THE REPUBLIC OF TURKEY
BAHÇEŞEHİR UNIVERSITY

A RE-ACTED AUDIO-VISUAL AFFECTIVE TURKISH DATABASE

Master's Thesis

ONUR ÖNDER

İSTANBUL, 2014

THE REPUBLIC OF TURKEY
BAHÇEŞEHİR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING

A RE-ACTED AUDIO-VISUAL AFFECTIVE TURKISH DATABASE

Master's Thesis

ONUR ÖNDER

Supervisor: Assoc. Prof. Çiğdem Eroğlu Erdem

İSTANBUL, 2014

THE REPUBLIC OF TURKEY
BAHCESEHİR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ELECTRICAL AND ELECTRONICAL ENGINEERING

Title of Thesis: A Re-acted Audio-Visual Affective Turkish Database
Name/Last Name of the Student: ONUR ÖNDER
Date of Thesis Defense: 18.03.2014

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. Devrim Ünay:	___________________
Asst. Prof. Tarkan Aydın:	___________________

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Assoc. Prof Çiğdem Eroğlu Erdem for all her support, insight, and precious contributions to this thesis.

I also would like to thank the members of my thesis committee for their time and valuable suggestions. In addition, I would like to thank my friends, İlkay Öksüz, Ömer Tura, Mehmet Bay, Mustafa Kılıç, Erinç Topdemir, Sara Zhalehpour, Sezer Ulukaya, Ali Üstüner and Mehmet Göksu for their constructive discussions on subproblems in the scope of this thesis work. I would like to thank all the valuable volunteers, who accepted to be the subjects for the database recordings.

Finally, my special thanks go to my father Ümit Önder, my mother İlknur Önder, my brother Olcay Önder, and to my best friend Hüseyin Samet Tanılmış for their support, and encouragements throughout this thesis study.

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

Onur ÖNDER İSTANBUL, March 2014

ABSTRACT

A Re-acted Audio Visual Affective Turkish Database

Onur Önder

Electrical and Electronics Engineering

Thesis Advisor: Assoc. Prof. Çiğdem Eroğlu Erdem

March 2014, 79 pages

Scientific research on emotion recognition have gained great interest from researchers in the past decade due to its importance in human-computer interraction and artificial intelligence. Extensive affective databases are needed to test the emotion recognition algorithms. Most of the existing databases available to researchers are collections of acted data. However, naturalistic and spontaneous data is needed for developing affect recognition algorithms which will work under realistic conditions.

In this thesis, we recorded and annotated a spontaneous audio-visual face database consisting of expressions of emotions as well as mental states. The targeted emotions are *happiness, sadness, anger, disgust, fear, surprise, contempt* and *boredom*. The targeted mental states are *interest* (including *curiosity*), *unsure* (including *confusion* and *undecidedness*), *bothered* (including *complaint*), *thoughfulness* and *concentration*.

The database is named as BAUM-1: Bahçeşehir University Multimodal Affective Face Database of Spontaneous Affective and Mental States. BAUM-1 has been collected from 31 subjects and it contains video clips recorded from two different angles (frontal stereo and half profile mono). The database contains about 25 hours of video and audio data. The database is being shared by researchers via a web site and we hope it will be a valuable resource for researchers working on audio and/or visual affect recognition.

Keywords: Emotion Recognition, Affect Recognition, Audio Visual Database, Affective Computing

ÖZET

Görsel İşitsel Türkçe Duygusal Veri Tabanı

Onur Önder

Elektrik-Elektronik Mühendisliği
Tez Danışmanı: Doç. Dr. Çiğdem Eroğlu Erdem

Mart 2014, 79 sayfa

Otomatik duygu tanıma üzerine yapılann çalışmalar, insan-bilgisayar etkileişimi ve yapay zeka çalışma alanlarındaki önemi sayesinde araştırmacılar tarafından büyük bir ilgi kazanmış durumdadır. Geniş duygu veri tabanları, duygu tanımlama algoritmalarının test edilmesi için gereklidir. Çoğu veri tabanı, rol yapılan verilerden oluşmaktadır. Ancak gerçekçi durumlarda çalışacak duygu tanıma algoritmaları geliştirmek için doğal ve spontan verilere ihtiyaç vardır.

Bu tezde, uygu ve zihinsel durum ifadeleri içeren spontan görsel-işitsel yüz veri tabanının kayıt ve etiketlemesini gerçekleştirdik. Hedeflenen duygular; *mutluluk, üzüntü, kızgınlık, iğrenme, korku, şaşırma, küçümseme* ve *sıkıntı*dır. Hedeflenen zihinsel durumlar; *ilgi (merak* dahil*), emin olamama (kafa karışıklığı* ve *kararsızlık* dahil*), rahatsız olma (şikayet etme* dahil*), düşünceli* ve *konsatre*dir.

Veri tabanı, BAUM1: Bahçeşehir Üniversitesi Spontan Duygusal ve Zihinsel Durum Çok Kipli Duygusal Yüz Veri Tabanı olarak isimlendirilmiştir. BAUM1, 31 denekten toplanmıştır ve iki farklı açıdan (önden stereo ve yarı profilden mono) kaydedilmiş video kliplerini içerir. Veri tabanı yaklaşık 25 saatlik görüntü ve ses verilerinden oluşmaktadır. Veri tabanı bir internet sitesi aracılığıyla araştırmacılarla paylaşılmaktadır ve umarız ki işitsel ve/veya görsel duygu tanımlama üzerine çalışan araştırmacılar için değerli bir kaynak olacaktır.

Anahtar Kelimeler: Duygu Tanıma, Etki tanıma, Görsel İşitsel Veri Tabanı.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AVI	:	Audio Video Interleaved
JPEG	:	Joint Photographic Experts Group
MOV	:	Quick Time File Format
WAV	:	Waveform Audio File Format
HTML	:	Hyper Text Markup Language
FPS	:	Frames per Second
MFCC	:	Mel-frequency Cepstrum Coefficients
PLP	:	Perceptual Linear Prediction
SVM	:	Support Vector Machine

1. INTRODUCTION

Emotions are important communication cues used in daily human-to-human interactions. How one says a message is sometimes more important than what is actually said. For naturalistic human-computer interaction scenarios, emotions should also be involved. There are many application areas of automatic emotion recognition such as tele-medicine [Gutierrez, 2012], gaming [Barakova, 2009], e-learning [Wang, 2009], ubiquitous / pervasive computing [Jungum, 2009], military applications [Clavel, 2008], smart home projects [Costoulas, 2008] etc.

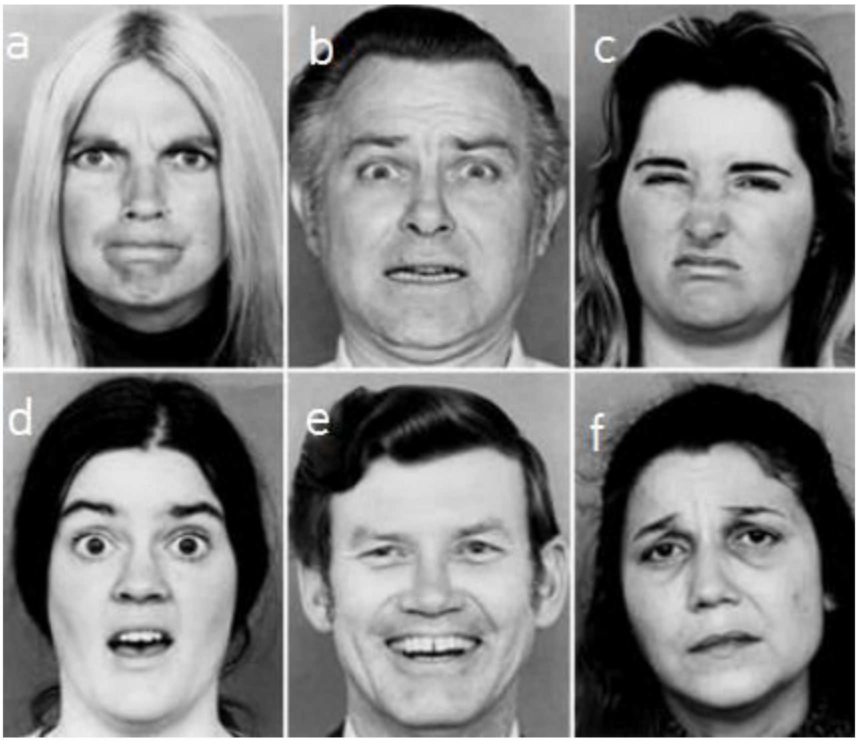
For example, in gaming industry, all game development communities constantly seek new approaches, techniques and tools that will permit them to easily incorporate AI (Artifficial Intelligence) in their games. In the action games like Half Life and F.E.A.R (First Encounter Assault Recon), emotion recognition approaches has been used to develop AI of enemy soldiers [Murphy, 2013]. The resulted AI had a dramatic effect on a players experience during game play.

Another example, in [Wang, 2009], authors explored how emotion evolves during learning process and how emotion feedback could be used to improve learning experiences. The article describes a cutting-edge pervasive e-Learning platform used in a Shanghai online college and proposed an affective e-Learning model, which combined learners' emotions with the Shanghai e-Learning platform. An experimental prototype of the affective e-Learning model was built to help improve students' learning experience by customizing learning material delivery based on students' emotional state. Experiments indicated the superiority of emotion aware over non-emotion-aware with a performance increase of 91 percent.

