THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

AUDIO-VISUAL AFFECT RECOGNITION

Master's Thesis

Sara ZHALEHPOUR

ISTANBUL, 2014

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ELECTRICAL AND ELECTRONICAL ENGINEERING

AUDIO-VISUAL AFFECT RECOGNITION

Master's Thesis

Sara ZHALEHPOUR

Supervisor: Assoc. Prof. Çiğdem Eroğlu ERDEM

ISTANBUL, 2014

THE REPUBLIC OF TURKEY

BAHCESEHİR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ELECTRICAL AND ELECTRONICAL ENGINEERING

Title of Thesis: AUDIO-VISUAL AFFECT RECOGNITION
Name/Last Name of the Student: Sara Zhalehpour
Date of Thesis Defense: 07.08.2014

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Tunç BOZBURA
Director

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Assoc. Prof. Çiğdem EROĞLU ERDEM
Program Coordinator

Examining Committee Members:	Signature
Assoc. Prof. Çiğdem Eroğlu Erdem:	_____
Asst. Prof. Kemal Egemen Özden:	_____
Asst. Prof. Tarkan Aydın:	_____

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my supervisor Assoc. Prof. Çiğdem Eroğlu Erdem for her continuous support to my master's studies and research, for her patience, motivation, enthusiasm, and immense knowledge. Without her guidance, this thesis would have been impossible.

I'm also very grateful to Assist. Prof Tarkan Aydın and Assist. Prof. Egemen Ozden for reviewing my thesis and attending my thesis defense.

I would like to thank all of my friends, Sezer Ulukaya, Dr. Zahid Akhtar, Onur Önder, Leonardo Iheme and İlkay Öksüz who supported me and whom I often disturbed.

Last but not the least, I would like to thank my family, Mom, Dad and Sasan, for their unconditional support throughout my degree.

This thesis was supported by the Turkish Scientific and Technological Research Council (TUBITAK) under project number EEAG-110E056.

Istanbul, 2014 Sara Zhalehpour

ABSTRACT

AUDIO-VISUAL AFFECT RECOGNITION

Sara Zhalehpour

Electrical and Electronics Engineering

Thesis Supervisor: Assoc.. Prof. Dr. Çiğdem Eroğlu Erdem

August 2014, 66 Pages

Humans express their emotions through multiple modalities, including facial expressions, speech prosody and body gestures and various biological signals. Therefore, multi modal emotion recognition has been a major interest in applications requiring natural man-machine interaction and ambient intelligence scenarios, such as security, driver safety, health-care, behavioral science, education, marketing and advertising, where the response of the system to the user depends on the estimated emotional and/or mental state of the user. In the literature, various state-of-the-art techniques have been employed for emotion recognition from single modality (mainly facial expressions and speech); but there are relatively few works that combine different modalities in a single system for the analysis of human emotional state. Recent research has started focusing on extraction of emotional features from each modality and then combining the outputs of each modality for improved recognition of the user's emotional state.

In this thesis, we present an effective framework for multimodal emotion recognition based on a novel approach for automatic peak frame selection from audio-visual video sequences. Given a video with an emotional expression, peak frames are the ones at which the emotion is at its apex, and hence are expected give higher emotion recognition results. The objective of peak frame selection is to summarize the expressed emotion over a video sequence. The main steps of the proposed framework consists of extraction of video and audio features based on peak frame selection, unimodal classification and decision level fusion of audio and visual results. We evaluated the performance of our approach on eNTERFACE'05 containing six basic emotional classes recorded in English and BAUM-1 audio-visual database containing eight emotional and mental state classes recorded in Turkish. Experimental results demonstrate the effectiveness and superiority of the proposed system over other methods in the literature.

Keywords: Multimodal Emotion Recognition, Peak Frame Selection, Decision Level Fusion, Affective Computing

ÖZET

YÜZ İFADELERİ VE SESTEN ÇOK-KİPLİ DUYGU TANIMA

Sara Zhalehpour

Elektrik-Elektronik Mühendisliği

Tez Danışmanı: Doç. Dr. Çiğdem Eroğlu Erdem

Ağustos, 2014, 66 Sayfa

İnsanlar arası iletişimde yüz ifadeleri, ses tonundaki değişiklikler, vücut duruşu ve hareketleri ve diğer biyolojik sinyaller gibi kipler duygularımız hakkında ipuçları taşırlar. Günümüzde gittikçe önem kazanmakta olan insan-bilgisayar etkileşimi ve yapay zeka uygulamalarının daha doğal ve etkin hale gelebilmesi için çok kipli duygu tanıma problemi ilgi odağı haline gelmiştir. Duygu tanımanın güvenlik, sürücü güvenliği, sağlık, davranış bilimleri, eğitim, reklam ve pazarlama gibi sistemin tepkisinin kullanıcının duygusal ve zihinsel durumuna göre değişebileceği alanlarda uygulamaları vardır. Literatürde, tek-kipli duygu tanıma yöntemleri mevcuttur (örn. yüz ifadeleri ve ses kullanarak). Fakat, birden fazla kipi birleştirerek duygu ya da zihinsel durum tanımaya çalışan yöntemler daha azdır. Yakın zamanda çok-kipli duygu tanıma çalışmaları daha yüksek tanıma başarımı elde etmek amacıyla önem kazanmıştır.

Bu tezde, yüz ifadelerinden ve sesten çok kipli duygu tanıma amacıyla tepe çerçeve seçimine dayalı bir sistem öneriyoruz. Duygusal ifade içeren bir yüz videosu verildiğinde, tepe çerçeveler, duygusal ifadenin maximum olduğu yerlerdir ve duygu tanıma için kullanıldığında daha yüksek tanıma oranları vermeleri beklenir. Bu nedenle tepe çerçeve seçiminin amacı, video parçasındaki duyguyu en iyi şekilde özetlemektir. Önerilen çok kipli duygu sisteminin ana basamakları, tepe çerçeve seçimine dayalı yüz ifadelerinden ve sesten öznitelik çıkarma, tek kipli sınıflandırma ve karar aşamasında birleştirme adımlarından oluşmaktadır. Sistemin performansını altı temel duyguyu içeren İngilizce eNTERFACE ve sekiz duygu ve zihinsel durum içeren Türkçe BAUM-1 veritabanları üzerinde test ettik. Deney sonuçları önerilen sistemin literatürdeki diğer yöntemlere göre etkinliğini göstermektedir.

Anahtar Kelimeler: Çok-Kipli Duygu Tanıma, Tepe Çerçevre Seçimi, Karar Aşamasında Birleştirme

TABLE OF CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AAM	:	Active Appearance Model
AFS	:	Audio Based Frame Selection
ANN	:	Artificial Neural Networks
ASM	:	Active Shape Model
AUs	:	Action Units
BAUM	:	Bahçeşehir University Multimodal Affective Database - 1
DCT	:	Discrete Cosine Transform
DEND CLUSTER	:	Clustering Based Peak Frame Selection
DFT	:	Discrete Fourier Transform
EIFS	:	Emotion Intensity Based Frame Selection
FACS	:	Facial Action Coding System
FAP	:	Facial Animation Parameters
FFT	:	Fast Fourier Transform
GMM	:	Gaussian Mixture Model
HCI	:	Human Computer Interaction
HCRFs	:	Hidden Conditional Random Fields
HMM	:	Hidden Markov Model
HoG	:	Histogram Of Oriented Gradients
JAFFE	:	Japanese Female Facial Expression
KDEF	:	Karolinska Directed Emotional Faces
KNN	:	K-Nearest Neighbor
LBP	:	Local Binary Pattern
LDC	:	Linear Discriminant Classification
LOSO	:	Leave-One Subject-Out
LPQ	:	Local Phase Quantization
MAXDIST	:	Peak Frame Selection Based On Maximum Dissimilarity
MFCC	:	Mel-Scale Frequency Cepstral Coefficient
MLP	:	Multilayer Perceptron
PCA	:	Principle Component Analysis
PIE	:	Pose, Illumination, And Expression
PLPCs	:	Perceptual Linear Predictive Coefficients
PSF	:	Point Spread Function
PSF	:	Point Spread Function
RASTA-PLP	:	Relative Spectral Perceptual Linear Predictive
SVM	:	Support Vector Machine
TAN	:	Tree-Augmented Naive Bayes

SYMBOLS

Distance score	:	d_j
Fourier transforms of the damaged image	:	$G(u)$
Fourier transforms of the original images	:	$F(u)$
Fourier transforms of the PSF	:	$H(u)$
M-by-M neighborhoods at each pixel position x	:	N_x
$N \times N$ dissimilarity matrix	:	M
The basis vector at frequency **u**	:	ω_u
The jth component of $G_{\mathbf{x}}$ after decorrelation	:	$g_j(\mathbf{x})$
The predicted output emotional label	:	$\widetilde{\omega}$
The true class emotional label	:	ω
The vector containing the values of all M^2 image samples	:	f_x

1. INTRODUCTION

1.1 MOTIVATION

Impact and influence of computers in human life is an indisputable fact. In less than the half a century, utilization of computers has extended from universities to industry, business environments and our houses. As a result of the widespread internet usage, computers are being utilized in many aspects of our daily lives. Computers are being used for communication, education and electronic trading by a wide range of people. With the help of artificial intelligence, systems have been developed that are capable of interacting with people in order to solve a variety of problems in many fields. In parallel to the attempts and research efforts that have been spent to produce and develop secure, efficient and cheap hardware and software, it has also become very important to develop human-computer interfaces which are easy to use. This issue has become so important that human computer interaction (HCI) has become an independent field of study. HCI is based on achievements of various scientific fields such as information technology, computer graphics and psychology to make a better and more natural environment for interactions between computers and humans.

Human to human communication is based on two channels of communication, linguistic and paralinguistic. Linguistic communication is based on words and sentences that are used by humans in their interactions. Paralinguistic communication refers to the gestures such as head and body movements or tone variations of speech, which expand and complete the expressed implication. Mehrabian [1] believes that over than 90 percent of the exchanged implications between humans are through the paralinguistic channel and the facial movements, speech tone and body gestures have the most distinct effects.

Expression of emotions is an important part of our paralinguistic communication, which is often reflected in our face and body gestures and also in our voice. Frown in a conversation indicates dissatisfaction and disagreement. Smile shows pleasure. We can also say unconscious occurrence of some emotions such as boredom in human interactions has an essential and determinative role. For instance, when a teacher sees the students' tired faces he/she may decide to change or diversify the topic. Similarly, a

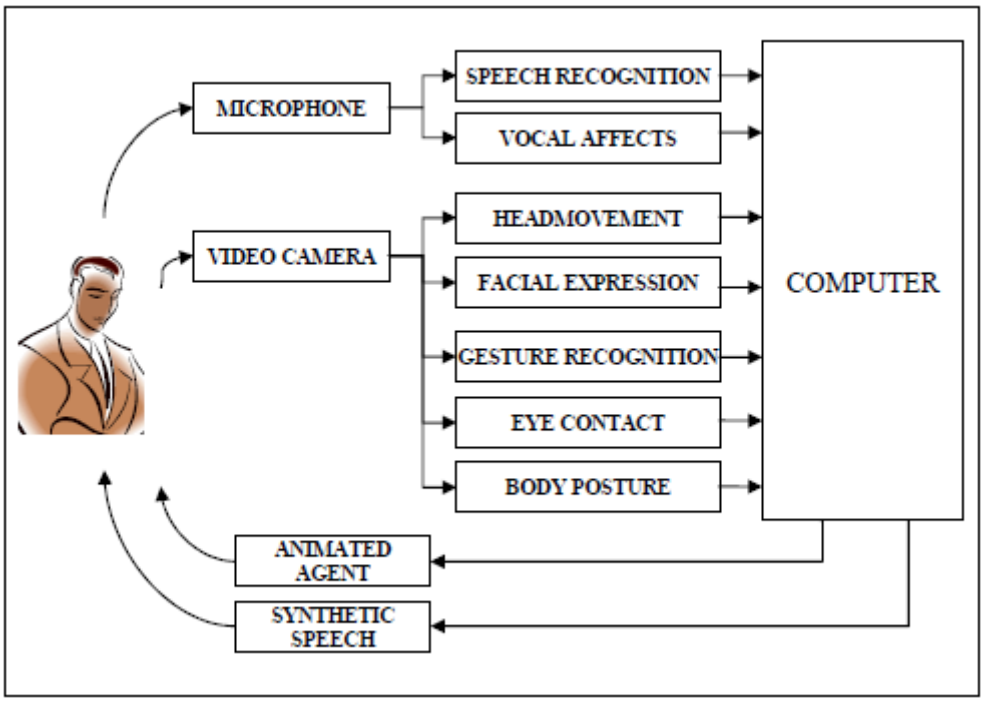
mother may understand the problem based on her infant's crying or unsettled behavior.. Therefore, the fact that we can understand each other's emotions from paralinguistic cues and react according to them enriches human-to-human communication. Using computers in such situations requires correct recognition of facial expressions, body gestures and vocal emotions of users. Meanwhile, for getting closer to natural human to human interactions, an HCI system needs to be able to respond to the emotional states. A general view of such system depicts in Figure 1.1. As we can see in the diagram, the input of the computer can be acquired from on board microphones and built in cameras which contain information of speech, facial movement, body gestures, etc. Computer on the other hand gives feedback by means of speech synthesis and animated agents. These systems can be used in applications in which computers have social roles such as a "tutor", "consultant" or even a "companion".

1.2 PROBLEM STATEMENT

Facial movements have a direct relation with emotional expressions and it has been shown that some emotional facial expressions are the same a across different societies, ages and cultures [2]. But face is not the only source of expressing emotions. Individuals are using a combination of speech information, facial movements and body gestures to show their emotions. In fact natural human-to-human emotional interactions are multimodal. Therefore, focusing on just one source or modality will give us uncertain results. .

A solution for the shortcomings of unimodal emotion recognition is employment of multiple modalities. Most of the work in the literature on multimodal emotion recognition has focused on the combination of speech and facial movements. The reason that body gestures and movements are not generally used is that humans are not able to use their hands and body freely in all situations [3]. For example if they hold something in their hands that would limit hand movements or they cannot show some gestures while sitting. Besides these limitations, technical difficulties of extraction and processing of body gestures are another reason for giving less consideration to this media. On the other hand, speech is a one dimensional signal and can be received easily by a microphone. Also, in most of the cases face can be detected and tracked by a camera form a frontal view.

Figure 1. 1: Multi-modal human-computer interaction [4]



Now the challenge that we address in this thesis is to train the computer to automatically recognize the users' emotional states by the means of audio and video inputs. Some questions that are investigated in this thesis are as follows:

i. Which features on the face and in the voice reveal a person's emotions the most?
ii. How well can we use these features to train the computer to recognize human emotions from each modality?
iii. What is the best way of combing the information from audio and video channel for having more accurate and efficient emotion recognition?

1.3 CONTRIBUTIONS AND ORGANIZATION OF THE THESIS

In this thesis we present a multi-modal emotion recognition system, based on fusion of facial expression and speech modalities. The facial expressions and speech features are classified separately and then they are fused at the decision level. The novelties of the thesis can be summarized as follows:

a) We present a facial expression recognition method from video based on a peak frame selection method. Peak frames are the frames at which the emotional expression is at its apex, which are detected automatically. We present and compare three automatic peak frame selection methods.

b) We present experimental results on a new database namely, BAUM-1, which an audio-visual database of affective and mental states recorded in Turkish. The BAUM-1 database has both acted and spontaneous recordings. The target emotions in the database are the six basic ones (**happiness, anger, sadness, disgust, fear, surprise**) and additionally **boredom** and **contempt**. We also aim to capture several mental states including **unsure (confused, undecided)**, **thinking**, **concentrating**, **interested (curious),** and **bothered**.

The publications from this thesis that have been published and in preparation are as follows:

[1]. O. Önder, O. Önder, C. E. Erdem, "BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States", *to be submitted*.

[2]. S. Zhalehpour, Z. Akhtar and C. E. Erdem, "Peak Frame Selection Based Multi-modal Affect Recognition", *to be submitted*.

[3]. S. Zhalehpour, Z. Akhtar and C. E. Erdem, "Multimodal Emotion Recognition with Automatic Peak Frame Selection", *IEEE Int. Symp. on Innovations in Intelligent Systems and Applications (INISTA)*, pp.116-121, Alberobello, Italy, June 2014.

[4]. O. Önder, S. Zhalehpour, C. E. Erdem, "A go", *IEEE Signal Processing and Applications Conference* (SIU), Girne, Northern Cyprus, April 2013.

[5]. C. Turan, C. Kansın, S. Zhalehpour, Z. Aydın, C. E. Erdem, "A Method for Extraction of Audio-Visual Facial Clips from Movies", *IEEE Signal Processing and Applications Conference* (SIU), Girne, Northern Cyprus, April 2013.

The organization of the thesis is as follows. In Chapter 2, a short literature survey on theories of emotion and several studies on speech emotion, facial expression and multimodal emotion recognition is presented. Applications of emotion recognition are discussed as well. Chapter 3 introduces the framework used in this thesis for recognizing emotional states from the audio, video and both modalities. Experimental results on two different audio-visual datasets are given in Chapter 4. Finally, conclusions and future directions for research are presented in Chapter5.

2. STATE OF THE ART IN EMOTION RECOGNITION

The main issue in constructing an automatic emotion recognition system is to define the "emotion". Most people have an instinctive perception about what emotion is, but for automatic recognition of emotions we need to define it systematically. There are various methods for automatic emotion recognition depending on the source of modalities. In this chapter, several theories for the definition of emotion and emotion recognition will be discussed. An overview of the state of art in emotion recognition is also given by presenting the prominent methods and studies conducted on facial expression recognition, emotional speech recognition and audio-visual emotion recognition.

