due to the emerging interest in a vast field of human endeavors. Despite the progress being recorded over times, the pressing needs for speed, accuracy and universality of the process continues to push the limits for the emergence of new approaches that could stand the test of time and keep up with the skyrocketing demands. Recently, algorithms that mimic the human visual cortex and enable image analysis and detection in a similar way to visual receptive fields with numerous overlapping visual neurons, have drawn a lot of interest. One of these algorithms is a Convolutional Neural Network (CNN), which opens a different frontier different to the conventional approaches including the ones proposed in the thesis. Exploitation of these new biologically inspired algorithms could potentially reveal yet one of the most promising universal solutions to most of the problems inherent in facial expression recognition. The studies of the CNN and likes would encompass design of a banks of filters effective and sensitive to the various emotions of humans across all races and cultures.

REFERENCES

A., A., & A. Eleyan, A. (2017). Facial Expression Recognition based on Compressive Sensing and Pyramid Processing. *Pamukkale University Journal of Engineering Sciences PAJES, 23*(5), 527-531.

Aleksic, P., & Katsaggelos, A. (2006). Automatic Facial Expression Recognition using Facial Animation Parameters and Multi-stream HMMS. *IEEE Transactions on Information Forensics and Security, 1*(1), 3-11.

Arof, H., & Deravi, F. (1998). Circular Neighbourhood and 1-D DFT Features for Texture Classification and Segmentation. *Vision, Image, and Signal Processing, 145*, 167-172.

Ashir, A., & Eleyan, A. (2017). A Monogenıc Local Gabor Bınary Pattern For Facıal Expressıon Recognıtıon. *Selcuk University Journal of Engineering Sciences and Technology, 5*(4), 414-422.

Ashir, A., & Eleyan, A. (2017). Facial Expression Recognition based on Image Pyramid and Single Branch Decision Tree. *Signal, Image and Video Processing, SIViP , 9*(1), 1-8.

Baochang, Z., & Shiguang, S. (2007). Histogram of Gabor Phase Patterns (HGPP):A Novel Object Representation Approach for Face Recognition. *Transaction on Image processing*, 57-68.

Baraniuk, R. G. (2007, August 8). Compressive Sensing [Lecture Notes]. *IEEE Signal Processing Magazine, 24*(4), 118 - 121.

Bartlett, M. ,., Littlewort, G., Fasel, I., & Movellan, J. R. (2003). eal-time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction. *Conference on Computer Vision and Pattern Recognition Workshop, CVPRW '03.* . Madison, Wisconsin, USA.

Belhumeur, P. N., Hespanha J, P., & Kriegman D, J. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. 711 – 720.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. (pp. 144-152.). IEEE.

Broomhead, D. S., & Lowe, D. (1988). Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. *ROYAL SIGNALS AND RADAR ESTABLISHMENT*.

Cai, L., & Yin, Z. (2009). A new Approach of Facial Expression Recognition based on Contourlet Transform. *International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR* , (pp. 275–280). Baoding, China.

Cai, L.-B., & Ying, Z.-L. (2009). A new Approach of Facial Expression Recognition based on Contourlet Transform. *International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2009.* , (pp. 275–280). Baoding, China.

Candès, E. J. (2006). Compressive Sampling,. *ICM,, 3*, 1433-1452.

Chao, W. L., Ding J, J., & Liu J, Z. (2015). Facial Expression Recognition based on Improved Local Binary Pattern and Class-regularized Locality Preserving Projection. *Signal Processing, 117*, 1-10.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing Multiple Parameters for Support Vector Machines. *Journal of Machine Learning, 46*, 131-159.

Coleman, J. C. (1949, .. 63, (1, Whole). Facial expressions of emotion. *63*(296), 1.

Cootes T, F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active Shape Models, Their Training and Application. *Computer Vision and Image Understanding, 61*(1), 38-59.

Darwin, C. (1872). *The expression of Emotion in Man and Animals.* Chicago: University of Chicago Press, 1965.

Davis, R. C. (1934). The specificity of facial expressions. *Journal of General Psychology, 10*, 42–58.

Dietterich, T. ,., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research, 2*, 263-286.

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research, 2*, 263-286.

Donato, G., Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence., 21*(10), 974–989.

Ekman, P. (1993). Facial Expression and Emotion. *Am. Psychologist, 48*, 384-392.

Ekman, P., & Friesen W, V. (1976). *Pictures of facial Affect.* Palo Alto, California: Consulting Psychologist.