1.1 PROBLEM DEFINITION AND MOTIVATION

First studies on emotion recognition was made by the French neurogolist Guillaume-Benjamin-Amand Duchenne (de Boulogne) in 1862. In 1971, Ekman and Friesel defined 6 universal emotions. These are "Anger", "Disgust", "Fear", "Happiness", "Sadness" and "Surprise".

Figure 1.1: Emotion Examples



Notes: a) Anger, b) Fear, c) Disgust, d) Surprise, e) Happiness f) Sadness

There are many different cues and classification methods for affect recogniton. Affect can be recognized using various channels such as facial expressions, vocal prosody, signals of the autonomous nervous system (hearth rate, skin conductivity, etc.), or other hand/body gestures and posture.

There is a vast amount of research on emotion recognition from facial expressions. Ekman and Friesen created a model known as the Facial Action Coding System (FACS). Ekman has argued that emotions are linked directly to the facial expressions, and there are six basic “universal facial expressions” corresponding to happiness, surprise, sadness, fear, anger, and disgust. In all of the visual works, some method to extract features from facial images is used and a classifier is used to detect the expressions. [Mase, 1991] used optical flow to extract the facial motion and used spatio-temporal templates to classify the expressions using a k-nearest neighbor classifier (kNN). [Black and Yacoob, 1995] used local parameterized models of image motion to estimate the nonrigid motion and a coarse-to-fine gradient-based optical flow to estimate large motions.The methods in the literature that try to recognize emotions from speech use features such as Mel-Frequency Cepstral Coefficientss (MFCC) [Gilke, 2012], Linear Predictive Coefficient (LPC) , Spectrcal Envelope (Formants) [Bozkurt, 2011], etc. Work on recognition of emotions from voice and video has been recently suggested and worked by [Chen 2000], and [DeSilva 1997] who studied human's ability to recognize six basic emotions by means of subjective evaluation. Chen analysed the same audio visual data that Desilva et al. showed to the other subjects and showed that two modalities complement each other. Chen's results showed potential advantages in using both modalities over either modality alone.

In order to test audio-visual affect recognition algorithms, databases contain sufficient variety are needed. Most of the databases in the literature available today are acted and only contain the six basic emotions. In this thesis our we introduce a spontaneous audio-visual face database containing expressions of emotions as well as several mental states.

1.2 CONTRIBUTION AND SCOPE

The main purpose of this thesis is to create a re-acted audio visual database of emotional and mental states. The targeted emotions are *happiness, sadness, anger, disgust, fear, surprise, contempt* and *boredom*. The targeted mental states are *interest* (including *curiosity*), *unsure* (including *confusion* and *undecidedness*), *bothered* (including *complaint*), *thoughfulness* and *concentration*. To the best of our knowledge, there are no databases in the literature that contain the above mental states recorded in a spontaneous way.

The novalities of our database are:

It is the first and only Turkish database.

We have stereo recordings.

Resolution and audio sample rate are higher than existing databases.
The database contains over 25 hours of video / audio.
As well as emotional states, mental states are included.

The database is named as BAUM-1: Bahçeşehir University Multimodal Affective Face Database of Spontaneous Affective and Mental States. BAUM-1 has been collected from 31 subjects and its contains video clips recorded rom two different angles (frontal stereo and half profile mono). The database contains about 25 hours of video and audio data. The database is being shared by researchers via a web site and we hope it will be a valuable resource for researchers working on audio and/or visual affect recognition.

The organization of the thesis is as follows. In Chapter 2, a literature survey on existing audio-visual databases is given. In Chapter 3, we describe the collection process of the BAUM-1 database. In Chapter 4, we give baseline audio-visual emotion recognition experiments. Finally, in Chapter 5, conclusions and discussion are presented together with possible future research directions.

2. PREVIOUS WORK

There are many single modal emotional databases in the literature. Generally these databases contain only face images/videos or audio data. There are also several multi-modal emotional databases, which have been collected recently. However, they are generally recorded in an acted way. There are also a few databases which contain spontaneous emotional expressions. Below, we briefly review the multi-modal emotional databases in the literature pointing out their shortcomings.

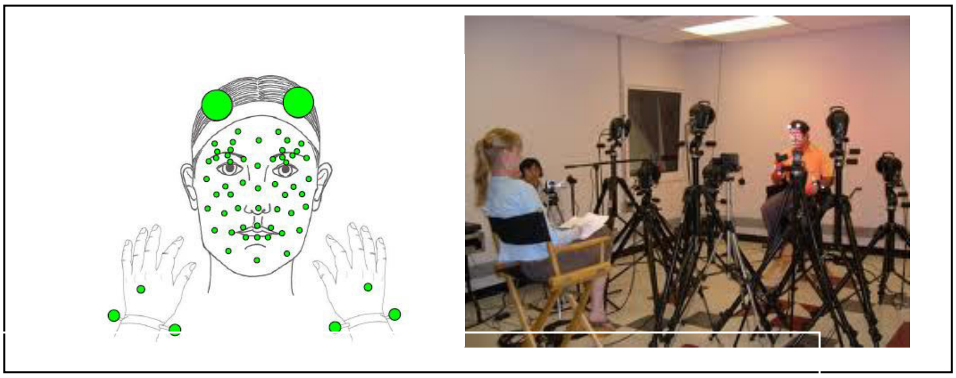
2.1 Existing Audio-Visual Emotional Databases

IEMOCAP Database [Busso, 2008]: Interactive Emotional Dyadic Motion Capture Database (IEMOCAP), is a multi-modal database, which is recorded by University of Southern California. This database consists of 12 hours of recordings of interactions between two actors out of a total number of ten actors. In order to be able to track the facial expressions in detail, markers have been placed on their head and face.

Recordings are divided into two parts, which are text based and improvisation based. In the text based part, professional actors (or students of performing arts) perform their role by memorising and rehearsing a scenario. And in the improvisation based part, they are asked to chat about a topic and share their ideas in an unscripted way.

During these interactions, only one of the actors is recorded. By placing cameras and microphone approximately one meter away from the subject, natural communication conditions are maintained. The camera and microphone are located in a way that they do not effect the actors (see Figure 2-1).

Figure 2.1 IEMOCAP, placed markers (left), b) recording environment (right)

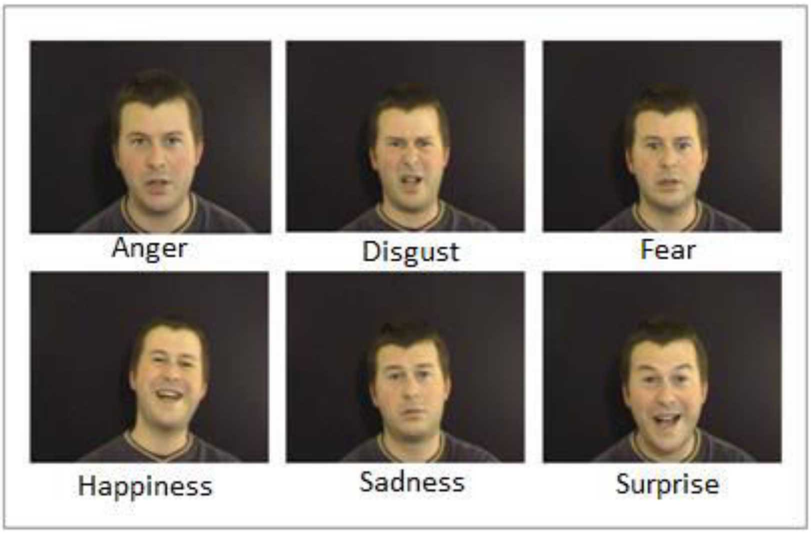


Placed markers (left), recording environment (right)
Reference: [Busso, 2008]

After recording the subjects, video and audio files are segmented into little segments and each of these segments are labeled by evaluators. Two different methods have been used for labeling. In the first one, each segment is assigned to one of the categories which consist of anger, sadness, happiness, disgust, fear, surprise, dissappointment, excitement and neutral. In the second method, labeling is done using a continuous 3D space by giving a score between 1 and 5 for the *valance*, *activation* and *dominance* dimensions. This three dimensional labeling technique has become popular in the last years since is it is possible to represent the intensities of emotional expresions. While a *valance* between 1-5 shows us how much negative or positive the emotion is, *activation* shows how excited the person is, and *dominance* shows us how weak or strong the emotion is. Only a part of the database, which contains only two actors, is currently shared with researchers and the shared facial recordings are not frontal, they are from an approximately 45 degree angle.

eNTERFACE'05 Database [Martin, 2006]: This database has audio visual recordings from 42 different subjects and 14 different nationalities. Each subject is asked to act regarding to a scenario and say the sentences, which they prepeare for the scenarios which are aimed to expose six basic emotions which are anger, fear, happiness, surprise, sadnes, disgust (see Figure 2-2). Since the subjects are not professional actors, some of them succeed to reflect the required emotion and the others failed to do it.

Figure 2.2: Examples from the eNTERFACE'05 database.



Reference: [Martin, 2006]

In this database, because of the emotions are given before recordings and subjects are asked to express that emotions, there are no labeling process after recordings (labeled are set at the beginning) and whole database is open sharing with researches.