2.1 EMOTION

Emotion is a topic that is very hard to comprehend. Currently, there is no consensus about the exact definition of emotion. A computer scientist would definitely have a different definition to that of a psychologist, behavioral scientist or an average person.

Expressing emotions is one of the ways that people show their feelings and intentions. Individuals express their emotions in the form of speech articulation, facial movements, body gestures and some biological signals like the heartbeat and body temperature. Each of these biological signs used for expressing emotions is called a mode.

Psychologists have been always interested in recognition emotion. First publications in this area belongs to Darwin which were published in 1872 [5]. He described emotions as instinctive reaction patterns that were shaped by evolution because of their survival value. He indicated that emotions were initiated in certain situations and were common in all humans and even animals. His work indicates that emotional expressions in the form of facial movements is similar among most people. The research about similarity or dissimilarity of facial expressions in people resulted in the concept of "basic emotions". Ekman confirmed the universality of facial expressions in support of the Darwin's view. He presented and defined the six basic emotions (anger, fear, disgust, sadness, surprise and happiness) that could be expressed by different facial, eye and mouth movements. According to this study each emotion is categorized based on its relevance to one of the facial expression, referred as "prototypes", Figure 2.1 shows an

example of these prototypes [6]. In 1998 Scherer [7] observed that even in congenitally blind individuals, facial expressions are similar to those without this disability. Even though seeing people improve the communication skills by interacting with the environment, the ways of expression emotions do not change that much. Such resemblances have motivated the studies on automatic emotion recognition. These studies have started from the time that basic emotions were proposed in 1992 and still continue with many progresses and developments. Most of the work have only used one channel such as facial movements or speech for emotion recognition. In recent years some studies also have been done on the combination of these channels or modalities.

Figure 2. 1: Prototypical six basic emotions according to Ekman [6]. From left to right and top to bottom: Anger, Fear, Disgust, Surprise, Happiness, and Sadness.

2.2 EXISTING AUDIO-VISUAL EMOTIONAL DATABASES

Recognizing the emotion of a person is a difficult task even for human observers. Thus, access to fully labelled databases prior to implementation of an emotion recognition system is essential. One of the solutions in this case is using the publicly available databases. These databases contain mostly acted (simulated), but sometimes elicited and natural emotional expressions. The acted emotion are obtained by asking the subjects to act a predefined emotion. The subjects are usually professional actors/actresses since

they can portray emotions more realistically. However, the acted emotions are often exaggerated compared to those expressed naturally. A way of partially overcoming this drawback is to use emotion-triggering texts and/or simulations to evoke the subject's emotion. This type of databases are called elicited emotional databases and are better than acted ones in the sense of being more realistic. Ideally, natural databases is a better choice for emotion recognition systems since they contain naturalistic and unbiased emotions [8].

Japanese Female Facial Expression (JAFFE) [9] is one of the first emotional databases. Ten subjects express 3 or 4 samples of the six basic facial expressions and a neutral face for comparison with emotional images. Karolinska Directed Emotional Faces (KDEF) database [10] contains 4900 images of 7 different emotional expressions collected from 70 amateur actors with an age range between 20 and 30 years. The Pose, Illumination, and Expression (PIE) database [11] contains 41368 images of 68 people from 13 different poses, in 43 different illumination conditions, and with 4 different expressions. These types of databases are "Still Image" databases. There are also databases which demonstrate emotions in the form of sequential series of frames such as the Cohn-Kanade [12], FEEDTUM [13] and MMI [14] databases. The Cohn-Kanade Facial Expression database contains sequences of images starting with a neutral expression and ending at the target emotion. There are a total of 327 sequences with emotion labels from 123 subjects in the extended version of the database (CK+) [15]. The FEEDTUM database contains spontaneous video clips with elicited emotions recorded from 18 subjects with six basic emotions and the neutral expression. The MMI is a very comprehensive facial expression database. It contains over 1250 videos and over 500 statistic images of about 50 subjects displaying the six basic emotions. It consists of both acted and spontaneous expressions.

There are also emotional speech databases available to researchers. The Berlin Emotional Speech database (EMO-DB) [16] contains 495 samples of acted emotional speech in German language by 10 actors. This database is comprised of six basic emotions as well as neutral speech. The AIBO database [17] contains audio samples of spontaneous emotional reactions in German and English in which children had to instruct a Sony AIBO robot to do specific tasks.

Since the current direction of research is toward accomplishing multimodal emotion recognition, recently several audio-visual databases have been collected. One of the early audio-visual databases is the Belfast database [8] which contains audio-visual clips extracted from television talk shows, current affairs programs and interviews conducted by the research team from 125 English speaking subjects. It contains a wide range of emotions. HUMAINE database [18] is a combination of 50 clips from Belfast and some other databases containing both spontaneous and acted data. The eNTERFACE'05 multimodal database [19] contains acted emotional recordings of 42 subjects of 14 different nationalities speaking in English showing prototypical emotions. Six basic emotional states are expressed in the video clips of the database in an acted way by uttering given sentences with target emotions. The Bahcesehir Universiy Multimodal Affective face database (BAUM-1) [20] is another database collected from 31 subjects containing all 6 basic emotions along with contempt, boredom and some mental states. All aforementioned databases are summarized in Table 2.1.

Table 2. 1: Some examples of available emotional databases in the literature.

Modality	Name	Number of subjects	Language	Description of emotions	Naturalness
Facial Expressions	JAFFE [9]	10	-	6 basic emotions + neutral	Acted
	KDEF [10]	70	-	6 basic emotions + neutral	Acted
	PIE database [11]	68	-	4 basic emotions	Acted
	Cohn-Kanade [12]	97	-	6 basic emotions	Acted
	FEEDTUM [13]	18	-	6 basic emotions + neutral	elicited
	MMI [14]	50	-	6 basic emotions	Acted + Elicited
Speech	EMO-DB [16]	10	German	6 basic emotions + neutral	Acted
	AIBO database [17]	51	German	Wide range	Elicited
Audio-visual	Belfast database [8]	125	English	Wide range	Naturalistic
	HUMAINE [18]		English French Hebrew	Wide range	Naturalistic + Acted
	Enterface'05 [19]	42	English	6 basic emotions	Acted
	BAUM [20]	31	Turkish	Wide range	Elicited

Source: It has been done by Sara Zhalehpour.

2.3 FACIAL EXPRESSION RECOGNITION METHODS

It is widely accepted that the human face is an important nonlinguistic means of human-to-human interaction and its changes convey individual's inner states, emotions, thoughts and even diseases. These changes are performed through movements of facial muscles. For example, when feeling surprised we open our mouth and eyes, controlling muscles of the lips and the chin are activated and the related emotion is expressed.

The most prominent work on facial expressions is the one conducted by Paul Ekman [21]. In 1987, Ekman and Freisen developed the Facial Action Coding System (FACS) to represent every human facial expression. FACS is a muscle based system which measures all visually perceivable changes in facial muscles and code expression in terms of some facial action units, called Action Units (AUs). Each facial expression may be described by an individual AU or a group of AUs (See Figure 2.2). Prior to the introduction of FACS, most of the facial behavior research relied on human observation which was not reliable and accurate. Ekman's work on studying the activities of the muscles inspired many researchers for facial expression recognition by means of image and video.

In general, every facial expression recognition system must perform a few steps before classifying the expression into a particular emotion. The first step is face detection to locate the face in a given image or video data. Once the face is detected, it needs to be tracked over the time in a video. . After the face is located, the system should extract emotion related features of the detected face. Final step is to classify the image/frame in set of emotion classes using the extracted features from previous step. Figure 2.3 summarizes three steps of a facial expression recognition system.

2.3.1 Face Acquisition

There are two basically methods for face acquisition in frontal and near frontal view images. These two methods are *face detection* and *head pose estimation*. Locating the face within a given image is called **face detection** and locating and tracking a face across multiple frames is termed as **face tracking**. Many detection and tracking methods have been employed for face detection. Among the most recent face detection algorithms, Viola and Jones method [22] has gained remarkable attention. The basic idea of that method is the use of a boosted cascade of Haar feature based classifiers. The

image regions that pass through all the stages of the classifiers are considered as the face. In order to handle the out-of-plane head motion, head pose estimation can be employed. The methods for estimating head pose can be classified as 2D image-based methods [23] and 3D model-based methods [24].

Figure 2. 2: Overview of Facial Action Coding System [25]. The facial muscles and the direction of motion are shown.

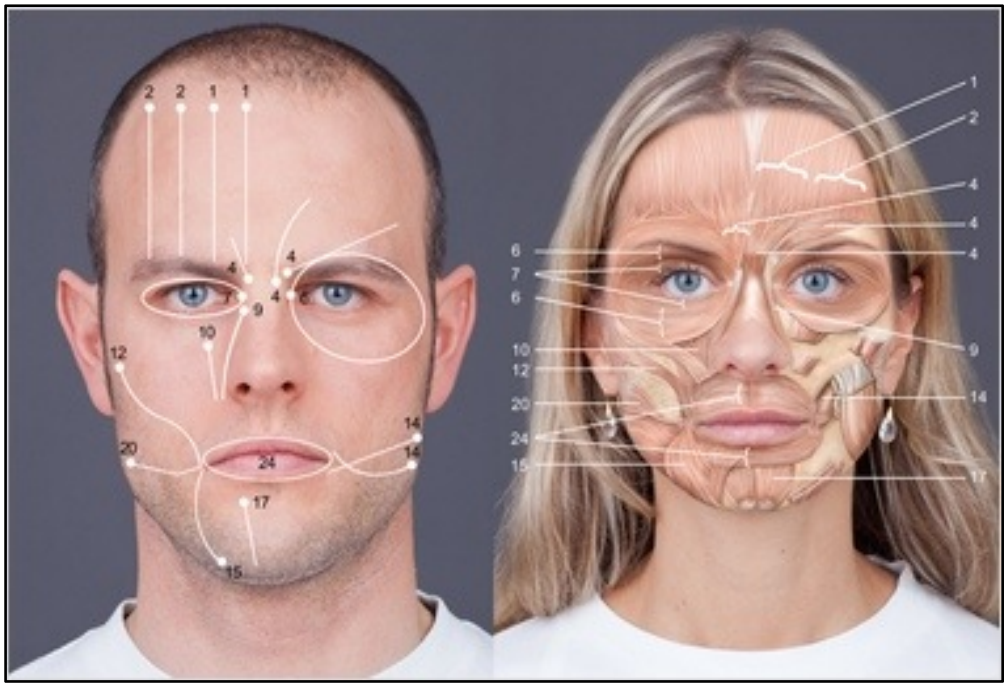
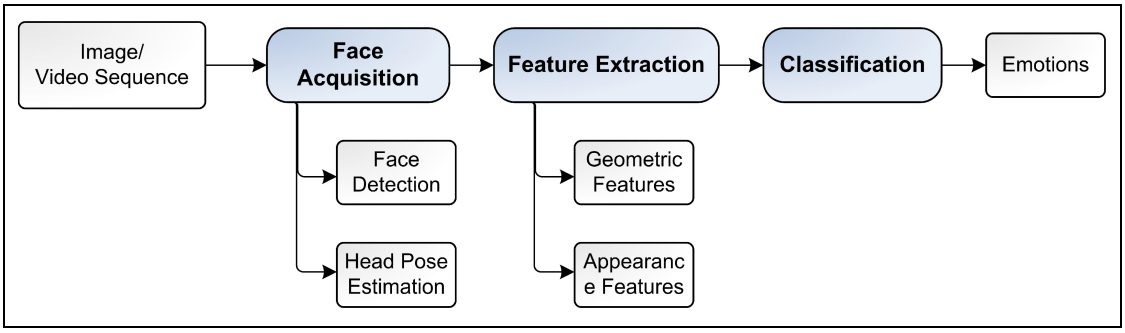


Figure 2. 3: Basic structure of a facial expression recognition system. The first step is capturing input images/frames, and detecting or estimating the location of the face in these images. The next step utilizes the provided information in the first step to extract some facial features related to the expressed emotion. These features are finally given an emotion label by using a classifier.



Source: It has been done by Sara Zhalehpour.

2.3.2 Feature Extraction

The next step after face detection is feature extraction to identify the facial expression. For feature extraction in the field of facial emotion recognition, two trends are dominant in the literature: **appearance features** and **geometric features.** Texture features appear temporarily in the face during any kind of facial expression (e.g. wrinkles, bulges, furrows, etc.). For appearance methods, image filters, such as Gabor wavelets can be applied to either the whole face or specific regions of the face in the image/frame. Geometric features are the ones always present in the face but deformed due to any kind of facial expression (e.g. contours of the eye, lips mouth, etc.). Geometric methods consist of tracking and processing the motion of facial points on the image/frame. Generally, geometric features have good tolerance to reasonable amounts of errors caused by variations in pose, size and location of the face, while texture features require a face normalization to handle such errors. However, geometric features suffer from some drawbacks such as losing regional texture and requiring accurate location and robust tracking of facial landmarks. Table 2.2 shows typical examples of the texture and geometric features used in previous studies.

The combination of the geometric and appearance features have also been used in the literature. For example, Zhang [45] tracked facial points on the face images while also using the Gabor wavelets around these points to take into account for facial expression recognition.

Table 2. 2: Texture and geometric features used for facial expression recognition in previous studies.

Feature type	Methods in the Literature
Texture (Appearance) Features	Gabor Wavelets [26, 27], Local binary pattern (LBP) [28], Haar Features [29, 30], Histogram of oriented gradients (HoG) [31], Discrete Fourier transform [32], Active appearance model (AAM) [33], Principle Component Analysis (PCA) [34], Candid Grid Node [35], Scale-invariant feature transform (SIFT) [36]
Geometric Features	Location of landmark points [37, 38], Point Distribution Model [39], Active shape model (ASM) [40], Optical flow [41], Active appearance models (AAM) [42], Shape movements [43], Facial animation parameters (FAP) [44],

Source: It has been done by Sara Zhalehpour.

2.3.3 Emotion Classification Methods

Accurate recognition of the emotional state of the input face is the aim of emotion classification. The classification in this case is a supervised learning process of a subset of features obtained at the feature extraction step. The selected classifier uses the learned patterns of emotion in training data to recognize the facial expression in test data. Emotion classifiers normally are based on either statistic image or video data input. The static classifiers such as the *k*-nearest neighbor (KNN), support vector machine (SVM) and artificial neural networks (ANN) attempt to recognize facial expressions using statistic images' features while temporal classifiers, such as Hidden Markov model (HMM) aims to utilize the temporal features of each frame.

2.3.4 Facial Expression Recognition Methods

There are many studies in the literature for recognizing emotions from face images and videos. Below we give a short overview of these methods.

Pantic and Rothkrantz [46] introduced a person independent system for recognition of six basic emotions from still images, which utilized a hybrid facial feature detection system for facial action recognition. The overall performance of the automatic system implies that the facial feature detection, the facial action unit recognition and the emotion classification are performed more accurately by the system.

Hu *et al.* [47] employed Gabor wavelets to extract features from difference images obtained by subtracting the first frame showing a frontal face from the current frame. A neural network, which takes the form of Multilayer perceptron (MLP), was used to classify the feature vector into different states of a HMM of a certain emotion sequence. After training, the output values of the NN were interpreted as the posterior of the HMM state and the Viterbi algorithm is applied to those values to estimate the best state path. When the states were generated, the neural network was analyzed by each of these HMM models and the network with the maximum probability was the winner. Their experiment over 240 samples with 4 emotions yielded a recognition rate of 95 percent.

Cohen *et al.* [48] Classified the expression from video sequences by focusing on changes in distribution assumptions, and feature dependencies. For this mean, they introduced different Bayesian network classiifers.in more details, they changed the

distribution of Naive-Bayes classifiers from Gaussian to Cauchy, and used Gaussian Tree-Augmented Naive Bayes (TAN) classifiers to learn the amount of dependency between different facial motion features. They also utilized the temporal cues for facial expression recognition from live video. They proposed a new structure of HMMs for automatic human facial expression recognition from video sequences.

Cowie *et al.* [49] classified facial expression to six basic emotion categories by a fuzzy rule based system. They used a novel confidence-based feature extraction system based on FAP extraction to feed the classifier. The average accuracy of their system was around 75 percent.

Zhang et al. [50] used a dynamic and probabilistic framework based on combining dynamic Bayesian networks (DBN) with FACS for modeling the dynamic and stochastic behaviors of spontaneous facial expressions. Their framework provided a coherent and unified hierarchical probabilistic framework to represent spatial and temporal information of facial expressions. It was also able to select visual cues with most emotional content from the available information sources to improve the recognition accuracy. The facial expressions recognition was performed by fusing the current visual observations and the previous visual evidences. The recognition results varied between 47 percent to 94 percent for different emotions.

Mansourizadeh *et al*. [51] introduced an expert system for emotion recognition from face image sequences. The system tracks a set of manually marked points of the face and classifies facial feature variations in terms of AUs. Then, the system classifies the detected action units in terms of a subset of basic emotions. Overall emotion recognition accuracy was about 80 percent.

Yeasin *et al*. [52] presented a spatiotemporal method for recognizing six universal facial expressions. Their approach relied on using a linear classification bank and optical flow vector to obtain decision and then merging them to produce a characteristic signature for each facial expressions. The discrete hidden Markov models are trained by the computed signatures from the training set to learn the underlying model for each facial expression. They achieved an accuracy around 91 percent on Cohn-Kanade database.

Kai *et al*. [53] proposed a graphical method to seamlessly couple and simultaneously analyze facial emotions and the action units. Their method was based on using the

hidden conditional random fields (HCRFs) where the output class label was linked to the underlying emotion of a facial expression sequence, and connected the hidden variables to the image frame-wise action units. As HCRFs are formulated with only the clique constraints, their labeling for hidden variables often lacks a coherent and meaningful configuration. They resolved the resulting difficulties in training by proposing a partially-observable HCRF model, and an efficient scheme via Bethe energy.