Ekman, P., & Friesen W, V. (1978). *The Facial Action Coding System: A Technique for the Measurement of Facial Movement.* San Francisco: Consulting Psychologist Press.

Ekman, P., & Friesen, W. V. (1982). *EMFACS.* San Francisco,: Unpublished manuscript.

Ekman, P., Friesen, W. V., & Tomkins, S. S. (1971). Facial Affect Scoring Technique (FAST) first validity study. *Semiotica, 3*, 37–58.

Eleyan, A., Demirel, H., & Özkaramanli, H. (2008). Complex Wavelet Transform-Based Face Recognition. *EURASIP Journal on Advances in Signal Processing*, 1-13.

Eleyan, A., Kose, K., & Cetin, A. (2014). Image Feature Extraction Using Compressive Sensing. *S. Choras R. (eds) Image Processing and Communications Challenges 5. Advances in Intelligent Systems and Computing* (pp. 177-184). Heidelberg: springer.

Essa, I., & Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 757–763.

Fasel, B., & Juergen, L. (2003). Automatic facial Expression Analysis. *The journal of Pattern Recognition Society, 36*(1), 259-275.

Friedman, J. H. (1996). *Another approach to polychotomous classification. Technical report.* Stanford, Department of Statistics,. Retrieved from http://wwwstat.stanford.edu/reports/friedman/poly.ps.Z

Frois-Wittman, J. (1930). The judgment of facial expression. *Journal of Experimental Psychology, 13*, 113–151.

Gonzales, R. C., & Woods, R. E. (2008). Wavelets and Multiresolution Processing. In *Digital Image Processing 2nd Ed* (pp. 483-492). USA.: Pearson publishers.

Goodenough, F. L., & Tinker, M. A. (1931). The relative potency of facial expression and verbal description of stimulus in the judgment of emotion. *Journal of Comparative Psychology*(12), 365–370.

Gu, W., Xiang, C., Venkatesh, Y. V., Huang, D., & Lin, H. (2012). Facial Expression Recognition using Radial encoding of local Gabor Features and Classifier Synthesis. *Pattern Recognition, 45*, 80–91.

Guo, G., & Dye, R. (2007). Facial Expression Recognition based on Gabor Histogram Feature and MVBoost. *Journal of Comput. Res. Dev, 44*, 1089–1096.

Huang, M.-W., Wang, Z.-w., & Ying, Z.-L. (2010). A new Method for Facial Expression Recognition based on Sparse Representation plus LBP. *International Congress on Image and Signal Processing (CISP)* , (pp. 1750–1754). Yantai, China.

imren, D., Madhav, S., Semih, D., Madhu, S. S., Marc, L. P., & Ramazan, S. A. (2014). Evaluation of Normalization and PCA on the Performance of Classifiers for Protein Crystallization Images. *IEEE conference on SOUTHEASTCON* (pp. 1-6). Lexington, KY, USA: IEEE.

John, C., Nello, C., & John, S.-T. (2000). Large margin DAGs for multiclass classification. *Advances in neural information processing systems*, (pp. 547-553).

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive Database for Facial Expression Analysis. *in Proc. FG2000-ICAFGR* (pp. 46-53). IEEE.

Knerr, S., Personnaz, L., & Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. (F. Soulié, & J. Hérault, Eds.) *Neurocomputing. NATO ASI Series (Series F: Computer and Systems Sciences), 68*, 41-50.

Landis, C. (1924). Studies of emotional reactions: II. General behavior and facial expression. *Journal of Comparative Psychology, 4*, 447–509.

Li, Z. S., Imai, J., & Kaneko, M. (2010). Facial Expression Recognition using Facial-component-based Bag of Words and PHOG Descriptors. *The Journal of The Institute of Image Information and Television Engineers, 64*(2), 230-236.

Li-Fen, C., & Yu-Shiuan, Y. (2007). *Taiwanese Facial Expression Image Database* . Taipei, Taiwan: Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University.

Lin, C., & Hsu, C. (2002). A Comparison of Methods for Multi-class Support Vector Machines. *IEEE transaction on Neural Network, 13*, 415 – 425.

Littlewort, G., Bartlett, M., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of Facial Expression Extracted Automatically from Video. *Image and Vision Computing, 24*(6), 615-625.

Liu, C., & Wechsler, H. (2003). Independent Component Analysis of Gabor Features for Face Recognition. *IEEE Transactions on Neural Networks, 14*(4), 919-928.