Belfast NaturalisticDatabase [Cowie, 2007] : This database is a collection of 239 videos which are between 10 and 60 seconds long and collected from various television shows, interviews, and TV programs. 209 of them are from television programs, 30 of them are from interviews which are made by researchers themselves. There are 31 male and 94 female subjects in the database and a wide range of emotions. The database has neutral clips of each subject and at least one emotional clip for each subject. Labeling is done using both categorical and dimensional approaches. FEELTRACE [Campbell, 2003] software is used for the annotation process.

Vera Am Mittag (VAM) Database 0: VAM is a database which takes its name from a German TV talk-show and consists of the recordings from that talk-show's 12 different episodes shown between December 2004 and February 2005. The recordings are partially spontaneous since the subjects did not know that they were going to be analysed emotionally. Also it is an advantage for database that thema of the show is related to friendship, family etc. which could provide great emotional data for the database. The other advantage of the database is, in an episode of the show, it is possible to record the same subject with different emotions and reactions. On the other hand, a TV show is an environment that everything is not under control so on the analysing stage, there can be many difficulties that can not be undone, this is the biggest disadvantage of the database. VAM has three parts, these are VAM-Video, VAM-Audio and VAM-Faces. Vam-Video has 1421 video segments from 104 subjects but it does not contain labels. Vam-Audio and VAM-Faces have labels with them. While segmentation process, audio is categorised as very good, good, fair and not useful. Then they have been labeled according to their emotional content. Examples from the VAM database can be seen in Figure 2-3.

Figure 2.3: Examples from the VAM database.

Reference: [Grimm, 2008]

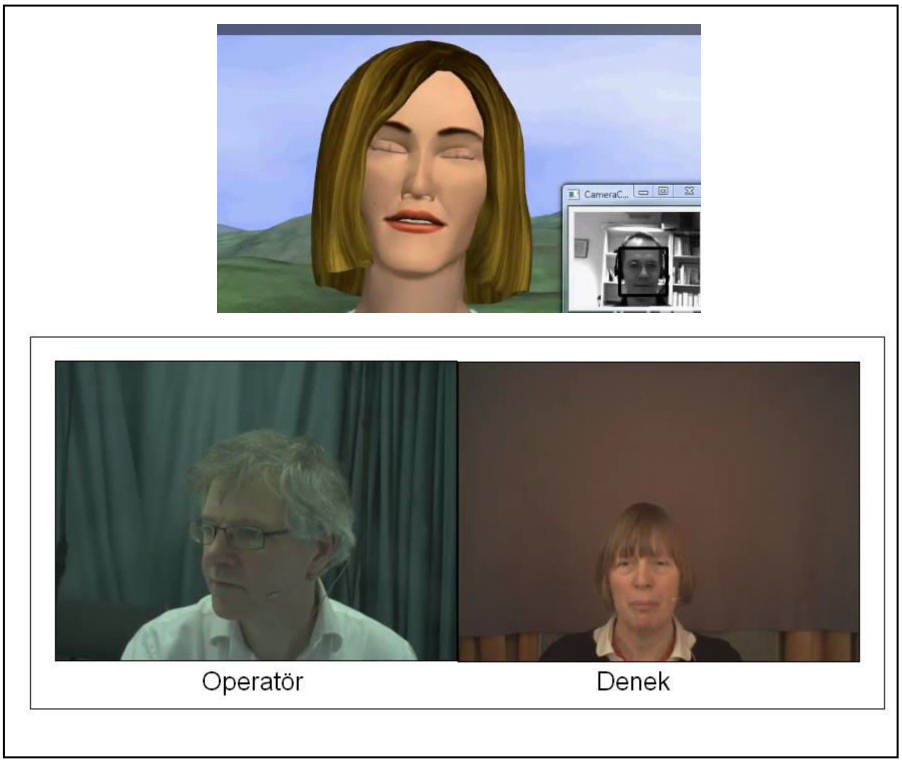
Humaine Database [Cowie, 2003]: Humaine is a database which consists of 50 clips of length between 5 seconds and 3 minutes. It is a combination of different databases and contains both spontaneous and artificial data. This database is partially labeled and two different methods are used to label the clips. In the first method, evaluators give one label for each clip and in the other they labeled the clips by giving scores to them on a continuous scale.

Semaine Database [McKeown, 2010]: Semaine is a database which takes its name from an EU project which is called “The Sensitive Agent Project”. The aim of this project is to create an artificial listener; which is able to listen, answer and change its reactions according to the speaker’s face expressions and voice stress. In these scenarios which is called as Sensitive Artificial Listener (SAL), a subject is being recorded while talking with the virtual character on the screen.

In Semaine database, subject which is recorded and the operator which animates five different Semaine characters are interacting. There are no pre-written scenarios and individuals speak by making improvisations. The operator and the subject see each other only from the monitor (see Figure 2.4). The four virtual Semaine characters that operator animates have different features. “Prudence” character is calm and sensitive, “Poppy” is cheerful and extrovert, “Spike” has angry and agressive, and “Obadiah” has a depressive character. Apart from the interaction with the virtual character, it is forbidden to ask any thing to operator himself.

Subjects are recorded by five hi-speed cameras and five hi-definition microphones. 24 recordings have been made from 20 subjects and these recordings have been segmented into 144 segments. On the data obtained, continuous labeling method has been used on five dimensions which are “Valence”, “Activation”, “Power”, “Expectation”, and “Intensity”. According to that dimensional labeling, they are categorized into four basic categories which are “Basic Emotion”, “Epistemic Status”, “Interaction Analysis”, “Validity”.

Figure 2.4: Examples from the SEMAINE



Virtual Character and Subject (upper), The operator and the subhect
Reference: [McKeown, 2010]

2.2 COMPARISON AND SHORTCOMINGS OF EXISTING DATABASES

IEMOCAP, SEMAINE and eNTERFACE'05 databases consist of their own recordings and VAM and Belfast Naturalistic , databases are collected from the videos which are chosen from television shows. Also, Humaine database is a compilation of some other databases. VAM and Belfast Naturalistic which make use of television shows, obtain fine emotional variations and spontaneous recordings which are not artificial and acted. But the uncontrolled parameters like camera and microphone variations are disadvantages for automatic affect recognition algorithms.

The difference betwee IEMOCAP and eNTERFACE'05, which both have their own recordings, is that IEMOCAP has more improvised, spontaneous recordings. For IEMOCAP recordings, professional actors were used and these actors have been recorded while both acting according to a scenario and improvising on a topic. The most distinguishing feature of IEMOCAP is that they used markers placed on the face and the hands. In eNTERFACE'05, non-professional people were used and these people were asked to look at the camera and say the sentence that was given to them to be memorized beforehand. While some of the subjects were successful in doing that, some of them acted artifically. Another shortcoming of the existing databases is that their video resolution is generally low (except Semanine), which poses a difficulty for autiomatic face detection algorithms.

In most of the existing databases in literature, emotion labeling is done by some annotators. While some of them do it by just categorizing into basic emotional groups, some of them was done by scoring them continuously and dimensionally. Two softwares are available to be used for the labeling process. One of them is ANVIL which is used to label the segments of IEMOCAP database, and the other one is FEELTRACE , which is used to label the segments of Belfast Naturalistic database.

In the Table 2.1, existing databases are summarized and compared according to their emotional content, the number of subjects, the clip length, language, labeling and if it is open to sharing to the researchers.

Table 2.1: Summary and comparison of existing databases in literature

Database	*Emotional Content*	*Acted/ Naturalistic*	*Number of Subjects*	*Record Length, Video and Audio Info*	*Language*	*Shared?* *Labeled?*
IEMOCAP	Extensive (Basics + frustration)	A & N	10	12 hours	Eng	Partially shared (2 subjects) Yes
eNTERFACE'05	Basic Emotions	A	42	1166 short clips Video: 720x576 @25fps, Audio: 48kHz	Eng	Yes Yes
VAM	Extensive (Valance, activation, dominance scores)	N	47	12 hours, Video:352x288 pixels @ 25fps, Audio: 16kHz	German	Yes Partially labeled
Humaine	Extensive	A & N	unspecified	50 clips	Eng., Fr., Hebrew	Yes, Partially labeled with ANVIL (16 clips)
Semaine	Extensive (Basic Emotions + Amusement, Epistemic states like certain, agreement)	A & N	20 subjects, 24 records, 144 segment	6,5 hours Video: 580x780 pixels @ 50 fps, Audio: 48kHz	Eng	Yes, Partially labeled

Reference: **[Önder, 2013]**

3. BAUM-1 DATABASE

The BAUM-1 database has been recorded in a studio, which is designed specifically as described in the following. In the studio, subjects first watch a stimuli video from a monitor and then express their feelings in their own words while they are being recorded with multiple cameras.