Chi *et al*. [54] proposed a facial expression system that analyzes the non-rigid morphing facial expressions and eliminates the person-specific effects through patch features extracted from facial motion due to different facial expressions. Classification and localization of the center of the facial expression in the video sequences were carried out by using a Hough forest.

Senechal *et al*. [55] proposed to use the combination of various types of features for automatic detection of AUs in facial images. They utilized a multi kernel SVM for each AU. The first kernel matrix was obtained using local Gabor binary pattern histograms and a histogram intersection kernel. The second kernel matrix was achieved from active appearance model coefficients and a radial basis function kernel. They combined these two types of features using the SimpleMKL algorithm in the training phase. SVM outputs were then averaged to obtain temporal information of each sequence. Their experiments show a promising results for GEMEP-FERA database.

Table 2.3 lists the details and results of the above mentioned studies for facial emotion recognition.

2.4 SURVEY ON SPEECH EMOTION RECOGNITION

Speech is one of the indispensable means for sharing ideas, observations, and feelings. People usually convey emotions by using speech information either explicitly through linguistic messages, or implicitly through paralinguistic messages that reflect of the way the words are spoken. Hence, considering only the verbal part, without taking into account the manner in which it was spoken, will cause loss of important aspects of the spoken message.

There are many application areas of speech emotion recognition. For example, speech recognition systems need to analyze the speech correctly in order to perform an

effectively under changes in emotions, states and tone of speakers. Doctors can diagnose possible diseases in patients. Psychologist can predict the human state of mind and human-computer interaction experts can enrich the communication between human and machines.

Table 2. 3: List of representative works in the field of visual emotion recognition.

Researcher(s)	Features	Classification	Database	Recognition rate
Pantic *et al*.[46]	FACS	Rule based system	Self-defined	91 %
Hu *et al*. [47]	Gabor wavelet	HMM + NN	Self-defined	95 %
Cohen *et al.* [48]	Facial movements	Tree-Augmented Naive–Bayes	Self-defined + Cohn-Kanade	66 %
Cowie *et al.* [49]	FACS	Tree Bayesian network	Cohn- Kanade	91 %
Zhang *et al.* [50]	FACS	Phase network	100,000 frames	84 %
Mansourizadeh *et al*. [51]	FACS	Rule based system	Cohn- Kanade	80 %
Yeasin *et al*.[52]	Optical flow	HMM	Cohn- Kanade	90 %
Feng *et al.* [28]	Face histogram	LP linear programing	Cohn- Kanade	91 %
Kai *et al.* [53]	FACS	HCRF	Cohn- Kanade	93 %
Senechal *et al.* [55]	LGBP histogram	SVM	GEMEP-FERA	65 %
Chi *et al.* [54]	Motion tracking	Hough forests	Cohn- Kanade	89 %

Source: It has been done by Sara Zhalehpour.

The general emotion recognition process from speech signals can be described in the following order: (1) capturing the speech signal and preprocessing, (2) extracting audio features, (3) recognition of emotions with an appropriate classifier. Preprocessing step of speech can include detection of voiced and unvoiced segments, end point detection, dividing signal into frames with predefined length and windowing them, etc.

2.4.1 Emotion Related Speech Features

The goal of feature extraction is to select some features related to the emotion, which can later be fed to the classification system. In the field of emotion recognition from speech, a variety of acoustic features have been investigated. However, the features extracted from speech signals can be divided into two most widely used groups: (1) **prosodic features**, which can be directly extracted from the signal itself, and (2)

spectral features, which are extracted after applying some mathematical transformations to the audio signal [56].

Prosodic features can be divided into three subclasses: *features related to pitch*, *energy*, and *temporal components*. Pitch, often referred to fundamental frequency, is the vibration rate of the vocal folds [57] . It is related to the vocal behavior and it is known as an effective emotion related feature. The features related to *pitch* can be presented by applying some statistical functions to pitch of speech signal, such as the values of maximum, minimum, mean, median, standard deviation, range and variance. *Energy-related features* are usually measured as the short term power of the speech signal and indicate the perceptual loudness of speech [57]. These features can also be presented by statistical functions. The last subclass of basic features uses the temporal characteristics of the considered speech signal, such as zero cross rate, rhythm, segment length and speaking rate [56]. The statistical properties of prosodic features carry important emotional cues and information. Table 2.4 lists the statistical properties of prosodic features (pitch, energy and speaking rate) for five basic emotions, with neutral being the reference point for comparison [57]. It can be observed from the table that the statistical properties of features for males and females may be different or even opposite for the same emotion.

Spectral features are spectral representation of speech signal in the frequency domain, therefore, providing useful information in additions to prosodic features. These features can be obtained by taking the discrete Fourier transform (DFT) of the speech signal. A classic example of spectral features is Mel-scale Frequency Cepstral Coefficients (MFCC), which is commonly reported as an effective feature for emotion recognition in literature [58-60]. In order to obtain MFCCs, a DFT is applied on windowed segments of the signal, the power spectrum calculated in the previous step is then mapped onto the N Mel-scale triangular shaped filters and converted to the logarithmic domain. Finally, discrete cosine transform (DCT) is applied to the log-energy values and the MFCCs are presented by the amplitudes of the resulting spectrum values. Another classic spectral features that have been widely used for emotion recognition are perceptual linear predictive coefficients (PLPCs). The PLPCs [61, 62] are obtained by using perceptual linear prediction. To calculate the coefficients, the power spectrum of the windowed speech signals are obtained and mapped to the Bark scale. Finally, the

mapped spectrums are processed by a cubic root compression, and approximated by all-pole model. The PLPCs are the resulting autoregressive coefficients.

Table 2. 4: Statistical properties of prosodic features for selected emotions [57].

Emotion	Pitch				Energy		Speaking Rate
	Average	Range	Variance	Contour	Average	Range	
Anger	>>	>	>>		>>M , >F	>	<M , >F
Disgust	<	>M , <F			=>		<<M ,<F
Fear	>>	>		↗	=>		
Joy	>	>	>	↘	>	>	
Sadness	<	<	<	↗	<	<	>M , <F

Explanation of symbols: '>', '<', and '=' mean an increase, a decrease, and no change compared to the neutral emotional state, respectively; double symbol indicates a considerable change; ↗ and ↘ stand for upward and downward inclines, respectively; subscripts *M* denotes male and *F* denotes female.

Source: It has been done by Sara Zhalehpour.

Besides the MFCCs and PLPCs, several other spectral feature have also been employed for speech emotion recognition in the literature, all reported to be useful but being employed less frequently as compared to the MFCCs e.g. speech formants and spectral measures. Formants are the concentrations of speech energy around a particular frequency shaping the spectrum. Usually position of the peak frequency, peak of amplitude and bandwidth of the first two or three formants are extracted as features [63-65]. Spectral measures are features extracted by measuring specific spectral characteristics, such as centroid, bandwidth, band energy, flux, and roll-off frequency [58, 65, 66].

2.4.2 Emotion Classification Methods

Many machine learning algorithms have been used for speech emotion classification, including support vector machines (SVM), neural networks (NN), *k*-nearest neighbor classifiers (KNN), discriminant analysis, decision tree naive Bayes (NB), Gaussian mixture models (GMM) and hidden Markov models (HMM). There is no definitive answer about which algorithm is the best one and every technique has its own strengths and weaknesses. But none can provide the best recognition performance under all situations. Hence, the selection should be based on the task at hand.

2.4.3 Speech Emotion Recognition Studies

Several works with different approaches has been done in the speech emotion recognition field. Comparing all these method is not practical due to the various emotions and also the difference of databases that have been used. Here, we will have a brief review of some of the works in this area.

Lee and Narayan [67] proposed an automatic speech emotion recognition by use of Linear discriminant classification (LDC) with Gaussian class-conditional probability distribution and k-nearest neighbors (K-NN) classifiers. These methods classified speakers into two basic emotional states, negative and non-negative. They used statistics of the fundamental frequency and energy of the speech signal as input features of these classifiers. To enhance performance of the classifiers, the dimension of the features were reduced while maximizing classification accuracy.

Vogt *et al.* [68] introduced a data-mining method for feature selection in automatic speech emotion recognition system. The most relevant features from more than 1000 features obtained from pitch, energy and MFCC time series, were selected from this feature set by removing correlated features.

Schuller *et al*. [69] presented a method for automatic speech emotion recognition by continuous hidden Markov models as the classifier. They introduced and compared two approaches. In the first approach, Gaussian mixture models was used to classify a global statistics framework of an utterance using extracted features of the raw pitch and energy contour of the speech signal. In the second approach they increased temporal complexity of the models by applying continuous hidden Markov models with several states using low-level instantaneous features instead of global statistics. Their methods (86 percent) gave promising results respect to human judgments (79.8 percent).

Lee and Narayan [64] again proposed method for emotion recognition in a call center based on a combination of three sources of information; acoustic, lexical, and discourse. Different feature sets were obtained followed by principal component analysis to accomplish the optimization of the acoustic correlation of emotion respect to classification error. Experimental results on their call center data exploited that the best accuracies were obtained by combination of acoustic and language information.

Mao *et al.* [70] proposed a method for speech emotion recognition using a hybrid of hidden Markov models (HMMs) and artificial neural network (ANN). They utilized both utterance and segment level information of speech. In their study the utterance was viewed as a series of voiced segments, and feature vectors extracted from the segments were normalized into fixed coefficients using orthogonal polynomial methods and distortions were calculated as an input of ANN. Meanwhile, the utterance as a whole was modeled by HMMs, and likelihood probabilities derived from the HMMs were normalized to be another input of ANN. The average recognition accuracy has reached 81.7 percent.

Ntalampiras *et al.* [71] studied three different methodologies namely, short-term statistics, spectral moments and autoregressive models for merging subsequent features. Additionally, they employed a set of parameters based on the wavelet decomposition. They evaluated fusion of these sets on the feature and log-likelihood levels. The classification stage is based on HMMs. They reported their results on the EMO-DB database for six emotional states.

Cheng *et al.* [72] presented an emotion classification method based on GMM for five basic emotions. They combined 60 basic features to form the feature vector. The features were extracted by PCA and sent into the improved GMM for classification. Results show that the selected features are robust and effective for the emotion recognition.

In Table 2.5 the results of some speech emotion recognition studies given above are summarized.

2.5 MULTIMODAL EMOTION RECOGNITION

One of the most important limitations of emotion recognition is that researches usually have focused on the recognition from only one modality. However, since the natural human-computer interaction is a multimodal process, using only one modality's observations lead us to uncertain results, which may be affected by other modalities' results. Therefore, integrating different modalities is a step towards a more realistic interaction. A perfect emotion recognition system would be the one which considers various modalities such as facial expressions, speech, body gestures and psychological reactions. Each modality provides information that complement each other.

Nonetheless, obtaining information from some modalities are not feasible or easy, such as the heart rate, which requires physical contact with the subject, thus, here, we narrow our focus on the recognition of emotion from audio-visual modalities.

Table 2. 5: A list of representative works in the field of audio emotion recognition

Researcher(s)	Features	Classification Method	Database	Recognition Rate
Lee *et al.* [67]	Pitch, Energy	LDC	Real users	77 %
Schuller *et al.* [69]	Pitch, Energy	GMM	Self-defined	86 %
Vogt *et al.* [68]	Pitch, Energy, MFCC	Naïve Bayes	Emo-DB + SmartKom	77.4 %
Lee *et al.* [64]	Acoustic features, Discourse information	LDC	Real users	Female : 40 % Male : 36.4 %
Mao *et al.* [70]	pitch, Energy, formant, LPCC, MFCC	HMM + ANN	BHUDES + Emo-DB	81.7 %
Ntalampiras *et al.* [71]	Temporal features	HMM	Emo-DB	91 %
Cheng *et al*. [72]	Pitch, MFCC	GMM	Self-defined	Female : 79.9 % Male : 89 %

Source: It has been done by Sara Zhalehpour.

2.5.1 Multimodal Fusion

The way of integration of emotions from various modalities is one of the biggest challenges in developing an efficient emotion recognition system. This is where multimodal fusion comes in handy. Multimodal fusion is a task of combining relevant information from multiple modalities. Three types of multimodal fusion strategies are usually applied for fusion of information from different modalities, namely **signal level fusion**, **feature level fusion** and **decision level fusion**. Figure 2.4 illustrates four possible multimodal fusion level approaches.

2.5.1.1 Signal level fusion

In this method the information signals of each modality are combined before extracting the features (Figure 2.4.a). This kind of combining information is usually not feasible to multimodal fusion due to the fact that the information must have same characteristics and be time synchronized. This method is instead useful for improving the recognition of one modality for example using an array of microphones for speech [73].

2.5.1.2 Feature level fusion

Feature level fusion is generally achieved by concatenating the extracted features from each modality and obtaining a joint feature vector to give to the classifier (Figure 2.4.b). This fusion method is the one humans use while obtaining information from various sources [74]. The feature level fusion is advantageous if multiple features from different modalities are uncorrelated or independent. Moreover, it only needs one classification phase on the joint features. However, it is criticized for not representing the differences in temporal structure of the multimodal features. Another drawback of this fusion technique is that the combined features must be represented in the same format before fusion. In addition, the increasing feature vector dimension makes fusion computationally more intense. Also, the increase in the number of modalities makes learning cross-correlation between modalities more difficult [75].

2.5.1.3 Decision level fusion

In this fusion strategy different modalities are analyzed individually. Then, the decisions from these modalities are combined (Figure 2.4.c). This method has many advantages over feature level fusion. For instance, there is no more synchronization issue between modalities. Also, for analyzing each modality the most suitable learning method can be used which provides more flexibility. Moreover, the decision level fusion offers scalability of the decision for each modality, which is hard to achieve in the feature level fusion methods [76]. Despite all the advantages of this method, it also has some disadvantages. One of the drawbacks is its inability to use the feature level correlation between modalities. In addition, using different classifiers for each modality for obtaining the local decision makes the learning process time consuming.

2.5.1.4 Hybrid multimodal fusion

To exploit the advantages of both feature and decision level fusion at the same time, several works have selected to use a hybrid fusion strategy of combination of these to level of fusion. An illustration of this approach is presented in Figure 2.4.d.

2.5.2 Multimodal emotion recognition studies

There are a number of studies which combine different modalities in literature. Below, we will give a brief overview of the approaches used in literature for multimodal emotion recognition.

Chen *et al.* [77] proposed a method for emotion recognition from joint audio-visual input of facial video and speech. The authors took into the account real time applications of their system, therefore, the recognition process was done on each frame. In the proposed system, there were three major modes of operation based on the presence of each modality. In the first mode, there is only audio signal for each frame. It happens when the user s face is out of camera's view or when it moves so fast that the video tracking is unreliable. In the second mode, the user does not speak, but there is a full view of the face. Finally, in the last mode, the speech signal and user's face are both available and the tracking is also reliable. The signal level fusion was used in the third mode for combining the features.

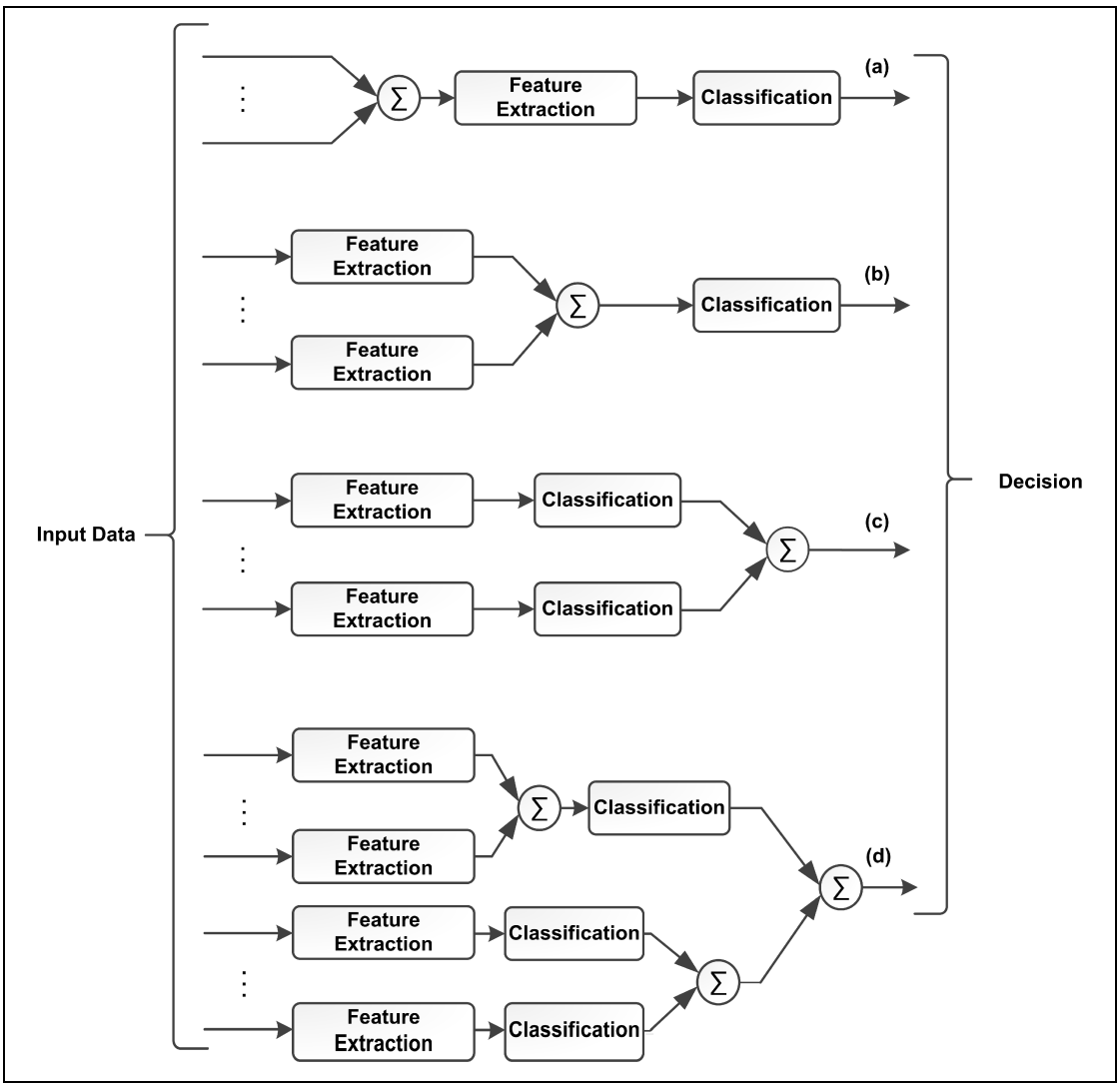
De Silva *et al*. [78] classified six basic emotions from both facial expressions and emotional speech and then combined the audio and video information using a rule-based system to improve the recognition rate. In their research the recognition rate for audio and video modalities were 32 percent and 62 percent, respectively and the overall results for both audio and video information were 72 percent.