Liu, S., Ruan, Q., Wang, C., & An, G. (2012). Tensor Rank one Differential Graph Preserving Analysis for Facial Expression Recognition. *Image and Vision Computing, 30*, 535–545.

Lowe, D. (1989). Adaptive radial basis function nonlinearities, and the problem of generalisation. *First IEE International Conference on Artificial Neural Networks* (pp. 171-175). London, UK: IET.

Lucey, P., Cohn, J. F., & Kanade, T. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),*, (pp. 9-14). San Francisco, CA, USA.

Maenpa, T., Ojala, T., Pietikainen, M., & Soriano, M. (2000). Robust Texture Classification by Subsets of Local Binary Patterns. *15th Int'l Conf. on Pattern Recognition*, (pp. 947-950).

Mase, K. (1991,). Recognition of facial expression from optical flow. *EICE Transactions E,, 74* (10), 3474–3483.

Michael, J. L., Shigeru, A., Miyuki, K., & Jiro, G. (1998). Coding Facial Expressions with Gabor Wavelets. *International Conference on Automatic Face and Gesture Recognition* (pp. 200-205). Nara, Japan: IEEE.

Michael, J. L., Shigeru, A., Miyuki, K., & Jiro, G. (1998). Coding Facial Expressions with Gabor Wavelets. *AFGR*, (pp. 200-205). chicago.

Min, T., & Feng, C. (2013). Facial Expression Recognition and its Application based on Curvelet Transform and PSO_SVM. *International Journal for Light and Electron Optics, 124*, 5401-5406.

Moeini, A., Faez, K., Sadeghi, H., & Moeini, H. (2016). 2D Facial Expression Recognition via 3D Reconstruction and Feature Fusion. *Journal of Visual Communication and Image Representation, 35*, 1-14.

Nandakumar, N., Anil, K., & Arun, A. A. (2005). Score Normalization in Multimodal Biometric Systems. *The Journal of Pattern Recognition Society, 38*, 2270-2285.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multi-resolution Gray Scale and Rotation Invariant Texture Analysis with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 971-987.

Oliver, J. (2018, 4 Monday). *THE MUSCLES OF FACIAL EXPRESSION*. Retrieved from Teach Me Anatomy: http://teachmeanatomy.info/head/muscles/facial-expression/

Rao, K., & Koolagudi, S. (2015, July). Recognition of emotions from video using acoustic and facial. *Signal, Image and Video processing SIViP*, *9*(5), 1029–1045.

Ribaric, S., & Fratric, I. (2006). Experimental Evaluation of Matching-Score Normalization Techniques on Different Multimodal Biometric Systems. *IEEE Mediterranean Electrotechnical Conference, MELECON*. (pp. 498-501). Malaga, Spain : IEEE.

Rivera, A. R., Castillo, J. R., & Chae, O. (2013). Local Directional Number Pattern for Face Analysis: Face and Expression Recognition. *IEEE Transaction on Image Processing, 22*(5), 1740 – 1752.

Rudovic, O., Pavlovic, V., & Pantic, M. (2012). Multi-output Laplacian Dynamic Ordinal Regression for Facial Expression Recognition and Intensity Estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2634-2641). Providence, RI, USA.

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. K., & Klerman, G. L. (1976). Facial muscle patterning to affective imagery in depressed and non-depressed subjects. *Science, 192*, 489–491.

Shan, C., Gong, S., & McOwan, P. (2005). Robust Facial Expression Recognition using Local Binary Patterns. *IEEE International Conference on Image Processing, ICIP*, (pp. 370–373). Genova, Italy.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial Expression Recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816.

Shannon, C. (1949, Jan.). Communication in the Presence of Noise. *Proceedings of IRE, 37*, 10-21.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379-423.

Shiqing, Z., Lemin, L., & Zhijin, Z. (2012, ," in proc. ICSP,). Facial Expression Recognition based on Gabor Wavelets and Sparse Representation. *International Conference of Signal Processing ICSP*, (pp. 816-819.).

Shishir, B., & Ganesh, K. V. (2008). Recognition of Facial Expressions using Gabor Wavelets and Learning Vector Quantization. *Journals of Engineering Applications of Artificial Intelligence, 21*, 1056-1064.

Tian, Y. (2004). Evaluation of Face Resolution for Expression Analysis. *Conference on Computer Vision and Pattern Recognition Workshop, CVPRW '04.* Washington, DC, USA.

Tomkins, S. S. (1962). Affect, imagery, consciousness. *Springer*, 1.