3.1 OVERVIEW OF THE RECORDING METHOD

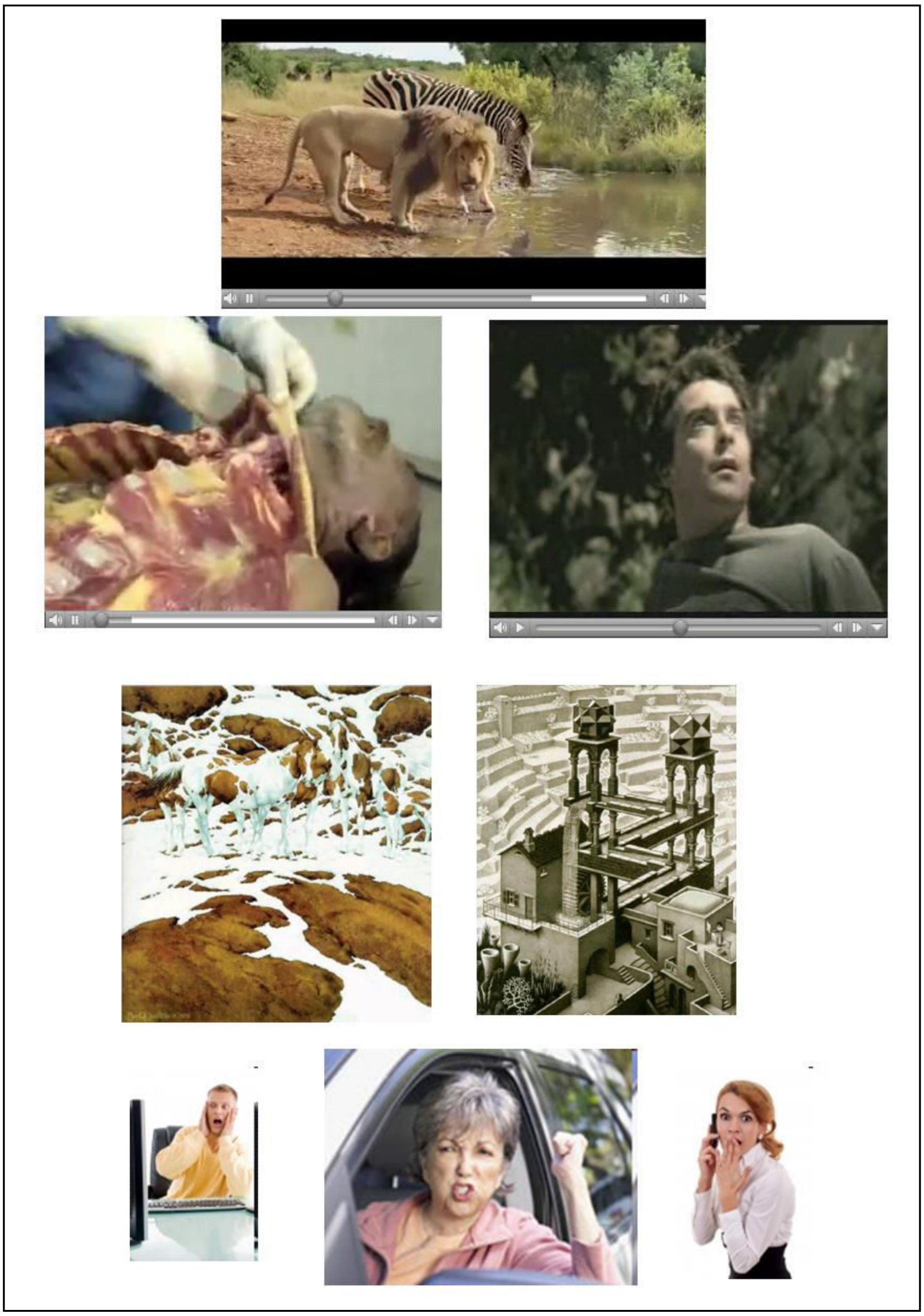
In BAUM-1 database, our main focus was to obtain spontaneous expressions of emotional and mental states. In order to bring a subject into the mood of the emotion, we first asked each subject to watch a video, which is called as “stimuli video”. While watching the video (or shortly after that), the subject is asked to comment on or explain their feelings about the video / image shown. While subjects are watching the video or expressing themselves, they are being recorded. After the recording process, the video is segmented and annotated. Below we explain each step in more detail.

3.1.1 Stimuli Video

Stimuli video is an audio-visual collection of materials, which contains video clips and images that will be shown to the subject to evoke target emotions in the subject. In order to create such a stimuli video, we seriously searched and them eliminated many candidate video and images. First, we made use of IAPS (International Affective Picture System). IAPS is an emotional stimulant set of images, which is created for the researchers working on emotion and attention relation research areas. This database is open to researchers and has a rich content. The pictures of IAPS have been labeled in three dimensions (valance, activation, dominance) with scores between 1-9 by people from different age groups. We have choosen the pictures with high valence scores.

Besides IAPS, we have also collected videos from television shows, from social networks, and also we have collected some confusing illusion images used in IQ tests. The first version of the stimuli video consisted of 56 pictures and 19 videos which had different lengths varying between 35 seconds and 12 minutes. Several examples of stimuli images are shown in Figure 3-1.

Figure 3.1: Examples from Stimuli Video



Reference: [Önder, 2013]

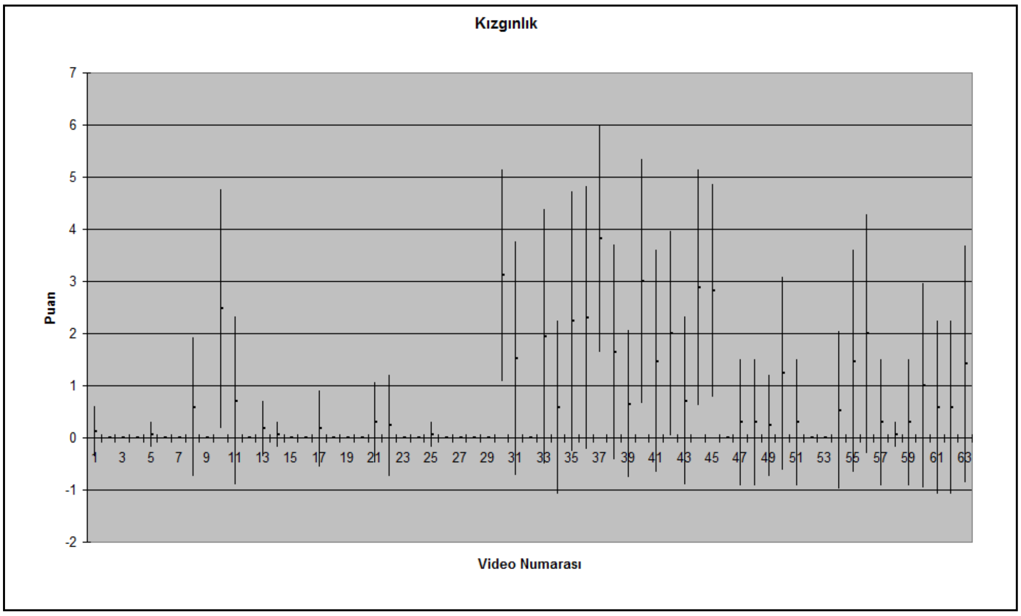
In order to pick the most stimulating images and video clips from a set of initial candidates, we did a collective presentation of the initially selected 56 pictures and 19 videos to an audiance (evaluator team) who were students of the physchology department. The 19 people, watched the videos and pictures carefully and scored them on a survey. There were 12 categories on that survey and each evaluator selected a category and also gave a score between 0 and 5 to each video or image. These 12 categories were: *Anger, Disgust, Fear, Happiness, Surprise, Neutral, Sadness, Confusion, Boredom, Curiosity, Serenity* and *Excitement*. After this presentation, we analysed the survey results statistically. For each category, we calculated the mean score and standart deviation of the scores given by the annotators. In Figure 3-2, the scores given by the 17 annotators to the stimulus number 12 are shown together with the averages and standard deviations of each category. In Figures 3-3 to 3-5, we give plots of scores of all videos for the anger, confusion and happiness, respectively.

Figure 3.2: An example of the surveys which are used for normalization

Değerlendirmen:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Average	Standard Deviation
Görüntü 12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kızgınlık	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sıkıntı	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kafa Karışıklığı	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tiksinme	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Korku	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,294118	1,212678125
Mutluluk	0	0	0	3	0	4	2	3	0	0	0	0	2	0	4	0	5	1,352941	1,800735144
Şaşırma	3	5	0	0	2	3	0	0	2	0	0	0	0	4	3	0	0	1,294118	1,72353945
Nötr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Üzüntü	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0,235294	0,9701425
Merak	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,117647	0,48507125
Huzur	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eğlence	4	4	5	5	4	5	5	4	2	3	5	5	5	5	4	5	5	4,411765	0,870260272

This example shows 17 evaluator's scored and their mean and standard deviations for image 12.

Figure 3.3: Video- Score Graphic for category "Anger".



X-axis shows the video/picture number and Y-axis shows their score for Anger. The points on the graphic are mean values of scores and the lines are standard deviations. First evaluations made according to the mean values, and the second made according to deviations.