Busso *et al.* [79] analyzed and compared decision level and feature level integration for fusing facial and speech modalities. In this research, they classified the emotions separately by capturing subjects' faces motion simultaneous speech data. Then they combined the outputs using two different methods of fusion and compared them with previous results. The outcome of their work showed that the fusion of the audio-visual information improved the performance of the system almost 5% compared to the facial expression recognition system alone. Also, the overall performance of the feature-level and decision-level bimodal classifiers were similar.

Chen *et al.* [80] proposed a method to combine both audio and visual features to extend the capability and performance of emotion recognition as compared to only single modal works.. They fused the information at feature level by using two ways of combination. In the first way, they combined the features directly by concatenating the

audio and visual information. In the other way, they tried to make the size of feature vectors of both modalities the same by duplicating the size of audio features since it had the lower dimension. There was only a slight difference in the performance of direct and balanced feature combination. However, the multimodal system outperformed both visual and audio results. Their experiments showed that recognition by only speech information had almost 63 percent accuracy and recognition for visual data gave almost 75 percent accuracy.

Figure 2. 4: Four basic fusion methods used in current multimodal emotion recognition systems. ∑ is not just symbol of simple adding. (a) Signal level (b) Feature level (c) Decision level and (d) Hybrid fusion [76].



Source: It has been done by Sara Zhalehpour.

Paleari and Lisetti [73] presented a framework for multimodal emotion recognition that could accept new recognition modules based on Scherer theories. The framework

almost automatically took into account the new recognition system and used it to enhance the emotion recognition results. It also used a buffer system to store the information from different modalities and then analyzed different possible fusion approaches and automatically chose the most stable one during a training phase.

In [81] Schuller and Wimmer proposed an audio-visual emotion recognition system that used the feature level fusion. Audio and video features were first derived as Low-Level-Descriptors. Synchronization and feature space combination was achieved by multivariate time-series analysis. Since the time interval for each sequences can be different they used the descriptive statistical analysis to obtain features with same dimension for feeding to the classifiers. Their overall recognition rate showed an improvement in audio-visual recognition rate with respect to the each modality's accuracies

Mansurizadeh *et al.* [82] introduced an asynchronous hybrid fusion approach which used both feature and decision level fusions. This approach was based on the fact that cues from facial image series and audio information of an audio-visual sequence are not temporally aligned, hence, features from audio and video modalities that are related to the same emotional event have more chance to be temporally overlapped and therefore should be fused together.

Gajsek *et al.* [83] proposed an audio-visual emotion recognition system, which used prosodic and cepstral coefficients as audio features and Gabor wavelets as video features followed by feature selection using a stepwise method. He used a multi-class classifier system to combine the outputs of the different classifiers. Datcu *et al.* [84] presented a multimodal semantic data fusion model. Their method used two types of geometric features for face depending on the presence or absence of speech. It also removed the influence of the speech on the face shape by using only the eye and eyebrow related features.

A summary of all mentioned fusion researches described above is provided in Table 2.6. As can be seen from the table, fusion of the modalities improved the recognition accuracy in all cases. However, the improvement rate is not the same for all cases. The reason can be due to the fact that each work used a different database with dissimilar sizes and also the various type of fusion method that have been used.

Table 2. 6: A list of representative works in the field of audio-visual emotion recognition

Researcher(s)	No. of Subjects	Fusion Level	Fusion Method	Recognition Rate		
				Audio	Video	Audio-visual
Chen *et al.* [77]	100	Signal	Selective Combining of Features	63 %	58 %	-
De Silva *et al.* [78]	2	Decision	Ruled Based	62 %	32 %	72 %
Busso *et al*. [79]	1	Feature/ Decision	Combination of Features/ Classifiers	71 %	85 %	89 %
Chen *et al.* [80]	2	Feature	Combination of Features	63 %	75 %	84%
Paleari and Lisetti [73]	-	Feature/ Decision	Combination of Features/ Classifiers	-	-	-
Schuller and Wimmer [81]	8	Feature	Combination of Features and Statistical Analysis	74%	61 %	81 %
Mansurizadeh *et al.* [82]	42	Hybrid	Asynchronous combination of Features/ Classifiers	33 %	37 %	71 %
Gajsek *et al*. [83]	-	Decision	Combination of Classifiers	63 %	55 %	71 %
Datcu *et al*. [84]	42	Decision	Combination of Classifier	56 %	38 %	56 %

Source: It has been done by Sara Zhalehpour.

3. MULTIMODAL EMOTION RECOGNITION SYSTEM

In this chapter, we present a new framework for multimodal emotion recognition using audio and visual channels of an expressive video. An emotional video consists of hundreds of frames, where the emotion is expressed with different intensities at different frames. Therefore, when it comes to emotion recognition from a video, it is a challenge to decide how to use these frames so that the facial expression recognition rate is maximized. One of the promising approaches is utilizing a frame or a subset of frames in the video, which represent the emotional content of the sequence best. These are frames at which the emotional expression is at its peak (i.e. at apex). Hereafter, we will refer to such frames as "peak frames". This approach is based on the assumption that there is a single emotion expressed in the video. The visual features of a video are extracted based on the fames at which the facial expression is at its apex.

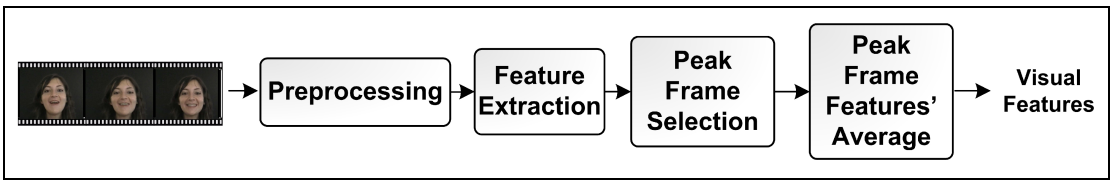
We propose three methods for automatically selecting peak frames from an emotional video, which is a problem that has been addressed by a very few researchers so far. The three methods we presented below are a method based on the dissimilarity between frames, a clustering based method and an emotion intensity based method.

Finally, the classification and multi-modal decision level fusion methods are described. In this thesis decision level fusion methods are chosen for their advantage respect to the feature level fusion methods, such as scalability and no need for time synchronizing between the modalities.

3.1 FEATURE EXTRACTION FROM VIDEO

In this thesis, we use appearance-based features of the face as our visual feature extraction method. In Figure 3.1, the overview of facial feature extraction process from a video is shown.

Figure 3. 1: General framework of visual feature extraction system

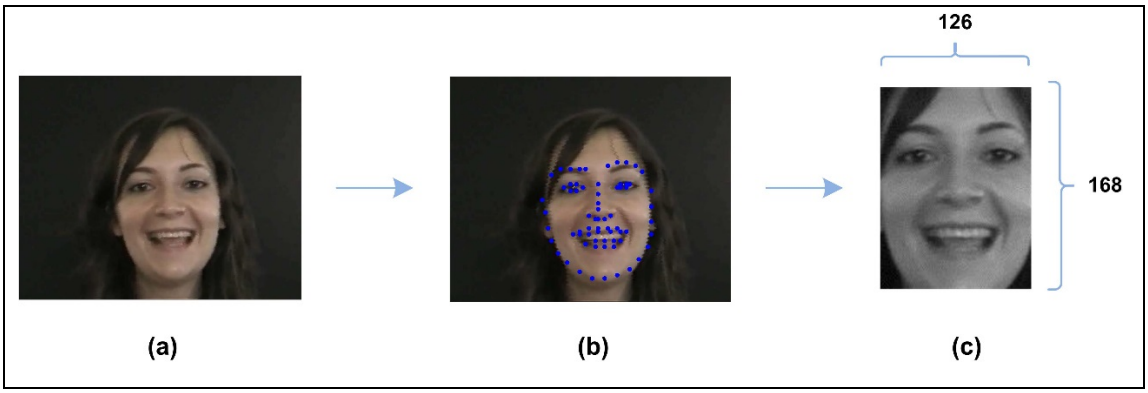


Source: It has been done by Sara Zhalehpour.

3.1.1. Preprocessing

As the first step in visual feature extraction, we need to detect the face in all frames of the video sequence and then align and crop these frames to eliminate unnecessary regions such as background and hair. First, we apply the Zhu's [85] face tracker to automatically detect the locations of the eyes resulting in 68 landmarks, for all frames. The eyes sub-model tracks the eyes using six landmark points around the eye. The coordinates of these six landmark points around each eye are averaged to give a central point location for each eye. In order to compensate for any scale differences between frames, images are scaled to obtain an inter-ocular distance of 64 pixels. Then, images are aligned based on eye localization and cropped in a way such that the face region has a size of 168×126 (see Figure 3.2).

Figure 3. 2: (a) Face image (b) Landmark point extraction (c) Aligned, scaled and cropped.



Source: It has been done by Sara Zhalehpour.

3.1.2 Local Phase Quantization Features (LPQ)

The blur insensitive Local Phase Quantization operator was originally proposed by Ojansivu and Heikkila [86] for texture classification. In this thesis, the LPQ operator is used to construct a face descriptor, which has recently been shown to be successful for emotion recognition [87-89]. Below we give the details of the computation of LPQ features.

Let us assume that we are given an image degraded image, on which we want to calculate a texture descriptor, i.e. in the frequency domain let $G(u) = F(u)H(u)$, where , $G(u)$, $F(u)$ and $H(u)$, are the Fourier transforms of the degraded image, the

original image and the point spread function (PSF) of the degradation function respectively. If a PSF, $h(x)$ is centrally symmetric, its Fourier transform $H(u)$ is always real valued and the phase angle $\angle H(u)$ is either 0 or π. First, local M-by-M neighborhoods N_x at each pixel position x of the image $f(x)$ are used to examine the phase. The local spectrum is obtained by a 2D short term Fourier transform of a window W around pixel x.,

$$F(\mathbf{u},\mathbf{x}) = \sum_{y \in W} f(\mathbf{x}-\mathbf{y}) e^{-j2\pi \mathbf{u}^T y} = \omega_{\mathbf{u}}^T f_{\mathbf{x}} \tag{3.1}$$

where ω_u denotes the basis vector at frequency $\mathbf{u}$ and f_x denotes the vector containing the values of all M^2 image samples, which come from N_x.

Then, four samples of $F(\mathbf{u}, \mathbf{x})$ are considered at the frequencies $\mathbf{u}_1 = [a, 0]^T$, $\mathbf{u}_2 = [0,a]^T$, $\mathbf{u}_3 = [a,a]^T$, $\mathbf{u}_4 = [a,-a]^T$, where a is a small frequency at which the FT of the blur function is positive ($H(\mathbf{u}_1) > 0$).

$$F_x = \left[F(\mathbf{u}_i, \mathbf{x}) \right],\ i = 1,2,3,4 \tag{3.2}$$

$$G_x = \left[Re\{F(\mathbf{u}_i, \mathbf{x})\}, Im\{F(\mathbf{u}_i, \mathbf{x})\} \right],\ i = 1,2,3,4 \tag{3.3}$$

The coefficient vectors are decorrelated using a whitening transform assuming that the image is a first order Markov process. This is done since it is known that information is preserved better in quantization if the quantized coefficients are statistically uncorrelated. After decorrelation, each element is quantized using a simple scalar quantizer to solve it.

$$q_i(\mathbf{x}) = \begin{cases} 1, & if\ g_i(\mathbf{x}) > 0 \\ 0, & otherwise \end{cases} \tag{3.4}$$

where $g_j(\mathbf{x})$ is the jth component of $G_{\mathbf{x}}$ after decorrelation.

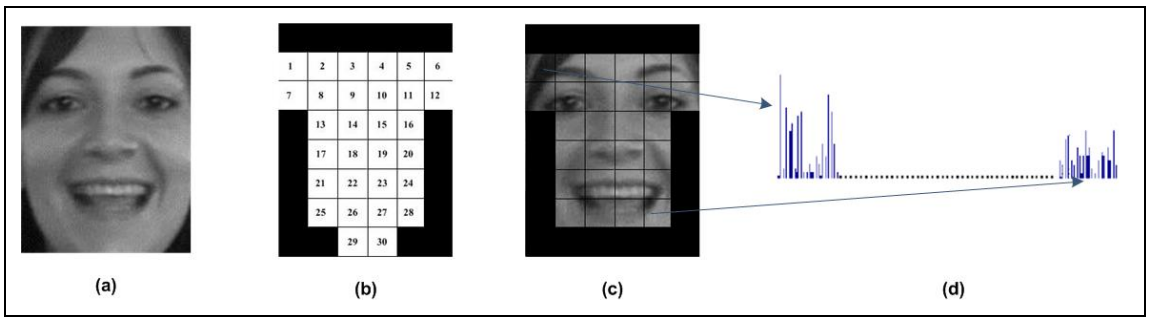
The resulting eight binary coefficients $q_j(\mathbf{x})$, $j = 1,\ldots,8$, are represented as integer values between 0-255 using binary coding as follows:

$$f_{LPQ}(x) = \sum_{j=1}^{8} q_i(\mathbf{x}) 2^{j-1} \tag{3.5}$$

For more details about the algorithm, the reader is referred to [86].

We compute the LPQ featues on the face image as follows. Since many regions of the face such as the upper forehead, outer sections of cheek and jaw do not carry information about the facial expression, we divide the face image into sub-blocks of size 8×6= 48 and discard 18 sub-blocks that are irrelevant as shown with black blocks in Figure 3.3(b). In the remaining (relevant) blocks, we extract the 256 bin histogram of LPQ features of each block. . The LPQ features of the 30 sub-blocks are concatenated into a long vector of histogram to form a single feature vector shown in Figure 3.3(d). The final histogram of the whole image is used as the facial expression feature. We used the implementation available from [90] and used it with default parameters (window size is 3 × 3, DFT is calculated using a uniform window, $a = 0.7$).

Figure 3. 3: (a) Cropped image (b) 30 used subregions for feature extraction (c) The considered subregions for feature extraction (d) concatenated histograms extracted from each 30 subregions.



Source: It has been done by Sara Zhalehpour.

3.1.3 Automatic Peak Frame Selection

We will calculate the visual features of a video based on peak frames. Therefore, peak frame have to be selected automatically. . Below we propose three novel methodologies for automatically selecting peak frames from an emotional video, which is an issue that has been addressed by a very few researchers so far [91, 92]. Yongjin *et al.* [91] selected the peak frames as the frames with the highest speech amplitude. Meghjani *et al.* [92] used a semi-supervised clustering technique to divide the frames into two major clusters; namely, the emotion frames and non-emotion frames. The cluster with the highest number of continuous frames is considered to be the emotion cluster.

In this work, the following three methods for peak frames selection based on visual features are proposed, MAXDIST, DEND CLUSTER and EIFS. Following that another

peak frame based on audio features (AFS) is presented. In all proposed peak frame selection methods we use LPQ features from all frames of each sequence, which will also be used later as peak frame features for emotion classification.

3.1.3.1 Peak frame selection based on maximum dissimilarity (MAXDIST)

This method of peak frame selection is based on the assumption that candidate peak frames are maximally dissimilar from neighbor frames. Therefore, first the dissimilarity between successive frames is computed by comparing the facial representation features. The method sorts the frames based on their average dissimilarity score with other frames, and selects those frames (peak frames) that correspond to the K largest average dissimilarity scores. We refer to this method as MAXDIST since peak frames are selected using a maximum dissimilarity criteria.

Let $\bar{V}$ be a video sequence and let $F = \{fe_1, fe_2, fe_3, \ldots, fe_N\}$ denote the LPQ features of N frames in the video $\bar{V}$. The steps of the proposed peak frame selection algorithm are as follows:

Step 1: An $N \times N$ dissimilarity matrix, M, is generated, where each element $M(i,j)$, $i, j \in \{1, 2, \ldots, N\}$ is the chi-squared distance score between LPQ features of frames i and j.

Step 2: For the jth frame, the average distance score, d_j, with respect to the other $(N - 1)$ frames is calculated, by finding the average of the elements in jth row of M.

Step 3: The average values obtained in step 2 are ordered in descending order and the top K frames that have the largest average distance scores are selected as peak frames, since they are the most "dissimilar" frames in the video.

The choice for the value of K is application dependent. In this thesis, we used $K = 6$ peak frames for each sequence. Figure 3.4 depicts the procedure of the proposed MAXDIST peak frame method.