Turk, M., & Pentland, A. (1991). Eigenfaces for Reccognition. *Journal of Cognitive Neuroscience, 3*, 71-86.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. *International Conference on Language Resources and Eval: Workshop on EMOTION*, (pp. 65–70).

Vapnik, V. (1991). Principles of Risk Minimization for Learning Theory. *Advances in neural information processing systems*, 832-838.

Vapnik, V. (1998). *Statistical Learning Theory.* New York: Wiley.

Wei-Lu, C., Jian-Jiun, D., & Jun-Zuo, L. (2015). Facial Expression Recognition based on Improved Local Binary Pattern and Class-Regularized Locality Preserving Projection. *An International Journal of Signal Processing*(117), 1-10.

Wenfei, G., Cheng, X., Venkatesh, Y. V., Dong, H., & Hai, L. (2012). Facial Expression Recognition using Radial Encoding of Local Gabor Features and Classifier Synthesis. *The Journal of Pattern recognition Society, 45*, 80-91.

Yeasin, M., Bullot, B., & Sharma, R. (2004). From Facial Expression to Level of Interest: a Spatio-temporal Approach. *IEEE Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, CVPR.*, (pp. 922–927). Washington, DC, USA.

Yimo, G., & Zhengguang, X. (2008). Local Gabor Phase Difference Pattern for Face Recognition. *Proceedings of ICPR* (pp. 1-4). IEEE.

Ying, Z., & Fang, X. (2008). Combining LBP and Adaboost for Facial Expression Recognition. *9th International Conference on Signal Processing, 2008. ICSP 2008*, (pp. 1461-1464). Beijing, China.

Ying-li, T., Kanade, T., & Cohn, J. F. (2001). Recognizing Action Units for facial Expression Analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 23*(2), 1-19.

Zavaschi, T. H., Koerich, A. L., & Oliveira, L. E. (2011). Facial Expression Recognition using Ensemble of Classifiers. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 1489-1492). Prague, Czech Republic.

Zavaschi, T., Koerich, A., & Oliveira, L. (2011). Facial Expression Recognition using Ensemble of Classifiers. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1489-1492). Prague, Czech Republic: IEEE.

Zhang, B., Shan, S., Chen, X., & Gao, W. (2007). Histogram of Gabor Phase Patterns (HGPP): A Novel Object Representation Approach for Face Recognition. *IEEE Transaction on Image Processing, 16*(1), 57–68.

Zhang, Z., Fang, C., & Ding, X. (2011). Facial Expression Analysis Across Databases. *International Conference on Multimedia Technology (ICMT)*, (pp. 317–320). Hangzhou, China.

Zhang, Z., Xu, C., Wang, J. X., & Chen, X. N. (2012). Facial Expression Recognition based on MB-LGBP Feature and Multi-level Classification. *Journal of Advanced Intelligence and Soft Computing, 129*, 37-42.

Zhang, Z., Xu, C., Wang, J. X., & Chen, X. N. (2012). Facial Expression Recognition based on MB-LGBP Feature and Multi-level Classification. *Journal of Advanced Intelligence Soft Computing*, *129*, 37–42.

Zhanyi, W., Weiran, X., Jiani, H., & Jun, G. (2010). A Multiclass SVM Method via Probabilistic Error-Correcting Output Codes. *International Conference on Internet Technology and Applications* (pp. 1-4). Wuhan, China: IEEE.

Zhao, G., & Pietikainen, M. (2007). Dynamic Texture Recognition using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 915–928.

Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., & Metaxas, D. N. (2012). Learning Active Facial Patches for Expression Analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 634-2641). Providence, RI, USA.

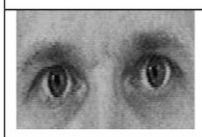
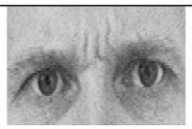
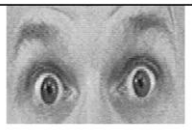
Zilu, Y., & Xieyan, F. (2008). Combining LBP and Adaboost for facial expression recognition. *International Conference on Signal Processing, 2008. ICSP* , (pp. 1661-1464). Beijing, China.