Figure 3.4: Video-Score Graphic for Confusion

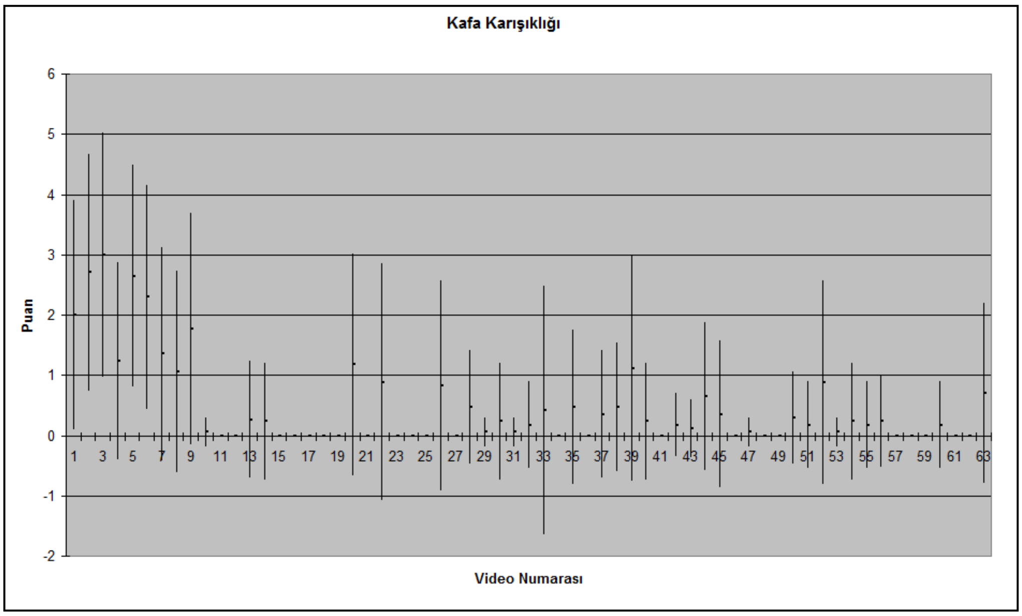
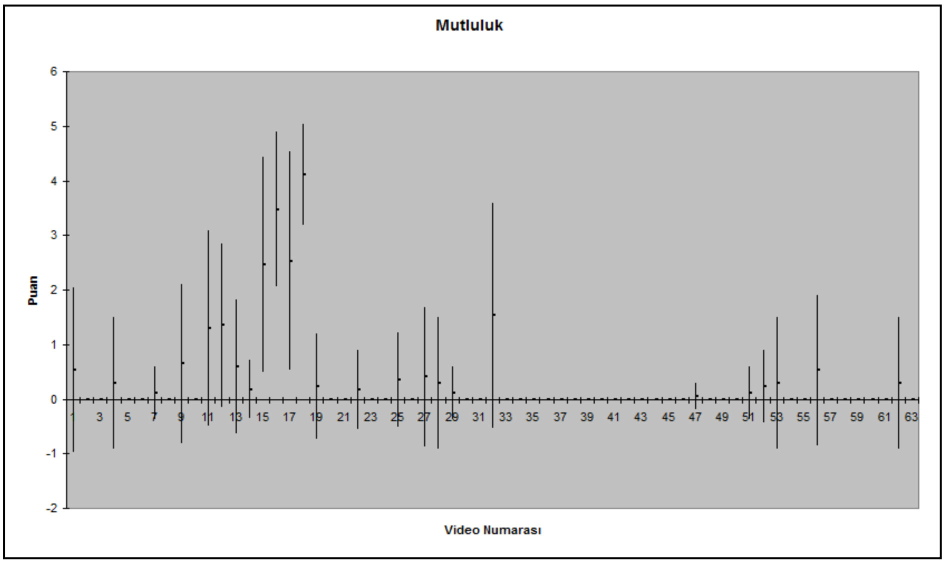


Figure 3.5: Video-Score Graphic for Happiness



After the statistical analysis, the videos / images which have the mean scores higher than 2.5 are selected and the others have been eliminated. While there are many clips which are above the treshold in the categories such as excitement and disgust, in the categories such as curiosity or serenity there were not many clips above the treshold. So for such categories, the treshold have been reduced to 2. And in the other hand, if a category had too many clips above the threshold, we considered the standard deviations, too and eliminated the ones with the highest standard deviations.

After this elimination proces, the last version of the stimuli video consisted of 29 video / images. After each of video or images, we inserted a 45 second break and asked the subject to talk in an unscripted way, to express his or her feelings about the seen image or video. In Table 3-1, we give a short description of each stimulus in the final video.

Table 3.1: Content of Stimuli Video

Video / Picture Number	Content	Target Emotion / Mental State
1	Horses	Confusion
2	Stair paradox	Confusion
3	Wheel illusion	Confusion
4	Face inside lines	Confusion
5	Dogs	Happiness
6	Advertisements	Entertainment
7	Cem Yılmaz	Entertainment
8	Space	Neutral / Boredom
9	Cars	Entertainment
10	Children	Happiness
11	Crazy sportsman	Entertainment / Surprise
12	Surprised man	Neutral / Boredom
13	A clip from a TV show	Sadness / Anger
14	Child and vulture	Sadness
15	Sick baby	Sadness
16	Dead child and father	Sadness / Anger
17	Murder of a cat	Anger
18	Fisherman	Neutral / Boredom
19	Angry person	Neutral / Boredom
20	Man with a gun	Anger
21	Shark	Fear
22	A man vomitting	Disgust
23	Waterfall	Neutral / Boredom
24	Car accident	Sadness
25	Car accident 2	Sadness / Anger
26	Injured hand	Disgust
27	Clips from horror movies	Fear
28	Autopsy	Disgust
29	An illusionist who cuts his wife accidentally	Fear / Surprise

We also recorded short acted sentences before the spontaneous session. In this part, we asked the subjects to utter the sentence written on the screen with a given emotion or mental state. The target emotions/mental states and the sentences are as follows:

(a) *Happiness:* You have won the lottery and you are telling it to a friend of yours.

I won! I won! I am rich now! I am rich!

(b) *Sadness:* You have to explain your friend that his father is passed away.

I don't know how to say that, it is better for you to sit down. Bad news… Your father… He is… He passed away.

(c) *Fear:* You are kidnapped and they are holding a gun towards you, you have to beg for your life.

Please don't kill me! I didn't do anything! Take all my money, I will do whatever you want, please don't kill me!

(d) *Anger:* You have caught the thief who has stollen your wallet.

Give it back! Who do you think you are stealing from? Give my wallet back!

(e) *Disgust:* You have discovered an insect in your soup.

Oww! Disgusting! Disgrace!

(f) *Confusion:* You didn't understand the lecture and asking to the lecturer.

Excuse me sir, could you explain that part again?

(g) *Boredom:* You have been waiting for a bus for at least an hour.

Come on! Where is this bus? I hope it comes soon.

(h) *Curiosity:* You want to learn your friend's secret.

Come on! I am not gonna tell that to anybody! Please, tell me!

3.1.2 Recording Environment

For the actual recordings, we designed a studio, by choosing the cameras and their locations, the lighting and the microphone. Below we describe this process in detail.

Cameras: First we had planned to record subjects with five different cameras. These cameras were: a Point Grey Bumblebee2 which is a stereo camera and planned to record the subject frontally, and Point Grey Fire Fly MV and/or Point Grey Grasshopper which would record from profile and halfprofile view. For these PointGrey Cameras, synchronization was done in two stages: pre and post processing. During pre-processing, Multisync, which is a software from Point Grey, provided the synchronization of data accusition. AutoIt [33] (see Figure 3-6), which is a script based software, also achieved synchronization by activating the camera and microphone recordings simultaneously. Another method, which we used later with the Sony HDR-XR200 camera is using a clapper board. All the cameras and the microphone were synchronized using the clapper board during post-processing by using various professional video editing softwares such as Sony Vegas [34], and AVID Studio [35]. Also during post processing, the synchronization was tested and if a frame drop was detected, it was corrected manually (see Figure 3-7).

Figure 3.6: AutoIt and MultySync during first tests

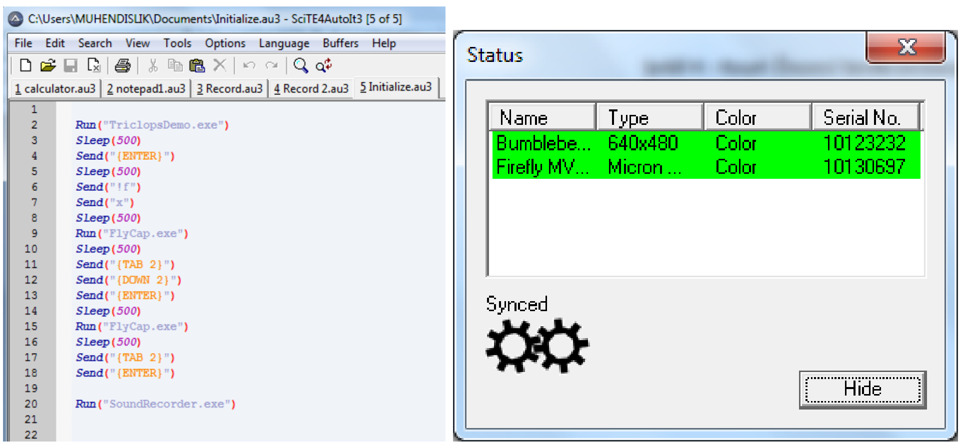


Figure 3.7: Post recording synchronization during first tests

Our first tests showed that instead of using Point Grey cameras, using camcorders will be more adventageous due to their better video quality and ease of synchronization. In the first system, cameras were capturing images and sending them to the computer via Firewire. Then we were processing them to create video files and then we were doing the synchronization process. If there is a frame drop occurred during the recording, it was hard to fix it manually. In the last setup, we decided to use Sony HDR-TD20 stereo camera instead of Point Grey Bumblebee 2 to record the subject frontally. To record from half profile, we decided to use Sony HDR-XR200 mono camera instead of Point Grey FireFly. By using the Sony cameras, cameras became independed from the computer and frame drop problem was solved. Also when we used Sony camcorders, we had the best results at post processing since the software Sony Vegas was fully compatible with the recordings of the Sony cameras. The syncronization of Sony camcorders using a clap board was easier than the old technique because there were no frame drops and there was no need for pre-recording synchronization. In Table 3-2 and Table 3-3, we compare the Point Grey and Sony cameras, which show that the spatial resolution of Sony cameras is better than Point Grey cameras, since they can record in HD resolution. In Figure 3.8, we compare the image qualities of Point Grey and Sony cameras. We can see that since Sony cameras perform some auto adjustments, the image quality is better since the face is better illuminated and the background color is closer to its actual color (green). We used a green cloth as the background.