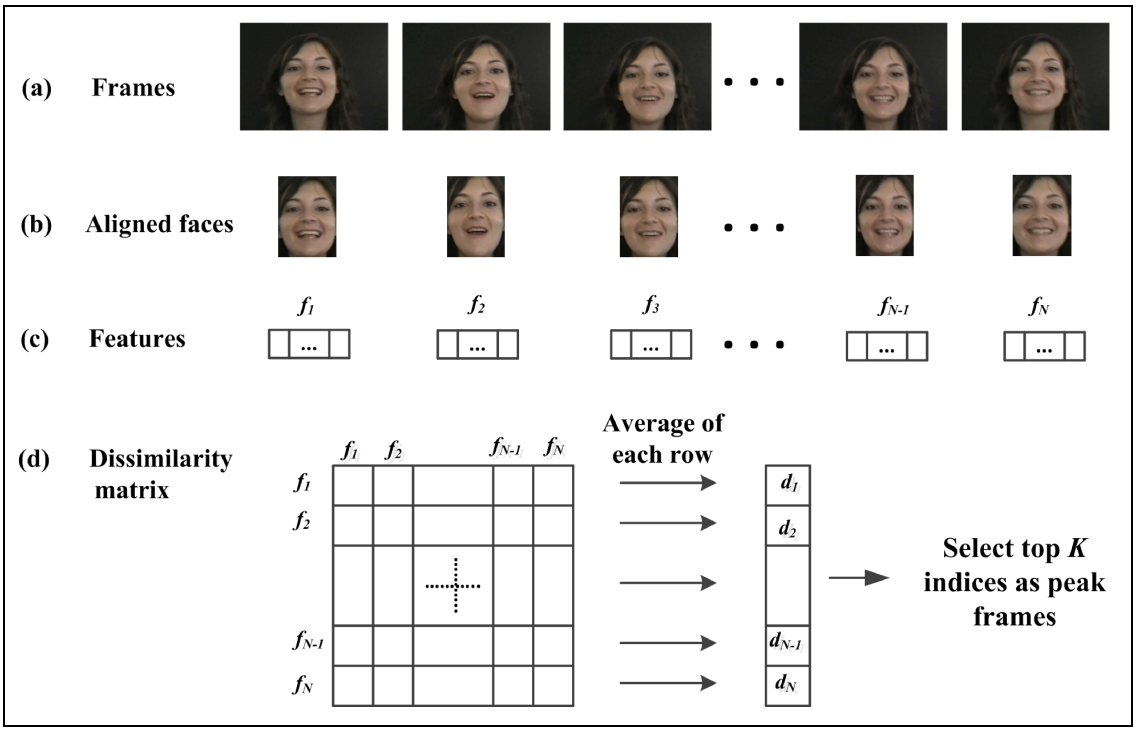
3.1.3.2 Clustering based peak frame selection (DEND CLUSTER)

In this method, the N frames are grouped into K clusters, in a way that frames of each cluster are more similar to each other than frames from the other clusters. Then for each cluster, a representative frame that represents all members of that cluster is chosen,

resulting in *K* peak frames. The above method is referred as DEND CLUSTER since it uses the dendrogram clustering method to choose the peak frames.

To perform clustering, it is first required to compute the dissimilarity scores between frames. First thus the dissimilarity between each pair of frames is computed by comparing the facial representation features. The comparison is done based on the Chi-squared histogram distance method. We then use hierarchical clustering [93], since our representation of the *N* frames is in the form of a $N \times N$ dissimilarity matrix instead of a $N \times d$ pattern matrix (*d* is the number of features). In particular, we use an agglomerative complete link clustering algorithm [94]. The output of this algorithm is a binary tree dendrogram, where each terminal node corresponds to a frame, and the intermediate nodes show the structure of clusters, as also illustrated in Figure 3.5.

Figure 3. 4: Overview of proposed peak frame selection. (a) Face detection and alignment, (b) The face region is cropped , (c) Extraction of the LPQ features for each frame and (d) Calculation of the dissimilarity matrix and selection of the peak frame based on the top K average distance scores.



Source: It has been done by Sara Zhalehpour.

The *K* peak frames are selected as follows:

Step 1: Find the pair-wise distance score between the *N* frames to form dissimilarity matrix *M*.

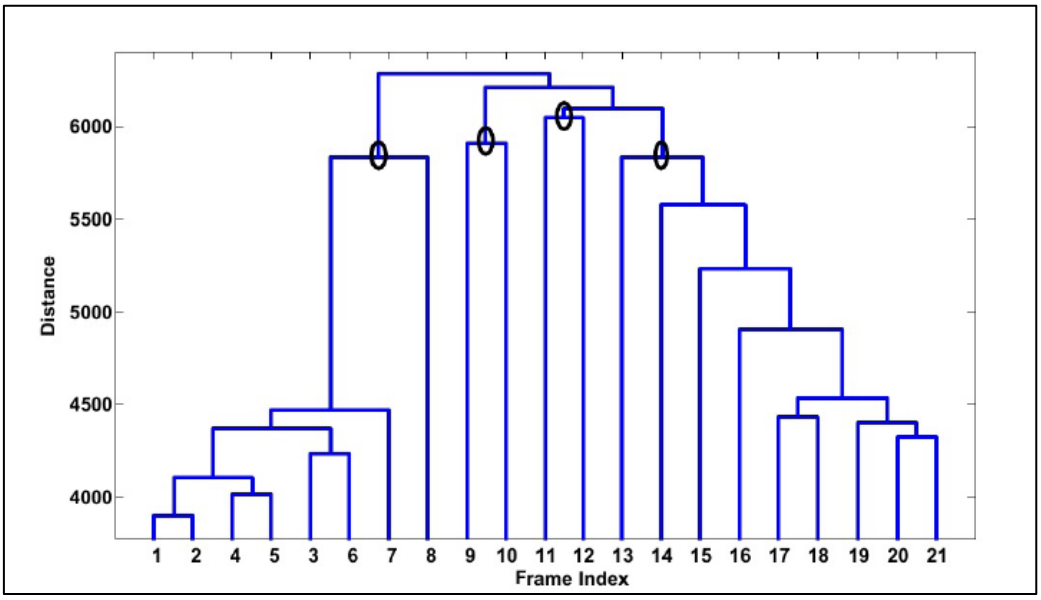
Step 2: Generate the dendrogram, *D* by applying the complete link clustering algorithm on *M*. Use the dendrogram *D* for identifying *K* clusters.

Step 3: In each of the clusters identified in step 2, select a frame whose average distance from the rest of the frames in the cluster is minimum. If a cluster has only 2 frames, choose any one of the two frames at random.

Step 4: The frames selected in step 3 make the peak frames set.

In step 2, DEND CLUSTER method automatically sets the threshold distance to cut the dendrogram and detects *K* clusters. For instance, for the dendrogam given in Figure 3.5, this distance is determined to be 6700. The choice for the value of *K* for this method is 6, means, we used $K = 6$ peak frames for each sequence.

Figure 3. 5: Dendrogram generated using the 21 × 21 dissimilarity matrix of a video consisting of 21 frames. The circles on the subtrees indicate cluster formations for $K = 4$ [93].



3.1.3.3 Emotion intensity based frame selection (EIFS)

This method is based on estimation the unknown neutral face for a given expressive face. Then, the difference between the feature vectors of the expressive face and the neutral face is used as the "emotion intensity". The frames which give the highest "emotion intensity" are selected as peak frames.

One important attribute of the proposed algorithm is the elimination of the requirement of a neutral expression frame of the same subject whose expression is to be recognized since it is applicable to any unknown subject. In other words, the method is completely subject independent.

The neutral face estimation method applies Karhunen-Loeve transform [95] to formulate a *neutral subspace* by eigenvector decomposition of neutral images from all subjects containing the variations present in neutral frame space. The neural frames are collected from other databases which contain neutral faces such as the Cohn-Kanade database. Whenever, any expressive face image is projected onto this subspace it is expressed as a linear combination of Eigenfaces from a space of neutral frame images. This subspace thus can be exploited to extract the neutral frame information present in an expression containing frame. The proposed methodology employs the property of eigenvector decomposition for synthesis of a virtual neutral frame of the subject whose frame with some expression is given. Once, the virtual neutral frame image is obtained for the sequence, it is subtracted from the given expression containing frame to estimate the emotion intensity.

The steps of the proposed peak frame selection algorithm are as follows:

Step 1: First we construct the neutral subspace. Let $A = [\Phi_1 \quad \Phi_2 \quad \dots \quad \Phi_M]$ be the matrix containing the neutral faces $\Phi_i, i = 1, \dots, M$ in its columns. If the image has a size of $N \times N$, then the size of Φ_i is $N^2 \times 1$ and the size of matrix A is therefore $N^2 \times M$. The covariance matrix is formed as $C = AA^T$ and the eigenvalues λ_i and eigenvectors $e_i, i = 1,2, \dots, M$ of C are estimated using a computationally efficient method as in [91]. Note that since there are M columns in A, there are at most M non-zero eigenvalues of the covariance matrix C [96]. We project all the neutral face images into the neutral subspace to obtain the vector of weight vectors $n_i, i = 1, \dots, M$, where n_i has a size of $M \times 1$.

Step 2: Given a sequence S with F frames $f_k, k = 1, \dots, F$ reflecting a single expression with different intensities at each frame, we want to select a single neutral image from the set of neutral images Φ_i, that represents all the frames in the sequence in the best way.

First, we project each frame f_k into the neutral subspace to obtain the corresponding weight vectors $t_k, k = 1, \ldots, F$. Then, we select the closest neutral image for each expressive frame by minimizing the following Euclidean distance:

$$i_k^* = \min_i t_k - n_i, i = 1, \ldots, M \tag{3.6}$$

The neutral frame which has been selected the most over all frames is selected as the single neutral frame "closest" to the whole sequence. Let us denote that neutral frame as n_S.

Step3: Determine the peak frame(s) of the sequence based on the Euclidean distance from the neutral face. That is we calculate

$$d(k) = t_k - n_S, \text{ for } k = 1, \ldots, F. \tag{3.7}$$

We order *d(k)* in descending order and select the top *K* frames as the peak frames of the sequence, since they are expected to be the highest emotion intensity be being the "farthest" from the neutral image.

3.1.3.4 Audio based frame selection (AFS)

The last method we present for peak frame selection uses the audio channel which is based on the heuristic such that peak frames occur when energy of the speech signal is at its maximum [91].

The speech energy of the video is calculated over 20 msec rectangular frames with an overlap ratio of 50 percent. Then, peak frames are selected as the ones with energy more than a threshold ε which is equal to the difference of the maximum energy of the sequence and one hundredth of it:

$$\varepsilon = Energy_{max} - (\frac{Energy_{max}}{100}) = 0.9 Energy_{max} \tag{3.8}$$

The frames corresponding to these maximum time instants are selected as peak frames.

Figure 3.6 illustrates the outcome of the above three methods for an example sequence from the eNTERFACE'05 database. We can see that all the selected six peak frames for the happiness emotion reflect the emotion happiness at its apex.

Figure 3. 6: Six selected peak frames for an example sequence from eNTERFACE'05 dataset by using the proposed method and the manual selection are shown for subject 7 and emotion happiness. Selected peak frames reflect the Happiness emotion at its apex.

Source: It has been done by Sara Zhalehpour.

3.1.4 VISUAL FEATURES

As our video features, we utilize the LPQ features which were previously extracted for peak frame selection. The values of LPQ features for all peak frames are averaged to get a single feature vector for the whole sequence and this value is used as the input of our classification system.

3.2 FEATURE EXTRACTION FROM AUDIO

Figure 3.7 illustrates the general flowchart of automatic speech emotion recognition system used in this study. The system consist of two processing stages. The first stage is preprocessing including the silence removal and windowing. In the second stage two audio spectral features are extracted from the processed speech signal.

3.2.1 Preprocessing

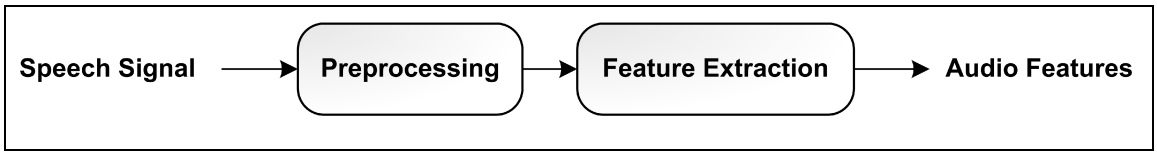
3.2.1.1 Silence removal

Since the database contains sentences and silent intervals within the sentences do not carry any useful information, those silent potions including the leading and trailing edges are eliminated by soft-thresholding the energy of short windows of the signal. The sampling signal energy was normalized for each frame with a length of 50 samples and no overlap as follows:

$$\bar{E}[k] = (E[k] - \tilde{E}[k])/\sigma \tag{3.9}$$

Where k is the frame number, $\tilde{E}[k]$ is the mean value of energy of all frames and σ is the corresponding standard deviation. The windows with normalized energy less than zero were removed and the remaining speech samples were concatenated.

Figure 3. 7: The structure of the speech emotion recognition system.



Source: It has been done by Sara Zhalehpour.

3.2.1.2 Windowing

The spectral analysis method is only reliable when the signal is stationary. Speech signals like any other audio signal are highly non-stationary, however; vocal tract can be considered stable over a very short period of time, typically around 10-30 msec. A signal $x(n)$ is divided into a succession of windowed sequences called frames. Each of these speech frames can then be processed individually by multiplying with a hamming window. The signal samples are segmented into frames of 25msec with 50 percent overlap ratio between consecutive frames.

3.2.2 Audio Features

A set of acoustic features have been experimentally investigated in this thesis including some prosodic features (zero crossing rate, pitch and energy) and spectral features

(MFCCs, Rasta-PLPs, Discrete wavelet coefficients, PMCCs and weighted mel-frequency cepstral coefficients). Based on the experimental results, combination of MFCCs and RASTA-PLP features have been chosen based on their superior performance. The MFCC and RASTA-PLP features are explained below.

3.2.2.1 MFCC Features

Mel Frequency Cepstral Coefficients (MFCCs) analysis is a speech feature extraction method that consider the human perception sensitivity with respect to frequencies. The Mel scale frequencies are distributed logarithmically in the higher range but linearly in the lower range. This is similar to the physiologic characteristics of the human ear [97]. The Mel Frequency Cepstral coefficient is one of the most common techniques for feature extraction and a quantitative representation of speech. The MFCC analysis can be described in the form of mathematical equation as following,

$$c(n) = DCT(\log(\left|FFT\left(s(n)\right)\right|)) \tag{3.10}$$

Where s(n) represents the speech window, and c(n) represents the MFCC coefficients. Figure 3.8 illustrates the overall procedure to generate the MFCCs from input signals after the preprocessing step. Each step is discussed briefly below [97]:

Fast Fourier Transform (FFT) - Since the original signal is in time-domain, FFT is used to transform the windowed signal into frequency domain.

Mel Filter Bank Processing - The frequencies range in FFT spectrum is very wide and speech signal does not follow the linear scale. In the Mel frequency filtering step, a set of band-pass filters is utilized to filter the frequency dependent segregated signals. Then, each filter output is the sum of its filtered spectral components.

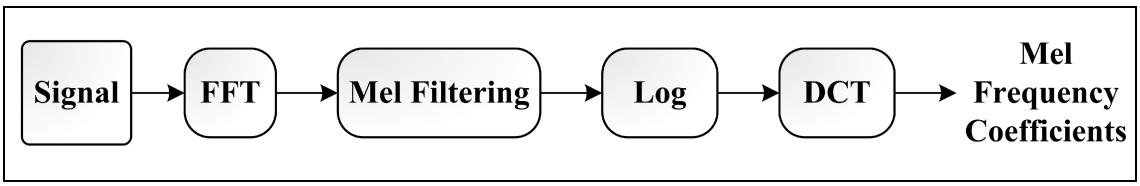
Logarithm - The Mel-frequency is obtained from equation (3.10) by taking the logarithm of given frequencies in *Hz*.

$$f_{Mel} = 2595 \times log_{10}(1 + \frac{freq}{700}) \tag{3.11}$$

Discrete Cosine transform (DCT) - The Discrete Cosine transform converts the Mel spectrum coefficients to coefficients in the time domain since the coefficients from

MFCC analysis are real numbers. The result of the conversion is called Mel Frequency Cepstrum Coefficient.

Figure 3. 8: The structure of MFCCs



Source: It has been done by Sara Zhalehpour.

The first 12 MFCCs by using a filter with order 12 is generated as MFCC features. As a common practice, delta and double-delta MFCCs (sometimes referred to as first and second derivatives, respectively) are calculated as well to capture local dynamics, forming a 36-dimensional feature vector. Then, nine statistical functions, namely, maximum, minimum, maximum position, minimum position, mean, variance, range, kurtosis and skewness are applied to the first 12 MFCCs and their first and second time derivatives in order to generate 12x3x9 = 324 MFCC features.

3.2.2.2 RASTA_PLP Features

RASTA-PLP (Relative Spectral Perceptual Linear Predictive) method is an enhancement of the traditional PLP (Perceptual Linear Predictive) method which makes PLP more robust to linear spectral distortions. This method replaces the conventional critical-band short-term spectrum in PLP and introduces a less sensitive to slowly changing or steady-state factors in speech. For the RASTA-PLP features, an additional filtering is used after decomposition of the spectrum into critical bands [98]. This RASTA filter control the low modulation frequencies which are supposed to stem from channel effects rather than from speech characteristics.

The steps for extraction of RASTA-PLP for each frame are the following:

1) Transform the windowed signal into frequency domain.
2) Calculate the critical-band power spectrum.

3) Take its logarithm and approximate the temporal derivative of log critical-band spectrum using regression line through five consecutive spectral values. Transform spectral amplitude through a compressing static nonlinear transformation.