APPENDIX

A-1 AUs Combination for Emotions in Lower Face

NEUTRAL	AU 9	AU 10	AU 12	AU 20
Lips relaxed and closed.	The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent.	Lip corners are pulled obliquely.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU15	AU 17	AU 25	AU 26	AU 27
The corners of the lips are pulled down.	The chin boss is pushed upwards.	Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched open and the mandible pulled downwards.
AU 23+24	AU 9+17	AU9+25	AU9+17+23+24	AU10+17
Lips tightened, narrowed, and pressed together.				
AU 10+25	AU 10+15+17	AU 12+25	AU12+26	AU 15+17
AU 17+23+24	AU 20+25			

A-2 AUs Combination for Emotions in Upper Face

NEUTRAL	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
				
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

CARRICULUM VITAE

PERSONAL INFORMATION

Name Surname	**:**	Abubakar Muhammad ASHIR
Nationality	**:**	Nigerian
Birth Place and Date	**:**	Gulu Nigeria, 01-10-1982
Phone	**:**	+905059931829
Fax	**:**	
e-mail	**:**	ashir4real@yahoo.com

EDUCATION

DEGREE		**Name, Province, Country**	**Completion Year**
High School	:	**Maitama Sule Science College, Gaya Kano, Nigeria**	**2001**
University	:	**Ahmadu Bello University, Zaria, Nigeria**	**2009**
Masters	:	**Mevlana University, Konya, Turkey**	**2014**
Doctorate	:	**Selçuk University, Konya, Turkey**	**2018**

WORK EXPERIENCES

Year	**Organization**	**Post**

AREA OF SPECIALIZATION

Pattern Recognition, Biometrics, Digital Signal Processing, Industrial Control and Automations.

FOREIGN LANGUAGE

Hausa, English, Turkish

OTHERS

PUBLICATION

Journals

1. **Ashir, A.,** & Eleyan, A. (2017). Facial Expression Recognition based on Image Pyramid and Single Branch Decision Tree. *Signal, Image and Video Processing, SIViP , 9*(1), 1-8 (sci-Made from PhD Thesis).
2. **Ashir, A.,** & Eleyan, A. (2017). Facial Expression Recognition based on Compressive Sensing and Pyramid Processing. *Pamukkale University Journal of Engineering Sciences PAJES*, 23(5), 527-531(Made from PhD Thesis).
3. **Ashir, A.,** & Eleyan, A. (2017). A Monogenıc Local Gabor Bınary Pattern For Facıal Expressıon Recognıtıon. *Selcuk University Journal of Engineering Sciences and Technology*, 5(4), 414-422 (Made from PhD Thesis).

4. **Ashir, A.,** Akdemir, B., & Bilal, A. (2018). Manhattan Penalty Based Multi-Modal System for Face Recognition, *Scholars Journal of Research in Mathematics and Computer Science*, 3(1), 2581-3064.
5. **Ashir, A.,** Akdemir, B., & Eleyan, A. (2018). Facial Expression Recognition with Dynamic Cascaded Classifier, In Review with Journal of Neural Computing and Application (NCA) (sci-Made from PhD Thesis).

Conferences

6. **Ashir, A**., & Akdemir, B. (2018). Facial Expression Recognition with an Optimized Radial Basis Kernel, *IEEE 6th International Symposium on Digital Forensic and Security*, 22-25 March 2018, Antalya Turkey (Made from PhD Thesis).
7. **Ashir, A**., & Eleyan, A. (2016). Compressive Sensing Based Facial Expression Recognition. *IEEE International Conference on Signal Processing and Communication Application Conference (SIU)*, 16-19 May 2016, 1-4, Zonguldak, Turkey(Made from PhD Thesis).
8. **Ashir, A.,** Atef A. Ata and Mohammad S. Salman," FPGA-Based Image Processing System for Quality Control and Palletization Applications", *IEEE International Conference on Autonomous Robot Systems and Competitions*, ICARSC, Espinho, Portugal, May 2014.
9. **Ashir**, **A**. and Eleyan, A. "A Multi-resolution Approach for Edge Detection using Ant Colony Optimization", *IEEE International Conference on Signal Processing and Communications Applications,(SIU2015)*, Malatya, Turkey, 2015.
10. Shehu, G. S., **Ashir, A.** and Eleyan, A." Character Recognition using Correlation & Hamming Distance", *IEEE International Conference on Signal Processing and Communications Applications, (SIU2015)*, Malatya, Turkey, 2015.
11. **Ashir**,**A**., and G. S. Shehu, "Adaptive Clustering Algorithm for Optical Character Recognition", *IEEE International Conference on Electronics, Computers and Artificial Intelligence (ECAI),*Bucharest, Romania, 2015.
12. **Ashir**, **A.,** and Abo-serie, E., "Least Squares Multi-Level Approach for Centroid Detection in Active Triangulation", *IEEE International Conference on Electronics, Computers and Artificial Intelligence (ECAI),* Bucharest, Romania, 2015.