Table 3.2: Comparison of BumbleBee2 and HDR-TD20 cameras.

	Camera	**Resolution**	**Frame Rate**	**Usage**
	BumbleBee2	Adjustable Max 648 x 488	Adjustable Max 48	With FlyCap software
	HDR-TD20	1920 x 1080	50 or 25 FPS	Automatic

Table 3.3: Comparison of Firefly and HDR-XR200 cameras.

	Camera	**Resolution**	**Frame/Rate**	**Usage**
	FireFly	Adjustable Max 752 x 480	Adjustable Max 30	With FlyCap Software
	HDR-XR200	1920 x 1080	25 FPS	Automatic

Figure 3.8: Comparison of outputs

TD20 (upper left), BumbleBee2 (upper right), XR200 (lover left), Firefly (lower right).

Microphone: Recording environment was an office, which was a slightly noisy environment we tried to choose a microphone, which should not be effected from the sound in the environment. One of the best choices was Rode NTG-2 (see Figure 3-9), which is a supercarioid microphone, which recorded the sound from one direction and was not so sensitive to the environmental sound. RODE NTG-2 also had a superior frequency response.

Figure 3-9: Rode NTG-2 microphone

Figure 3.10 The recording setup in the studio.

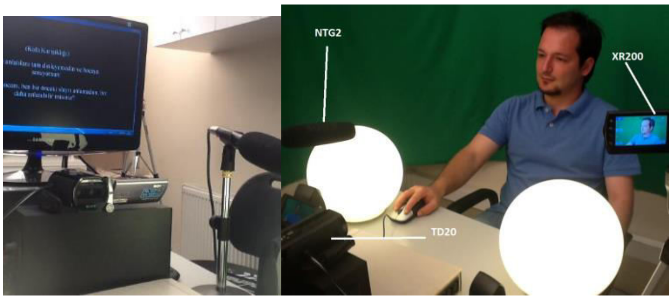
Illumination and background: In order to illuminate the room, we used 3 Read Head 1000Watt spot lights. We located them so as to minimize the shadows in the background and not to create flare on the subject. Indirect lighting has been used by reflecting the spotlights from the ceiling and front wall to create diffuse ligting as much as possible. In the media industry, green or blue background is commonly used in studios to seperate the foreground and the background easily. So we have decided to use a green cloth as the background.

Figure 3-11: 1000W Red Head Spot Lights

3.1.3 Recording Procedure

Before the recording, we explained each subject the procedure and warned them about possible disturbing scenes. Each subject first signed a consent form which states that the subject has understood and accepted the procedure and whether all recordings of the subject can be used and shared for research purposes. Then the subjects watched the 50 minute length stimuli and expressed their thoughts and feelings with their own words.

Figure 3.12: Studio during a recording session.



3.2 POST PRODUCTION

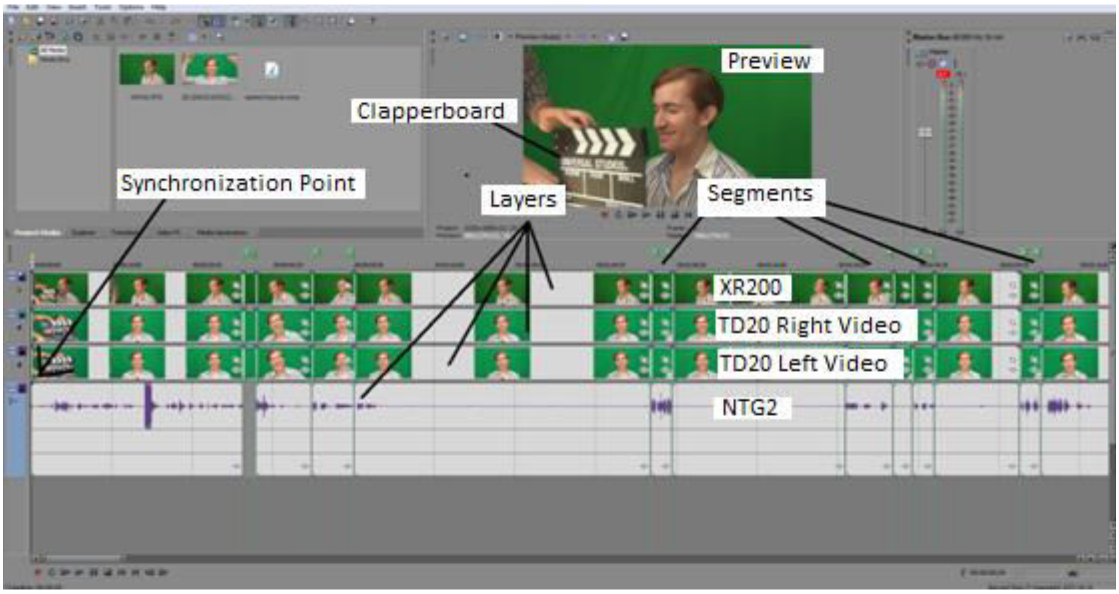
After the video of a subject is recorded. It is divided into smaller segments and then annotated. Below we give the details of the segmentation and annotation processes.

3.2.1 Segmentation

To segment the videos, Sony Vegas 11 software is used (see Figure 3-13). Because of the manufacturer of the software is same with cameras, the software is fully compatible with the recordings. When video and audio files are added to a project in the software, each of them is presented by a layer. At the beginning of each recording, clapper board was used to insert a sudden sign in both videos and audio tracks. In video the frame at which clapper board is closed is synchronzed with the instant in audio where there is a peak signal. By using that sign, synchronization of all data (two camereas and the microphone) can be done. This method has been used to synchronize video and audio in the movie industry since 1920s. The precision of the synchronization is 1 frame so it means 1/30 of a second (0.033 miliseconds).

After the synchronization process is done, videos are segmented to little video clips so that in each clip a single mental state or emotion exists. These clips are then rendered and given a name indicating the subject and video numbers.

Figure 3.13: Sony Vegas 11 User Interface



3.2.2 Annotation

Labels: As well as the 6 basic emotions, mental state labels are used to annotate the segments. The list of labels are we used are:

Neutral

Happiness

Sadness

Anger

Surprise

Disgust

Fear

Boredom

Contempt

Concentrating

Thinking

Unsure (including confusion, undecidedness)

Interest (including curiosity)