4) Filter the time trajectory of each transformed spectral component.

5) Transform the filtered speech through expanding static nonlinear transformations using a first order IIR system.

6) Multiply by the equal loudness curve and raise the power 0.33 to simulate the power law of hearing.

7) Take the inverse logarithm of this relative log spectrum, yielding a relative auditory spectrum.

8) Compute an all-pole model of the resulting spectrum by taking inverse furrier transform and using Levinson-Durbin recursion technique.

In general case, the whole derivative-reintegration process is equivalent to the band pass filtering of each frequency channel through the IIR filter with transfer function denoted by equation (1)

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})} \tag{3.12}$$

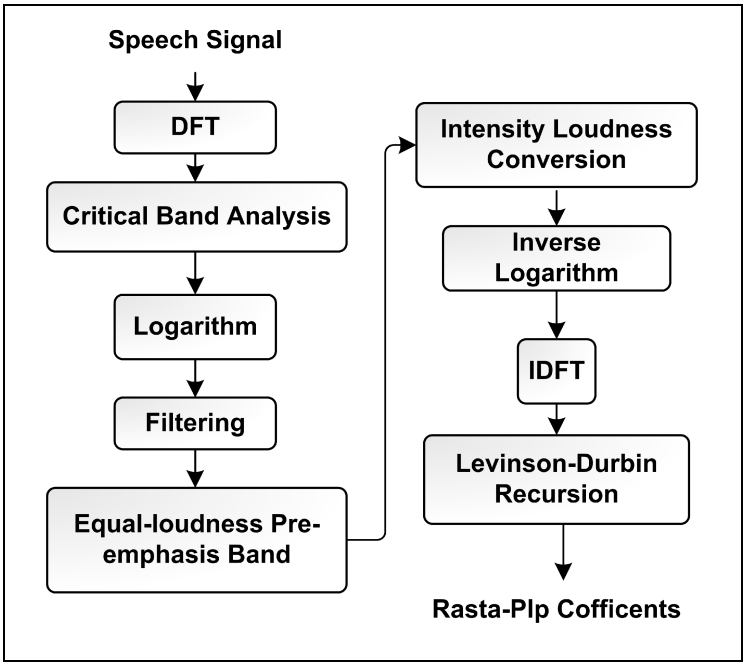
The low cut-off frequency of the filter determines the fastest spectral change of the log spectrum which is ignored in the output, while the high cut-off frequency determines the fastest spectral change which is preserved. The whole process of RASTA-PLP feature extraction is illustrated in Figure 3.9.

The 13 RASTA-PLP coefficients are calculated by using filter of order 20 and are augmented by their delta and double delta features. The same statistical parameters as used for MFCCs are then calculated for the RASTA-PLPs and their deltas, giving 13x3x9 = 351 -dimenstion feature vector. The MFCC and RASTA-PLP feature vectors are then concatenating in order of having the audio feature vector of dimension 324 + 351 = 675.

3.3 CLASSIFICATION

In recent years, support vector machines (SVMs) have been widely used for analyzing data and recognizing patterns. SVMs are based on the structural risk minimization principle, closely related to the regularization theory. The binary SVM tries to construct an *N*-dimensional hyperplane in a multidimensional vector space that can optimally separate vectors that belong to two classes. A good separation is obtained by the hyperplane that has the largest distance to the nearest training vectors of each class. The two-class SVM method can be expended to a multi-class problem. It is usually done by reducing the single multi-class problem into multiple binary classification problems. Each of the binary classifier is assumed to produce an output function that gives relatively large values for examples from the positive class and relatively small values for examples belonging to the negative class.

Figure 3. 9: The flow diagram for calculation of RASTA-PLP speech features.



Source: It has been done by Sara Zhalehpour.

In this thesis, we used an SVM classifier implemented in the LIBSVM toolbox [99] to classify each of the audio and video features. In order to classify the audio features, we used an SVM classifier with a radial basis kernel function and one-against-all method. Before classification, we normalized the numerical values of audio features to the

interval [0, 1] [75] to prevent features with large numeric values dominate features with small numeric values during classification. For the classification of the video features, we used an SVM classifier with a linear kernel to avoid the curse of dimensionality problem, since the dimension of the features is high (i.e. 7680).

3.4 MULTIMODAL FUSION

Examining and fusing the modalities at the fusion level enable us to gain intuition about how multimodal cues interplay during emotional expression and to incorporate in-domain information in the decision process [100]. Decision level fusion focuses on the combination rules for the outputs of several classification models. The available feature set is divided into subgroups (e.g. one classifier per modality) and the partitions are used to form classifiers. The outcomes of these slim classifier models are considered for the final decision making process. The term decision-level fusion sums up a variety of methods designed in order to merge the decisions of classifiers into one single decision. We utilized several algebraic decision level techniques where the output is the mathematically computed final decision of the output label from each classifier [76]. They will be described briefly below. In Figure 3.10 we illustrate the overview of the proposed multimodal framework.

The Maximum Rule and Minimum Rule - These techniques simply choose the minimum or maximum of the conditional probabilities derived from classifier *i*. The decision for an observed sample *x* is chosen to be the class ω^* for which has the largest probability between all *K* probabilities as shown below,

$$P(\omega_k \mid x) = \underset{i=1,2}{Max \text{ or } Min} \{P(\tilde{\omega}_k \mid x, \lambda_i)\} \ , k = 1,..., K \tag{3.13}$$

$$\omega^* = \max_k \{P(\omega_k \mid x)\}, \quad k = 1,..., K \tag{3.14}$$

Where x = { $x^{\lambda=1}$, $x^{\lambda=2}$} is a set of future from both audio and video channels, ω is the true class label and $\tilde{\omega}$ is the predicted output label. Finally, $P(\tilde{\omega}_k | x, \lambda_i)$ is the probability of the test vector x to belong to each class $\tilde{\omega}_k$ for each individual classifier λ_i.

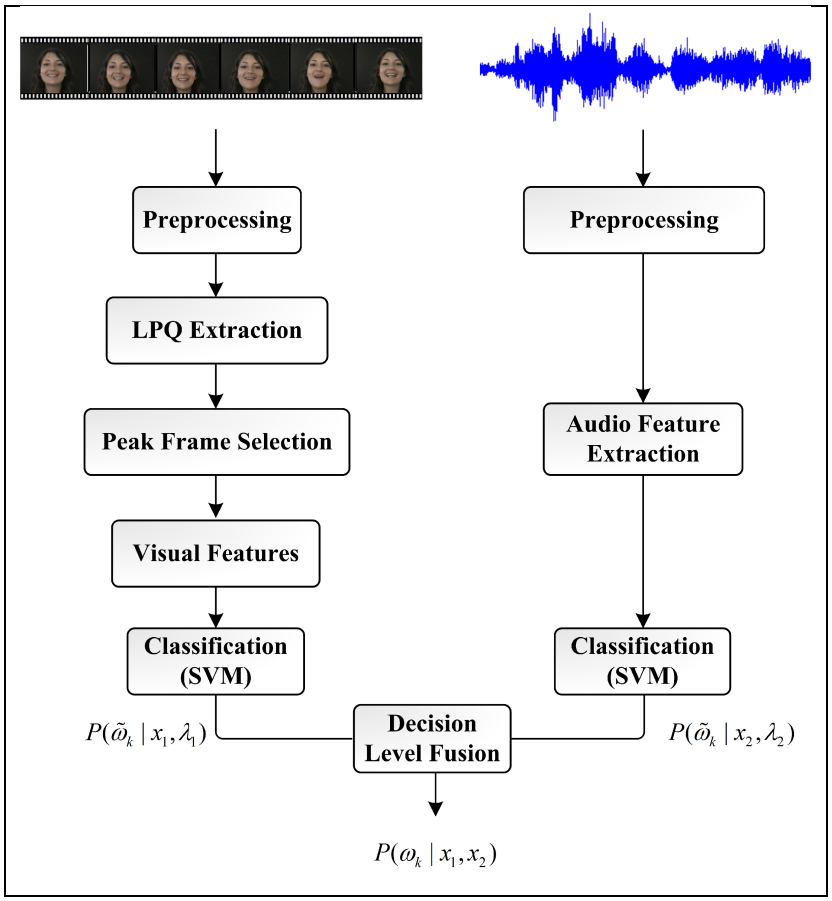
Sum Rule, Mean Rule and Weighted Average Rule - the sum rule sums up the probabilities given to each class in order to generate total probabilities of both classifiers. By averaging the probabilities ($\frac{1}{i}$ serves as normalization factor) given to each class, we obtain the Mean Rule. By additionally adding classifier weights $P(\lambda_i \mid x)$, we would have the weighted average method.

$$P(\omega_k \mid x) = \frac{1}{2}\sum_{i=1}^{2}[P(\tilde{\omega}_k \mid x, \lambda_i)P(\lambda_i \mid x)] \quad , k = 1,\ldots, K \tag{3.15}$$

$$\omega^* = \max_k \{P(\omega_k \mid x)\}, \quad k = 1,\ldots, K \tag{3.16}$$

Where $P(\lambda_i|x)$ is equal to the weight assigned to i-th classifier in the combination and controls the influence of each channel on the fusion results.

Figure 3. 10: An overview of the proposed multimodal emotion recognition system.



Source: It has been done by Sara Zhalehpour.

Product Rule - In the product rule the probabilities obtained from the classification of each modality is multiplied for given a test vector and we predict the final label as the one which gives the maximum product as follows:

$$P(\omega_k \mid x) = \prod_{i=1}^{2} [P(\tilde{\omega}_k \mid x, \lambda_i)]^{P(\lambda_i|x)} \quad , k = 1,..., K \tag{3.17}$$

$$\omega^* = \max_k \{P(\omega_k \mid x)\}, \quad k = 1,...,K \tag{3.18}$$

Bayesian Framework - Bayesian method is based on combining the conditional probability derived from the confusion matrix of the validation subset to weight the output of each classifier before fusing them for the final decision. Both uni-modal emotion recognition systems provide the confusion matrix on a validation subset and per class conditional probability distributions for each test sequence. These conditional probabilities are combined based on a Bayesian weighting framework. Then using the probabilities for each class a decision is made for the overall emotional state of the each sequence. The output of the fusion system can be expressed in terms of a marginal distributions as below [101],

$$P(\omega_k \mid x) \approx \sum_{i=1}^{2} [\sum_{k=1}^{K} P(\omega_k \mid \tilde{\omega}_k, \lambda_i) P(\tilde{\omega}_k \mid x, \lambda_i)] P(\lambda_i \mid x) \ , k = 1,..., K \tag{3.19}$$

$$\omega^* = \max_k \{P(\omega_k \mid x)\}, \quad \mathrm{k} = 1,...,\mathrm{K} \tag{3.20}$$

This method can be considered as a generalized framework of the classic methods for fusion which are using different strategies such as sum and product.

4. PERFORMANCE EVALUATION AND RESULTS

We have done experiments on the eNTERFACE'05 database, which is a well-known and popular acted database for emotion recognition. We also carried out experiments on the BAUM-1 acted (BAUM-1a) database, which were collected at Bahçeşehir University [102, 103]. Below, we first give a brief description of the databases. Then, we present and compare audio-visual emotion recognition results on these two databases using the framework described in Chapter 3. The audio-visual results obtained from some classic decision level fusion methods are also presented and compared.

4.1 DATABASES

4.1.1 eNTERFACE'05

eNTERFACE'05 audio-visual dataset [19], which contains audio-visual clips of 44 subjects from 14 different nationalities, speaking in English. Six basic emotional states anger, disgust, fear, happiness, sadness and surprise are expressed in the video clips of the database in an acted way by uttering given sentences with target emotions.

The final version of the database contains 42 subjects, coming from 14 different nationalities. Among the 42 subjects, a percentage of 81 percent were men, while the remaining 19 percent were women. A percentage of 31 percent of the total set wore glasses, while 17 percent of the subjects had a beard.

4.1.2 BAUM-1

BAUM-1 (Bahçeşehir University Multimodal Affective Database - 1) [20] is a collection of audio-visual facial clips of acted and spontaneous (re-acted) affective expressions. The audio-visual clips have been recorded from 31 subjects, who express a rich set of emotional and mental states in an unscripted way in Turkish. The database contains synchronous facial recordings of subjects with a frontal stereo camera and a half profile mono camera.

The subjects first watch visual or audio-visual stimuli on a screen in front of them, which are designed and timed to elicit certain emotions and mental states. The subjects

answer questions and express their feelings about the visual stimuli in their own words. The target emotions that have been elicited are the five basic ones (happiness, anger, sadness, disgust, fear) and additionally boredom and contempt. We also aim to elicit several mental states including being unsure (including confusion, undecidedness), being thoughtful, concentration, interest (including curiosity), and bothered (inc. complaint). The database also contains short acted recordings of each subject. The video clips have been categorically annotated by five labelers. Also a score between 0-5 is given to each video clip indicating the activation level at the peak frame of the emotion or mental state expressed in the video clip.

4.2 EXPERIMENTAL SETUP

Experiments can be conducted in subject dependent or subject independent manner. In subject dependent tests, data in the testing and training sets may contain data from the same subjects, while in subject independent tests data in the testing sets are from different subjects not existing in the training set. The performance can be obtained using different validation strategies, the most widely used ones in emotion recognition are *k*-fold cross-validation and leave-one-subject-out (LOSO) cross-validation.

In *k*-fold cross-validation, all samples are randomly partitioned into *k*-folds. In each validation cycle, one fold is used for testing and the rest of the folds are used for training. The process is repeated *k* times and the average result over all validation cycles is used as the final performance.

In leave-one-subject-out cross-validation (LOSO), the samples belonging to one subject in turn is reseved for testing. The remaining samples are used for training of the classification models, which then are tested against the test data of the reserved subject. LOSO validation scheme is computationally demanding but generally gives more reliable results, since more data is used for training. . In both cases, the average recognition accuracy is reported as the final emotion recognition accuracy.

4.2.1 Experimental Results on eNTERFACE'05 Database

We employed subject-independent LOSO cross-validation to conduct our set of experiments on 42 subjects of eNTERFACE'05 database. First we evaluate and compare the visual emotion recognition results using the peak frames selected by the

four proposed methods, the audio based method and manually selected peak frames (see Table 4.1). From the results, it is clear that clustering based peak frame estimation methods provide a higher emotion recognition accuracy as compared to the other proposed methods as well as the audio based peak frame selection method, even though it is not as high as the emotion recognition rate when the peak frames are selected manually. The advantage of our video-based peak frame selection methods to the audio based one could be the fact that there is sometimes a time shift between expressing the emotion by speech and facial expressions.

Table 4. 1: Video emotion recognition accuracies on eNTERFACE'05 database for all proposed peak frame selection and the manual peak frame selection based on LOSO cross-validation technique.

Peak Frame Selection Method	Recognition Accuracy
Manual Frame Selection (MFS)	47.05 %
Maximum Dissimilarity based Frame Selection (MAXDIST)	38.22 %
Emotion Intensity based Frame Selection (EIFS)	39.38 %
Clustering Based Peak Frame Selection (DEND CLUSTER)	40.00 %
Audio based Frame Selection (AFS)	34.46 %

Source: It has been done by Sara Zhalehpour.

Next, we report the emotion recognition accuracy of our multimodal framework using decision level fusion strategy for the peak frame selection methods. Results are given in Table 4.2, which is split into two parts. In the top part of the Table single channel results are shown for each modality using various peak frame selection methods. Audio modality clearly outperforms the video modality on this database.

The second part of Table 4.2 shows the results of decision level fusion approaches that we investigate. All decision level fusion strategies aim at utilizing differences in single modalities in order to enhance the overall performance. It can be well observed that product and sum based fusion methods perform better with respect to the others for almost all the peak frame selection methods. Using weighting in the sum or product rules cause a strong reliance on the dominant modality, which in this case is the audio modality. If one compares the four peak frame selection schemes, in most of the cases DEND CLUSTER peak frame selection method gives us the highest accuracy. This behavior can be explained by the better performance of the video modality on this

dataset and the resulting influence on the fusion. For the eNTERFACE'05 database, the weighted product method using the second power of the audio modality in (3.16) gives the best results. As we can see in the second row from below of Table 4.2, the accuracy for the manual peak frame selection is 79.57 percent and the accuracy using DEND CLUSTER peak frame selection method is78.26 percent, which are really close to each other and comparable. Hence, we can conclude that our peak frame selection method is effective in selecting reasonable peak frames from an expressive sequence.

Table 4. 2: Single and multi-modal emotion recognition accuracies on eNTERFACE'05 database for different decision level fusion techniques and peak frame detection methods, using LOSO cross validation. Maximum value of each row is shown in bold.

	MFS	MAXDIST	EIFS	DEND CLUSTER	AFS
Single modality results					
Audio	72.95 %	72.95 %	72.95 %	72.95 %	72.95 %
Video	47.05 %	38.22 %	39.38 %	**40.00** %	34.46 %
Decision level fusion results					
Max Rule	74.75 %	72.21 %	72.21 %	**72.38** %	71.33 %
Min Rule	74.65 %	**71.35** %	70.10 %	70.88 %	69.56 %
Sum Rule	78.25 %	75.00 %	74.31 %	**75.47** %	75.43 %
Average Rule	78.25 %	75.00 %	74.31 %	**75.47** %	75.43 %
Weighted Average Rule(1)	78.72 %	76.16 %	75.00 %	76.48 %	**76.63** %
Bayesian Rule (1)	78.56 %	**76.24** %	75.31 %	76.17 %	75.98 %
Product Rule	80.19 %	**76.39** %	74.76 %	75.93 %	74.39 %
Weighted Product (2)	79.57 %	76.94 %	76.08 %	**78.26** %	77.23 %
Weighted Product (3)	78.95 %	76.79 %	76.24 %	**77.41** %	76.11 %

(1) The weight is equal to 0.6 for audio modalities and selected empirically. (2) The second power of audio probabilities is used in (3.16) (3) The third power of audio probabilities is used in (3.16)

Source: It has been done by Sara Zhalehpour.