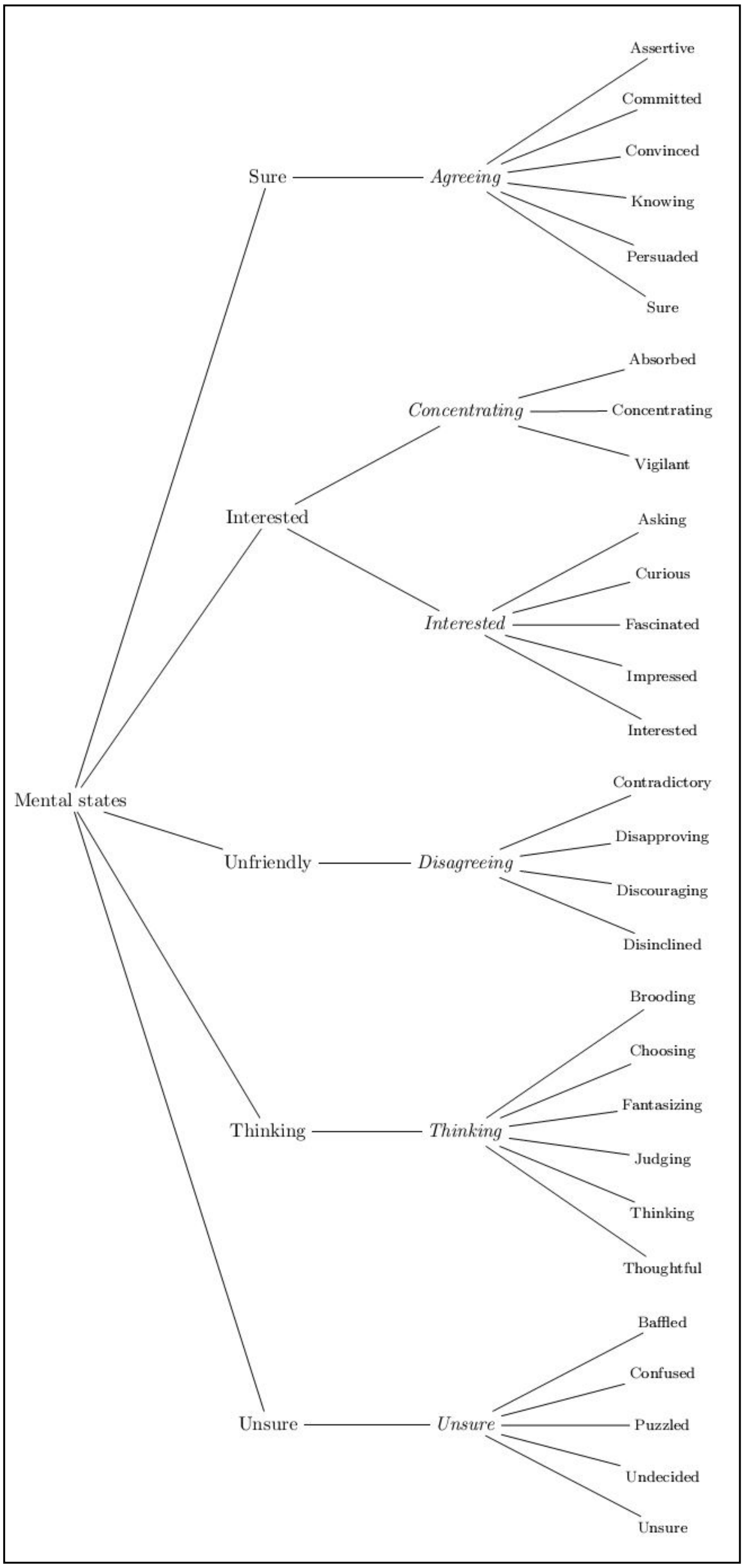
Bothered

In order to determine the labels that will be used, we used the the mental state categorization in the technical report entitled “Mind-reading machines: automated inference of complex mental states” [Kalioubi, 2005]. In this report, mental states are categorized to five categories as seen in Table 3-4 and Figure 3-14. And these categories have their own subcategories.

Table 3.4: Emotional categories used in [Kalioubi, 2005], the labels shown in bold are the ones we used.

Fear	**Anger**	**Boredom**	**Complaint**	Distrust	**Disgust**
Excited	Exaggerated	Wounded	**Interested**	**Happiness**	Kind
Liked	Romantic	Sneaky	Apologise	**Sad**	Sure
Surprised	**Thinking**	Impressed	Unfriendly	**Unsure**	Wanted

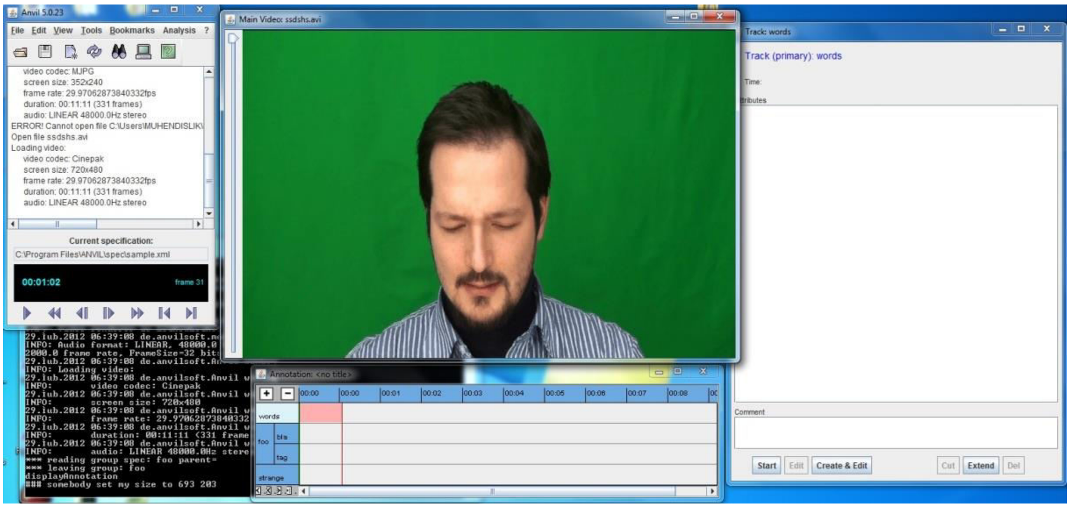
Figure 3.14: Mental State categories



Reference: Kalioubi, 2005

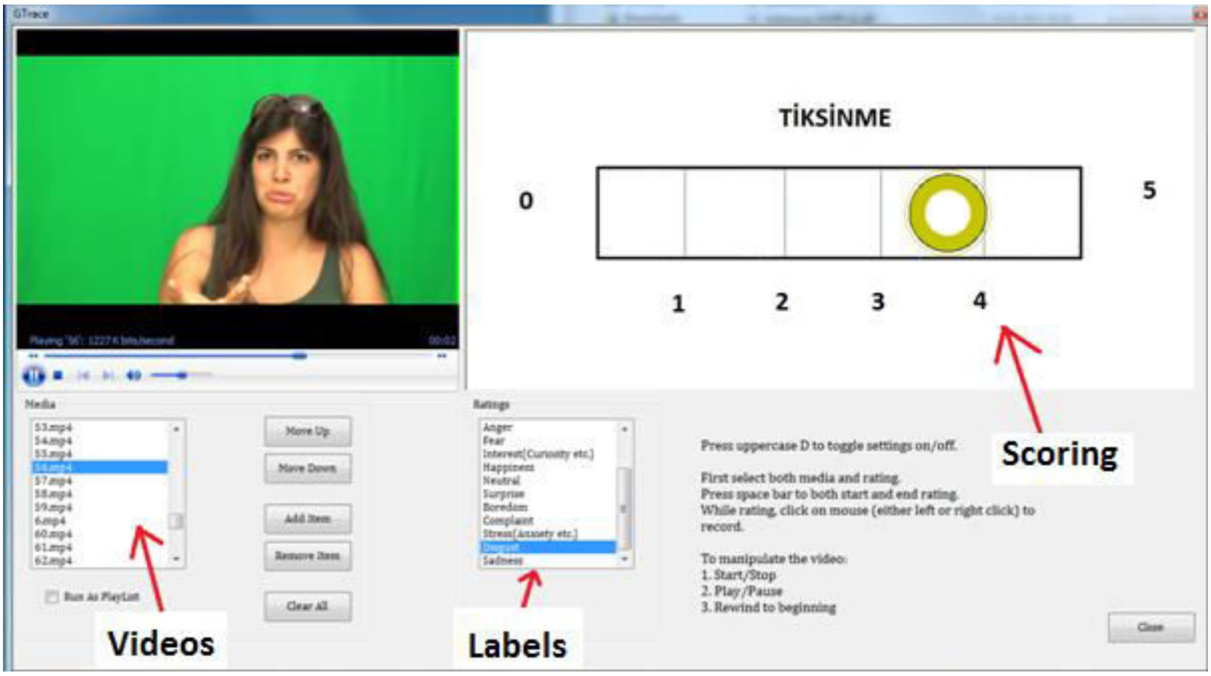
Annotation Tool: There are two common annotation tools: ANVIL and FEELTRACE. Although these two softwares have the same main purpose, the methodology of using them were too different. These two different methods has been tested for labeling and segmentation. These are “First label, than segment” and “First segment, than label” approaches. In the first method, the ANVIL annotaion tool is used (see Figure 3.15), which takes the whole video as input, and the labeler annotates while watching the whole record which is about 50 minutes long. Than, ANVIL outputs the time stamps of the related emotions. By using these data, it is possible to do segmentation. On the other hand, a standart segmentation could be done first and these segments can be labeled by using GTrace annotation tool , which is a new version of FEELTRACE. After testing both of them, we have decided to use GTrace, because it supports more video formats than ANVIL and also, its interface is more user friendly. Also it is possible to change and configure the emotion categories of GTrace.

Figure 3.15: User interface of the ANVIL Annotation Tool



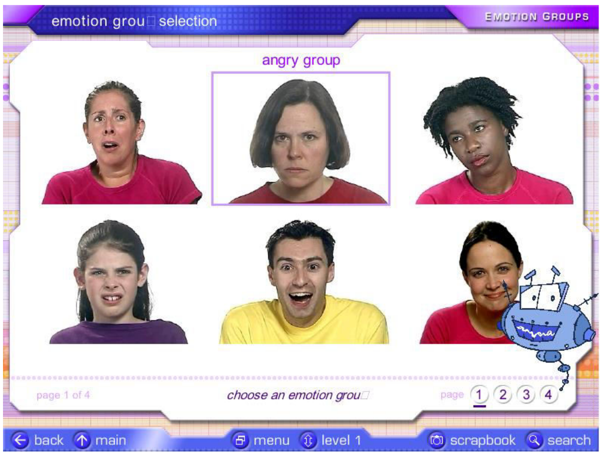
For labeling, we preferred GTrace software due to its user friendly interface. In this software, annotators are able to select the video clip that is going to be annotated from the lower left menu and the video is displayed at upper left of the interface. From the rating list menu, annotators are able to select the emotion that belongs to the selected clip and give it a score between 0 and 5 from the upper right screen.

Figure 3.16: User interface of the Gtrace annotation tool



Annotators: Annotators are the people who is watching, labeling and scoring the videos. Before the process, these people pass a small training session to have better understanding on emotions and mental states. This training was made by using Mind Reading software [Kingsley, 1996]. These software is a great reference work covering the entire spectrum of human emotions. By using these software it is possible to explore over 400 emotions, seeing and hearing each one performed by six different people.