In order to learn more about the contribution of the single modalities to the fusion performance the confusion matrices of the audio, video and audio-visual modalities are given in Table 4.3, Table 4.4 and Table 4.5, respectively. If we compare the fusion and audio results for each emotion, we can see that the multimodal fusion increases the recognition accuracy for all the emotions. The emotions that have the highest two

emotion recognition accuracies are anger (88.37 percent) and happiness (75.35 percent) for the audio modality and disgust (63.72 percent) and happiness (53.49 percent) for the video modality. Disgust and happiness are the two emotions that benefit the most from the fusion of audio and video modalities. As the result, he overall system can discriminate anger, disgust and happiness better than the other emotions.

We also compared the performance of our framework with other methods in the literature that report emotion recognition results on the eNTERFACE'05 dataset, as shown in Table 4.6. We can observe from the table that the performance of our method (78.26 percent) is better than the other methods in the literature (to the best of our knowledge) among the methods that ensure subject independence.

Table 4. 3: Confusion matrix for the 6 basic emotions using eNTERFACE'05 database for the audio modality with the average accuracy of 72.95 percent.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**88.37%**	1.40%	3.26%	2.33%	2.33%	2.33%
Disgust	5.58%	**71.16%**	6.98%	4.19%	6.05%	6.05%
Fear	7.44%	9.77%	**64.19%**	4.65%	7.44%	6.51%
Happiness	6.98%	4.65%	1.86%	**75.35%**	6.98%	4.19%
Sadness	3.26%	6.51%	5.12%	6.05%	**72.56%**	6.51%
Surprise	4.19%	5.12%	7.44%	6.51%	10.70%	**66.05%**

Source: It has been done by Sara Zhalehpour.

Table 4. 4: Confusion matrix for the 6 basic emotions using eNTERFACE'05 database for the video modality and DENDCLUSTER frame selection method with the average accuracy of 40.00 percent.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**25.58%**	13.02%	11.16%	15.35%	12.56%	22.33%
Disgust	5.58%	**63.72%**	7.91%	15.81%	4.19%	2.79%
Fear	17.67%	15.81%	**13.95%**	9.77%	23.26%	19.53%
Happiness	6.98%	14.42%	6.05%	**53.49%**	3.26%	15.81%
Sadness	12.56%	11.63%	14.88%	6.05%	**37.67%**	17.21%
Surprise	14.42%	5.12%	11.63%	13.49%	9.77%	**45.58%**

Source: It has been done by Sara Zhalehpour.

4.2.2 Experimental Results on BAUM-1a Database

We also carried out the experiments on the BAUM-1a acted database. We used a total of 275 short acted recordings in the BAUM-1a database representing 8 emotional and mental states. For these experiments, we used a 5-fold subject independent cross-validation strategy. . We conducted two sets of experiments on BAUM-1a database. In the first set, we used the five basic emotions (anger, disgust, fear, happiness, and sadness). In the second set of experiments we added boredom, interest and unsure to the five basic emotions. For each of these experiments we compared the proposed peak frame selection methods (see Table 4.7 and Table 4.12). We can see that the highest recognition rates belong to clustering based peak frame selection method DEND CLSUTER(55.70 percent and 36.33 percent), which are even slightly better than manual peak frame selection results (55.61 percent and 31.32 percent) for both experimental sets. Furthermore, the other two peak frame selection methods still outperform the audio based peak frame selection method.

Table 4. 5: Confusion matrix for the 6 basic emotions using eNTERFACE'05 database for the audio-visual decision level fusion and DENDO CLUSTER frame selection method with the average accuracy of 78.26 percent.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**89.30%**	0.93%	4.19%	2.33%	1.40%	1.86%
Disgust	2.79%	**80.93%**	7.91%	3.72%	2.33%	2.33%
Fear	6.51%	8.37%	**68.37%**	3.26%	7.44%	6.05%
Happiness	3.26%	3.72%	0.93%	**84.19%**	4.19%	3.72%
Sadness	2.79%	4.19%	6.98%	1.86%	**76.28%**	7.91%
Surprise	3.26%	2.33%	7.44%	5.12%	11.16%	**70.70%**

Source: It has been done by Sara Zhalehpour.

Then, decision level fusion methods are applied to assess the best fusion method. As we can see in Table 4.8 and 4.13, the best accuracies for almost all peak frame selection methods are obtained for the DEND CLUSTER peak frame selection method by the weighted product rule. (The previous sentence is not true for Table 4.8) In this decision level fusion technique, the reliance of the overall performance to the audio channel is proportional to the third power of the audio probabilities. The video based recognition accuracy is inferior to the audio based accuracy for both 5 and 8 emotion cases. The gap

between audio (71.71 percent) and video accuracies (55.70 percent) becomes larger when we include boredom, interest and unsure to the experiments (63.53 percent and 36.33 percent for audio and video modalities, respectively).

Table 4. 6: Comparison of our method and other works on eNTERFACE'05 database (CV: Cross-Validation, NI: No Information)

Authors	Number Of Subjects	Subject Independency	Audio Based Recognition Accuracy	Video Based Recognition Accuracy	Audio-Visual Recognition Accuracy
M. Paleari [104]	44	NI	35.0 %	25.0 %	67.0 %
B. Schuller *et. Al* [105]	42	Yes – (5 Fold CV)	72.5 %	None	None
M. Mansoorizadeh [82]	42	No – (10 Fold CV)	33.0 %	37.0 %	71.0 %
R. Gajsek [106]	NI	No – (5 Fold CV)	62.9 %	54.7 %	71.3 %
D. Datcu [84]	42	Yes – (3 Fold CV)	55.9 %	37.7 %	56.3 %
Y. Wang *et. Al* [107]	43	No – (10 Fold CV)	38 %	58 %	76 %
K.Huang *et. Al* [108]	NI	Yes – (6 Fold CV)	56.4 %	52.3 %	61.1 %
Our approach	**43**	**Yes – (LOSO)**	**72.95 %**	**40.00 %**	**78.26 %**

Source: It has been done by Sara Zhalehpour.

The confusion matrices for the best case after fusion are given in Table 4.9, Table 4.10 and Table 4.11 for the 5 emotion case. The confusion matrices for the 8 class experiments are given in Table 4.14, Table 4.15 and Table 4.16.

Table 4. 7: Video emotion recognition accuracies for all proposed peak frame selection methods and the manual peak frame selection on BAUM-1a dataset based on 5-fold subject independent cross-validation technique for 5 basic emotions.

Peak Frame Selection Method	Recognition Accuracy
Manual Frame Selection	55.61 %
Maximum Dissimilarity based Frame Selection (MAXDIST)	46.60 %
Emotion Intensity based Frame Selection (EIFS)	52.06 %
Clustering Based Peak Frame Selection (DEND CLUSTER)	**55.70** %
Audio based Frame Selection (AFS)	42.18 %

Source: It has been done by Sara Zhalehpour.

Table 4. 8: Single and multi-modal emotion recognition accuracies on BAUM-1a database for different decision level fusion techniques and peak frame detection methods using 5-fold subject independent cross-validation technique for 5 basic emotions.

	MFS	MAXDIS	EIFS	DEND CLUSTER	AFS
Single modalities					
Audio	71.71 %	71.71 %	71.71 %	71.71 %	71.71 %
Video	55.61 %	46.60 %	52.06 %	**55.70** %	42.18 %
Decision level fusion					
Max Rule	74.00 %	71.79 %	73.14 %	72.07 %	**73.51** %
Min Rule	55.17 %	55.45 %	56.04 %	**60.20** %	56.61 %
Sum Rule	71.35 %	**73.66** %	71.01 %	70.41 %	71.37 %
Average Rule	71.35 %	**73.66** %	71.01 %	70.41 %	71.37 %
Weighted Average Rule(1)	73.71 %	**73.27** %	73.15 %	72.57 %	72.29 %
Bayesian Rule	74.11 %	**73.27** %	73.15 %	71.01 %	70.14 %
Product Rule	67.48 %	**69.21** %	66.28 %	64.78 %	70.30 %
Weighted Production(2)	72.11 %	73.94 %	71.39 %	72.56 %	**73.97** %
Weighted Production(3)	74.18 %	74.42 %	74.64 %	74.61 %	**75.91** %

(1) The weight is equal to 0.6 for the audio modality (2) The second power of audio probabilities is used in (3.16) (3) The third power of audio probabilities is used in (3.16)

Source: It has been done by Sara Zhalehpour.

Table 4. 9: Confusion matrix for the 5 basic emotions using BAUM-1a database for the audio modality with the average accuracy of 71.70 percent.

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**87.88**%	6.26%	1.82%	0.00%	4.04%
Disgust	10.00%	**73.33**%	5.00%	3.33%	8.33%
Fear	23.69%	8.19%	**46.62%**	7.00%	14.50%
Happiness	10.00%	6.19%	12.38%	**60.57%**	10.86%
Sadness	2.00%	2.50%	5.33%	0.00%	**90.17%**

Source: It has been done by Sara Zhalehpour.

Table 4. 10: Confusion matrix for the 5 basic emotions using BAUM-1a database for the video modality and DEND CLUSTER frame selection method with the average accuracy of 55.70 percent.

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**53.38%**	17.95%	13.91%	12.93%	1.82%
Disgust	5.83%	**75.33%**	0.00%	11.50%	7.33%
Fear	27.02%	7.83%	**41.90%**	6.67%	16.57%
Happiness	17.05%	9.52%	6.67%	**62.76%**	4.00%
Sadness	17.67%	18.44%	18.78%	0.00%	**45.11%**

Source: It has been done by Sara Zhalehpour.

Table 4. 11: Confusion matrix for the 5 basic emotions using BAUM-1a database for the audio-visual decision level fusion and DEND CLUSTER frame selection method with the average accuracy of 74.18 percent.

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**87.47%**	6.26%	2.22%	0.00%	4.04%
Disgust	5.00%	**80.83%**	5.00%	5.83%	3.33%
Fear	27.40%	4.50%	**59.10%**	4.50%	4.50%
Happiness	10.00%	6.67%	19.52%	**54.10%**	9.71%
Sadness	2.00%	2.50%	5.33%	4.00%	**86.17%**

Source: It has been done by Sara Zhalehpour.

Table 4. 12: Video emotion recognition accuracies for all proposed peak frame selection methods and the manual peak frame selection on BAUM-1a dataset based on 5-fold subject independent cross-validation technique for 8 basic emotions.

Peak Frame Selection Method	Recognition Accuracy
Manual Frame Selection (MFS)	31.32 %
Maximum Dissimilarity based Frame Selection (MAXDIST)	26.30 %
Emotion Intensity based Frame Selection (EIFS)	29.55 %
Clustering Based Peak Frame Selection (DEND CLUSTER)	**36.33 %**
Audio based Frame Selection (AFS)	20.83 %

Source: It has been done by Sara Zhalehpour.

Table 4. 13: Single and multi-modal emotion recognition accuracies on BAUM-1a database for different decision level fusion techniques and peak frame detection methods using 5-fold subject independent cross-validation technique for 8 basic emotions.

	MFS	MAXDIS	EIFS	DEND CLUSTER	AFS
Single modalities					
Audio	63.53 %	63.53 %	63.53 %	63.53 %	63.53 %
Video	31.32 %	26.30 %	29.55 %	**36.33 %**	20.83 %
Decision level fusion					
Max Rule	64.34 %	60.55 %	62.04 %	60.61 %	**63.09 %**
Min Rule	36.90 %	39.91 %	39.01 %	**40.27 %**	38.97 %
Sum Rule	64.15 %	62.43 %	62.67 %	**62.98 %**	62.30 %
Average Rule	64.20 %	63.35 %	63.90 %	**64.05 %**	63.05 %
Weighted Average Rule(1)	64.20 %	63.35 %	63.90 %	**64.05 %**	63.05 %
Bayesian Rule	63.56 %	63.38 %	63.16 %	63.16 %	**62.47 %**
Product Rule	58.06 %	61.00 %	58.59 %	58.40 %	**61.77 %**
Weighted Production(2)	62.20 %	**65.46 %**	63.72 %	65.01 %	63.28 %
Weighted Production(3)	65.44 %	**65.06 %**	64.51 %	64.80 %	64.78 %

(1) The weight is equal to 0.6 for the audio modality (2) The second power of audio probabilities is used in (3.16) (3) The third power of audio probabilities is used in (3.16)

Source: It has been done by Sara Zhalehpour.

Table 4. 14: Confusion matrix for the 8 emotions using BAUM-1a database for the audio modality with the average accuracy of 63.53 percent.

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**82.75%**	4.72%	8.48%	0.00%	0.00%	0.00%	4.04%	0.00%
Boredom	18.00%	**47.50%**	12.50%	0.00%	8.00%	0.00%	10.00%	4.00%
Disgust	10.00%	2.50%	**76.67%**	2.50%	5.83%	0.00%	2.50%	0.00%
Fear	17.69%	4.00%	11.19%	**56.62%**	2.00%	0.00%	4.50%	4.00%
Happiness	10.00%	0.00%	6.19%	6.67%	**66.95%**	0.00%	6.86%	3.33%
Interest	10.00%	2.86%	6.19%	5.00%	0.00%	**59.29%**	10.00%	6.67%
Sadness	2.00%	0.00%	11.17%	2.00%	0.00%	3.33%	**76.78%**	4.72%
Unsure	3.33%	2.00%	7.67%	13.67%	4.00%	5.00%	22.67%	**41.67%**

Source: It has been done by Sara Zhalehpour.

Table 4. 15: Confusion matrix for the 8 emotions using BAUM-1a database for the video modality and DEND CLUSTER frame selection method with the average accuracy of 36.33 percent.

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**48.23%**	0.00%	19.07%	7.37%	15.15%	0.00%	6.14%	4.04%
Boredom	28.00%	**0.00%**	5.00%	4.00%	0.00%	0.00%	22.00%	41.00%
Disgust	5.83%	0.00%	**77.83%**	2.50%	6.50%	0.00%	7.33%	0.00%
Fear	21.67%	0.00%	5.33%	**31.40%**	6.67%	2.50%	10.86%	21.57%
Happiness	17.05%	0.00%	9.52%	6.67%	**59.43%**	0.00%	4.00%	3.33%
Interest	27.86%	0.00%	6.19%	15.71%	21.67%	**0.00%**	16.19%	12.38%
Sadness	19.89%	0.00%	18.44%	2.00%	0.00%	0.00%	**53.44%**	6.22%
Unsure	34.67%	0.00%	1.67%	19.00%	7.67%	1.67%	15.00%	**20.33%**

Source: It has been done by Sara Zhalehpour.

Table 4. 16: Confusion matrix for the 8 emotions using BAUM-1a database for the audio-visual decision level fusion and DEND CLUSTER frame selection method with the average accuracy of 65.01 percent.

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**83.74%**	2.22%	7.78%	2.22%	0.00%	1.82%	2.22%	0.00%
Boredom	10.00%	**45.00%**	10.00%	10.50%	4.00%	0.00%	10.00%	10.50%
Disgust	7.50%	2.50%	**76.67%**	2.50%	8.33%	0.00%	2.50%	0.00%
Fear	12.36%	0.00%	15.19%	**57.95%**	0.00%	4.00%	4.50%	6.00%
Happiness	10.67%	4.00%	6.19%	3.33%	**66.29%**	3.33%	2.86%	3.33%
Interest	10.00%	0.00%	5.71%	0.00%	0.00%	**61.43%**	13.33%	9.52%
Sadness	0.00%	0.00%	6.94%	2.00%	0.00%	3.33%	**80.78%**	6.94%
Unsure	3.33%	2.00%	7.67%	17.67%	0.00%	1.67%	14.67%	**53.00%**

Source: It has been done by Sara Zhalehpour.

5. CONCLUSION AND FUTURE DIRECTION

In this thesis, we developed a fully automatic framework for audio-visual emotion recognition. We proposed three novel methods for peak frame selection methods to be used for emotion recognition from video. These methods are namely, maximum dissimilarity based frame selection, emotion intensity based frame selection and clustering based peak frame selection.

The audio-visual framework utilized local phase quantization (LPQ) features to represent visual information and a set of spectral features to represent audio information. The investigated various decision level fusion methods to achieve audio-visual emotion recognition. For experimental evaluation of the proposed method, experiments were carried out on the well-known acted eNTERFACE'05 database. The experimental results show that our results are superior to the results reported in the literature. The performance of the proposed system was also evaluated on the BAUM-1 database consisting of the five basic emotions as well as additional emotions and mental states. We have recently created this naturalistic audio-visual database in our laboratory at Bahcesehir University. The experimental results confirmed the good performance of the proposed audio-visual emotion recognition method.

In the experiments we used SVM classifiers. We believe that experimental results could be significantly increased by utilizing temporal models such as Hidden Markov Models or Dynamic Bayesian Networks to model the temporal variations of emotional expressions better (as compared to SVM).Fusion of audio and visual data could also be improved using temporal statistical models.

REFERENCE

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

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology,* vol. 17, p. 124, 1971.

[3] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 39, pp. 64-84, 2009.

[4] L. S.-H. Chen, "Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction," Citeseer, 2000.

[5] P. Ekman, "Facial expressions," *Handbook of cognition and emotion,* vol. 16, pp. 301-320, 1999.

[6] P. Ekman, "Basic emotions," *Handbook of cognition and emotion,* vol. 98, pp. 45-60, 1999.

[7] D. Galati, K. R. Scherer, and P. E. Ricci-Bitti, "Voluntary facial expression of emotion: comparing congenitally blind with normally sighted encoders," *Journal of personality and social psychology,* vol. 73, p. 1363, 1997.

[8] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication,* vol. 40, pp. 33-60, 2003.

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

[10] D. Lundqvist, "The Karolinska directed emotional faces (KDEF)."

[11] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, 2002, pp. 46-51.

[12] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000, pp. 46-53.

[13] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll, "Efficient recognition of authentic dynamic facial expressions on the feedtum database," in *Multimedia and Expo, 2006 IEEE International Conference on*, 2006, pp. 493-496.

[14] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, p. 5 pp.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94-101.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517-1520.

[17] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell*, et al.*, "" You Stupid Tin Box"-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus," in *LREC*, 2004.

[18] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie*, et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*, ed: Springer, 2007, pp. 488-500.

[19] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, 2006, pp. 8-8.

[20] O. Onder, S. Zhalehpour, and C. E. Erdem, "A Turkish audio-visual emotional database," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, 2013, pp. 1-4.

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

[22] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision,* vol. 57, pp. 137-154, 2004.

[23] Y.-l. Tian, L. Brown, A. Hampapur, S. Pankanti, A. Senior, and R. Bolle, "Real world real-time automatic recognition of facial expressions," in *In Proceedings of IEEE workshop on*, 2003.

[24] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear Models for 3-D Face and Facial Expression Recognition," *Information Forensics and Security, IEEE Transactions on,* vol. 3, pp. 498-511, 2008.

[25] http://www.mimik-lesen.com/mimik-lesen-buchung.html

[26] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, 2006, pp. 223-230.

[27] H.-Y. Chen, C.-L. Huang, and C.-M. Fu, "Hybrid-boost learning for multi-pose face detection and facial expression recognition," *Pattern Recognition,* vol. 41, pp. 1173-1185, 2008.

[28] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis,* vol. 17, pp. 592-598, 2007.

[29] J. Whitehill and C. W. Omlin, "Haar features for facs au recognition," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, 2006, pp. 5 pp.-101.

[30] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting encoded dynamic features for facial expression recognition," *Pattern Recognition Letters,* vol. 30, pp. 132-139, 2009.

[31] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Computer Vision and Image Understanding,* vol. 115, pp. 541-558, 2011.

[32] T. Xiang, M. K. Leung, and S.-Y. Cho, "Expression recognition using fuzzy spatio-temporal modeling," *Pattern Recognition,* vol. 41, pp. 204-216, 2008.

[33] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin*, et al.*, "The painful face–pain expression recognition using active appearance models," *Image and vision computing,* vol. 27, pp. 1788-1796, 2009.

[34] L. H. Thai, N. D. T. Nguyen, and T. S. Hai, "A facial expression classification system integrating canny, principal component analysis and artificial neural network," *arXiv preprint arXiv:1111.4052,* 2011.

[35] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *Image Processing, IEEE Transactions on,* vol. 16, pp. 172-187, 2007.

[36] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4125-4128.

[37] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+ 3D facial action and expression recognition," *Pattern Recognition,* vol. 43, pp. 1763-1775, 2010.

[38] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the 9th international conference on Multimodal interfaces*, 2007, pp. 38-45.

[39] C.-L. Huang and Y.-M. Huang, "Facial expression recognition using model-based feature extraction and action parameters classification," *Journal of Visual Communication and Image Representation,* vol. 8, pp. 278-290, 1997.

[40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding,* vol. 61, pp. 38-59, 1995.

[41] A. Sánchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernández, and J. J. Pantrigo, "Differential optical flow applied to automatic facial expression recognition," *Neurocomputing,* vol. 74, pp. 1272-1282, 2011.

[42] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, 2008, pp. 1-6.

[43] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 258-273, 2010.

[44] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream HMMs," *Information Forensics and Security, IEEE Transactions on,* vol. 1, pp. 3-11, 2006.

[45] Z. Zhang, "Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 13, pp. 893-911, 1999.

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

[47] T. Hu, L. C. De Silva, and K. Sengupta, "A hybrid approach of NN and HMM for facial emotion classification," *Pattern Recognition Letters,* vol. 23, pp. 1303-1310, 2002.

[48] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding,* vol. 91, pp. 160-187, 2003.

[49] R. Cowie, E. Douglas-Cowie, J. G. Taylor, S. Ioannou, M. Wallace, and S. Kollias, "An intelligent system for facial emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, p. 4 pp.

[50] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 699-714, 2005.

[51] M. MANSOURIZADEH, N. MOGHADAM CHARKARI, and E. KABIR, "AN EXPERT SYSTEM FOR EMOTION RECOGNITION FROM FACE IMAGE SEQUENCES," *THE CSI JOURNAL ON COMPUTER SCIENCE AND ENGINEERING,* 2006.

[52] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Multimedia, IEEE Transactions on,* vol. 8, pp. 500-508, 2006.

[53] C. Kai-Yueh, L. Tyng-Luh, and L. Shang-Hong, "Learning partially-observed hidden conditional random fields for facial expression recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 533-540.

[54] H. Chi-Ting, H. Shih-Chung, and H. Chung-Lin, "Facial expression recognition using Hough forest," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, 2013, pp. 1-9.

[55] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 42, pp. 993-1005, 2012.

[56] J. Rong, Y.-P. Chen, M. Chowdhury, and G. Li, "Acoustic features extraction for emotion recognition," in *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*, 2007, pp. 419-424.

[57] J. K. Casper and R. Leonard, *Understanding voice problems: A physiological perspective for diagnosis and treatment*: Lippincott Williams & Wilkins, 2006.

[58] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-941-IV-944.

[59] I. Mohino-Herranz, R. Gil-Pita, S. Alonso-Diaz, and M. Rosa-Zurera, "MFCC based Enlargement of the Training Set for Emotion Recognition in Speech," *arXiv preprint arXiv:1403.4777,* 2014.

[60] E. Vijayavani, S. Lavanya, P. Suganya, and E. Elakiya, "Emotion Recognition Based on MFCC Features using SVM," *International Journal,* vol. 2, 2014.

[61] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7527-7531.

[62] S. Scherer, F. Schwenker, and G. Palm, "Classifier fusion for emotion recognition from speech," in *Intelligent Environments, 2007. IE 07. 3rd IET International Conference on*, 2007, pp. 152-155.

[63] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication,* vol. 48, pp. 1162-1181, 2006.

[64] L. Chul Min and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on,* vol. 13, pp. 293-303, 2005.

[65] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner*, et al.*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals."

[66] T. S. Tabatabaei, S. Krishnan, and A. Guergachi, "Emotion recognition using novel speech signal features," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 345-348.

[67] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 240-243.

[68] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 474-477.

[69] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, 2003, pp. I-401-4 vol.1.

[70] M. Xia, C. Lijiang, and F. Liqin, "Multi-level Speech Emotion Recognition Based on HMM and ANN," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, 2009, pp. 225-229.

[71] S. Ntalampiras and N. Fakotakis, "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition," *Affective Computing, IEEE Transactions on,* vol. 3, pp. 116-125, 2012.

[72] X. Cheng and Q. Duan, "Speech Emotion Recognition Using Gaussian Mixture Model," in *Proceedings of the 2012 International Conference on Computer Application and System Modeling*, 2012.

[73] M. Paleari and C. L. Lisetti, "Toward multimodal fusion of affective cues," in *Proceedings of the 1st ACM international workshop on Human-centered multimedia*, 2006, pp. 99-108.

[74] M. S. Bartlett, J. R. Movellan, G. Littlewort, B. Braathen, M. G. Frank, and T. J. Sejnowski, "Towards automatic recognition of spontaneous facial actions," 2003.

[75] P. Atrey, M. A. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems,* vol. 16, pp. 345-379, 2010/11/01 2010.

[76] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems,* vol. 16, pp. 345-379, 2010.

[77] L. Chen, H. Tao, T. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information," in *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, 1998, pp. 83-88.

[78] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000, pp. 332-335.

[79] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh*, et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 205-211.

[80] C.-Y. Chen, Y.-K. Huang, and P. Cook, "Visual/Acoustic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 1468-1471.

[81] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. II-733-II-736.

[82] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications,* vol. 49, pp. 277-297, 2010.

[83] R. Gajsek, x030C, V. truc, and F. Mihelic, "Multi-modal Emotion Recognition Using Canonical Correlations and Acoustic Features," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4133-4136.

[84] D. Datcu and L. J. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*, 2011, pp. 122-128.

[85] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2879-2886.

[86] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and signal processing*, ed: Springer, 2008, pp. 236-243.

[87] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 878-883.

[88] Y. Songfan and B. Bhanu, "Facial expression recognition using emotion avatar image," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 866-871.

[89] A. Cruz, B. Bhanu, and S. Yang, "A psychologically-inspired match-score fusion model for video-based facial expression recognition," in *Affective Computing and Intelligent Interaction*, ed: Springer, 2011, pp. 341-350.

[90] Machine vision group, matlab codes for local phase quantization, http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab. Last Accessed: 01/07/2013.

[91] W. Yongjin and G. Ling, "Recognizing Human Emotional State From Audiovisual Signals," *Multimedia, IEEE Transactions on,* vol. 10, pp. 659-668, 2008.

[92] M. Meghjani, F. Ferrie, and G. Dudek, "Bimodal information analysis for emotion recognition," in *Applications of Computer Vision (WACV), 2009 Workshop on*, 2009, pp. 1-6.

[93] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*: Prentice-Hall, Inc., 1988.

[94] H. Husna, *Models to Combat Email Spam Botnets and Unwanted Phone Calls*: BiblioBazaar, 2012.

[95] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience,* vol. 3, pp. 71-86, 1991.

[96] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 19, pp. 711-720, 1997.

[97] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint arXiv:1003.4083,* 2010.

[98] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1992, pp. 121-124.

[99] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[100] K. P. Körding and D. M. Wolpert, "Bayesian decision theory in sensorimotor control," *Trends in cognitive sciences,* vol. 10, pp. 319-326, 2006.

[101] Y. Ivanov, T. Serre, and J. Bouvrie, "Error weighted classifier combination for multi-modal human identification," 2005.

[102] O. Onder, S. Zhalehpour, and C. Erdem, "A Turkish audio-visual emotional database," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, 2013, pp. 1-4.

[103] Onur Önder, "A Re-acted Audio-Visual Affective Turkish Database," *M.Sc. Thesis, Bahçeşehir University,* March 2014.

[104] M. Paleari and B. Huet, "Toward emotion indexing of multimedia excerpts," in *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, 2008, pp. 425-432.

[105] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 552-557.

[106] R. Gajsek, V. Struc, and F. Mihelic, "Multi-modal emotion recognition using canonical correlations and acoustic features," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4133-4136.

[107] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel Cross-Modal Factor Analysis for Information Fusion With Application to Bimodal Emotion Recognition," *Multimedia, IEEE Transactions on,* vol. 14, pp. 597-607, 2012.

[108] H. Kuan-Chieh, H. Y. S. Lin, C. Jyh-Chian, and K. Yau-Hwang, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, 2013, pp. 1-6.

CURRICULUM VITAE

Sara Zhalehpour
Date of birth June 2, 1986
Gender: female
Nationality: Iranian

Address:
Department of Electrical and Electronics Engineering
Bahcesehir University
Ciragan Cad., 34349, Besiktas,
İstanbul, Turkey
Mobile.: +90 (507)6013369
E-mail: s.zhalehpour@gmail.com

EDUCATIONAL BACKGROUND

02/2012 - present ***Bahcesehir University, Istanbul, Turkey***
M.Sc., Electrical and Electronic Engineering
GPA: 3.89/4
Thesis topic: "Audio-visual Affect Recognition"
Advisor: Assoc. Prof. Cigdem Eroglu Erdem
Program Language: English

09/2009-01/2012 ***Tabriz University, Tabriz, Iran***
M.Sc., Electrical Engineering-Telecommunication
GPA: 15.4/20
Thesis title: "Emotion Recognition from the Speech Signal"
Advisor: Dr. Mohammad Ali Tinati

09/2003-09/2009 ***Tabriz University, Tabriz, Iran***
B.Sc., Electrical Engineering-Telecommunication
GPA: 14.38/20
Thesis: "A Multiband Reconfigurable Antenna with Radiation Pattern Control by MEMs Switches"
Advisor: Dr. Saeid Nikmehr

RESEARCH INTERESTS

- Signal Processing, Audio/ Image / Video Processing
- Human-Computer Interaction (HCI)
- Pattern Recognition
- Machine Vision
- Wireless Communication and Networking
- MAC Layer Based on IEEE 802.11 Standard
- Ad Hoc Networks
- Antennas, Microwave, Computational Electromagnetics

PUBLICATIONS

[1]. S. Zhalehpour, Z. Akhtar, C. E. Erdem, "Multimodal Emotion Recognition with Automatic Peak Frame Selection", IEEE Int. Sym. on Innovations in Intelligent System and applications (INISTA) , Italy, June 2014.

[2]. O. Onder, S. Zhalehpour, C. E. Erdem, "A Turkish Audio-Visual Emotional Database", IEEE Int. Conf. Signal Processing and Communications Applications (SIU), Northern Cyprus, April 2013.

[3]. C. Turan, S. Zhalehpour, C. Kansın, Z. Aydın, C. E. Erdem, "A Method for Extraction of Affective Audio-Visual Facial Clips from Movies", IEEE Int. Conf. Signal Processing and Communications Applications (SIU), Northern Cyprus, April 2013.

[4]. S. Zhalehpour, H. Shahriar Shahhoseini, S.Zhalehpour, "Performance evaluation of adaptive Backoff algorithms in Ad Hoc network", in Proc. IEEE International Conference on Computer Technology and Development, 2009.

ACADEMIC EXPERIENCE

Bahcesehir University, Electrical and Electronic Department, Istanbul, Turkey

Teaching Assistant of:

- EEE 2180 Electronic Devices and Circuits Lab. **Spring 2012**
- EEE 2101 Circuit Theory Lab. **Spring 2012**
- EEE 2101 Circuit Theory Lab. **Fall 2012**
- EEE 2101 Circuit Theory Lab. **Spring 2013**
- EEE2314 Computational Analysis Lab. **Spring 2013**
- EEE 3304Feedback Control Systems Lab. **Spring 2013**

Research Assistant at:

- Signals & Communications Laboratory **Fall 2012 - present**

ACCOMPLISHED RESEARCHES AND PROJECTS

Bahcesehir University, Dept. of Electrical & Electronics Engineering, Istanbul, Turkey

- Performed course projects on the following topics, **2012 - present**
- Statistical models of shape and appearance
- Human emotion recognition from facial expressions
- Model matching algorithms

Tabriz University, Dept. of Electrical & Computer Engineering, Tabriz, Iran

- Performed research and projects on the following subjects, **2006 – 2012**
- Human emotion recognition from speech
- Square microstrip patch antennas
- Wireless Sensor Network Protocols In MAC Layer

Free Study:

- MAC Layer Of IEEE 802.11 Standard
- Ad Hoc Networks
- Mobile Networks

SELECTED COMPLETED COURSES IN M.SC PROGRAM

- Introduction to Digital Image and Video Processing (Grade : A$^-$) **Spring 2012**
- Digital Audio Processing (Grade : A) **Spring 2012**
- Optimization (Grade : A$^-$) **Spring 2012**
- Artificial Intelligence (Grade : A) **Fall 2012**
- Random Processes and Estimation Theory (Grade : A) **Fall 2012**
- Advanced Communication (Grade : 18/20) **Fall 2010**
- Digital Signal Processing (Grade : 16.5/20) **Spring 2010**

SCIENTIFIC CONTRIBUTIONS

Student member of:

- IEEE organization **2010-2012**
- Electrical And Electronic Workshop of Shahid Fahmideh Elite Foundation of Northwest, Tabriz, Iran **2011-2012**
- Electrical Engineering Scientific Society of University of Tabriz **2004-2012**

Article reviewer for:

- 14th Iranian Student Conference on Electrical Engineering, Kermanshah university, Kermanshah, Iran **2011**
- 15th Iranian student Conference on Electrical Engineering, Kashan, Iran **2012**

COMPUTER SKILLS

Technical Softwares

- MATLAB (+GUI+Simulation), PRAAT, C/C++, familiar with Network Simulator (NS2), HFSS, AWR, PASCAL, PSPICE, ORCAD.

General software

- Linux, Windows Vista/7, Microsoft office.

AWARDS

- Graduate program scholarship, Bahcesehir University, Istanbul, Turkey. **2012 – 2013**
- Graduate program scholarship, Tabriz University, Tabriz, Iran. **2009 – 2012**
- Undergraduate program scholarship, Tabriz University, Tabriz, Iran. **2003 – 2009**

WORK EXPERIENCE

Shahid Rajae Communication Center, Tabriz, Iran **2009**

- Internship,
  I was working as a communication engineer at switching support section.

LANGUAGES

Persian – Native.

English – Toefl (IBT) : 93.(Reading:21, Listening:25, Speaking:22, Writing:25)

Azerbaijani – Fluent

Turkish - Average

REFERENCES

- Dr. Cigdem Eroglu Erdem

  Email: cigdem.eroglu@bahcesehir.edu.tr

  Associate professor in the Department of Electrical and Electronics Engineering at Bahcesehir University, Istanbul, Turkey.
- Dr. Ufuk Tureli

  Email: mehmetserdarufuk.tureli@bahcesehir.edu.tr

  Associate professor in the Department of Electrical and Electronics Engineering at Bahcesehir University, Istanbul, Turkey.
- Dr. Baris Bozkurt

  Email: baris.bozkurt@bahcesehir.edu.tr

  Associate professor in the Department of Electrical and Electronics Engineering at Bahcesehir University, Istanbul, Turkey.