Figure 3.17: Mind Reading Software



Reference: Kingsley, 2005

Inter-Annotator Agreement: We used the Kappa statistics **[Carletta, 1996]** to measure the agreement between annotators. This statistic eliminates the chance factor and tests the reliability of annotation of categorical data. It is a measure of agreement between annotators above the chance factor.

$$\text{KAPPA} = \frac{\text{Relative Observed Agreement Along Raters} - \text{Hypothetical Probability of Chance Agreement}}{1 - \text{Hypothetical Probability of Chance Agreement}}$$

Po=Relative Observed Agreement Along Raters

Pc= Hypothetical Probability of Chance Agreement

$$KAPPA = \frac{Po - Pc}{1 - Pc}$$

A Kappa calculator has been designed for 5 observers 14 classes by using MATLAB. All clips of the BAUM1a database (acted recordings), have been annotated by 5 different annotators. Kappa value has been calculated as 0.064, which is at a fair level. The best agreement was found between annotators 3 and 5, which has a Kappa value of 0.74. In Figure 3-18, we give the Kappa values for pairs of annotators. In Table 3-5 we give an example of an excel sheet, which shows the annotations of the first two evaluators.

The Kappa value of BAUM1s database (spontaneous recordings) was calculated as 0.54. It was expected that Kappa value for acted recordings would be greater than Kappa for spontaneous recordings. That is because mental states such as “confusion”, “unsure”, “thinking” are hard to annotate, and difficult to decide on a label.

Figure 3.18: KAPPA statistics for different pairs of annotators

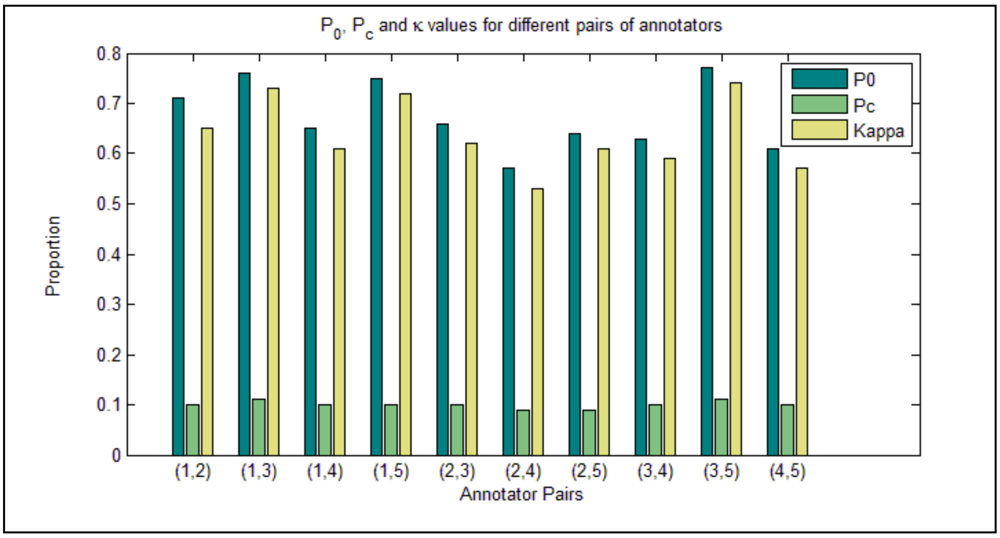


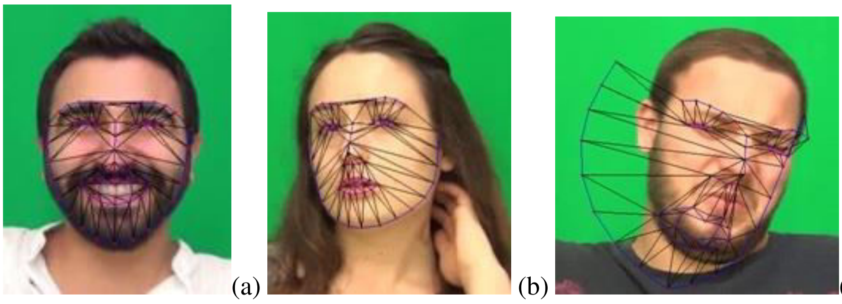
Table 3.5: Example Annotation Excel Sheet

Subject 1			Evaluator 1		Evaluator 2	
Video Number	Slang language?	Hand(1)/Head Gestures(2)	Emotion	Rating	Emotion	Rating
1	No	0	Anger	2,348879	Anger	3,572049
2	No	0	Disgust	4,509154	Disgust	4,134312
3	No	0	Boredom	3,266256	Boredom	2,408065
4	No	0	erest(Curiosity e	2,595486	Interest(Curiosity etc.)	3,098564
5	No	0	Unsure	4,084991	Concentrating	3,848248
6	No	0	Unsure	3,157749		
7	No	0	Complaint	3,187342	Concentrating	3,266256
8	No	0	Thinking	2,792771	Concentrating	3,927162
9	No	0	Thinking	3,729877	Unsure (Confusion, undecided etc.)	3,147885
10	No	0	Happiness	2,625079	Happiness	3,157749
11	No	0	Happiness	4,025805	Happiness	4,972775
12	No	0	Concentrating	3,443813	Concentrating	3,295849
13	No	0	Concentrating	2,960464	Concentrating	3,226799
14	No	0	Concentrating	3,601641	Sadness	2,506708
15	No	0	Anger	4,371054	Anger	2,506708

3.2.3 Facial Landmarks

In BAUM1 database, we also provide the geometric face feature point locations at each frame of all video sequences. These facial points are very useful in facial expression recognition algorithms. We used the facial point tracker by Jason Saragih **[Wang, 2010]** has been used to detect 66 points for each subject’s each video, which are shown as the nodes of the mesh shown in Figure 3-19. Saragih's Face Tracker is a C++ based open source software. The software needs the list of the file names and the video files should be located in the same folder as the list. Tracker detects the locations of facial landmarks and gives them as an output in txt file format.

Figure 3.19: The tracked facial points are at the nodes of the mesh.



(a) (b) (c)

(a),(b): Successfully tracked face examples, (c): Face tracking with errors

3.3 CONTENT AND FOLDER STRUCTURE OF THE BAUM-1 DATABASE

The database contains spontaneous audio-visual recordings which have been collected and named as BAUM-1 or BAUM-1s (**BA**hçeşehir **U**niversity **M**ultimodal Database 1) and a short acted recording for each subject named as BAUM-1a. There are 31 subjects (18 male, 13 female) and the age range is 18-66 (mostly between 20-30).

Some of the recordings are acted recordings and the collection of these acted recordings is called BAUM-1a. In this part of database, subjects acted through the scenario, uttered a given script with a target emotion. Collection of spontaneous recordings is called BAUM-1s. In this part of database, the subjects watched the stimuli and expressed their own feelings with their own sentences.

In Figure 3-20, we give examples from BAUM-1 database with various subjects and emotions. We can see that the videos are quie naturalistic. Some of them even contain spontaneous hand gestures (e.g. subject 10). In Table 3-6, we compare the BAUM-1 database with other databases in the literature. We can see that BAUM-1 is the only database with expressions of mental state. It is also the only database recorded in Turkish. In Figure 3-21, we provide examples from recordings of frontal and half profile views. In Table 3-7, we give an exract from the list of subjects in the database. In Table 3-8, we list the properties of BAUM-1 database.

Table 3.6: Comparison between existing databases and BAUM1

Database	*Emotion Content*	*Artificial / Natural*	*Number of Subjects*	*Record Length, Video and Audio Info*	*Language*	*Open sharing?* *Labeled?*
IEMOCAP **0, 0**	Extensive	A & N	10	12 hours	Eng	Partially open sharing (2 subjects) Yes
eNTERFACE'05 **[0, 0**	Basic Emotions	A	42	1166 short clips Video: 720x576 @25fps, Audio: 48kHz	Eng	Yes Yes
VAM **0**	Extensive	N	47	12 hours, Video:352x288 pixels @ 25fps, Audio: 16kHz	German	Yes Partially labeled
Humaine **0**	Extensive	A & N	unspecified	50 clips	Eng., Fr., Hebrew	Yes, Partially labeled with ANVIL (16 clips)
Semaine **0, 0**	Extensive	A & N	20 subjects, 24 records, 144 segment	6,5 hours Video: 580x780 pixels @ 50 fps, Audio: 48kHz	Eng	Yes, Partially labeled
Belfast Naturalistic Error! Reference source not found.	Extensive	N	298 clips (209 TV recordings, 30 interviews), 125 subjects (31 male, 94 female)	Record lengths 10-60 seconds.	Eng	Yes, Partially labeled with FEELTRACE
BAUM1	Emotions + Mental States	A & N	31	25 Hours of Video: 1080p 720x576 480p @30fps Audio: 96kHz	Tr	Yes Yes

The detailed explanation of database folder structure is below:

Annotations: In this folder, for each subject, there is an Excel file which includes annotation data (Subject**XX**Annotations.xlsx).

Audios: In this folder, there are subfolders named s1, s2, s3, etc. Each of these folders belong to a subject. Each of them contains audio clips which are in wav format.

Face Trackings: This folder contains txt files for facial feature locations for each video frame. There are two subfolders, one for Full HD videos and the other for low resolution videos.

- *Full HD:* Structure is same as SD Resolution folder, but the videos are in Full HD.
- *SD Resolution:* There is a folder for each person, and this folder contains txt files for each clip (for instance s1\01a.txt). Each row of these text files represents coordinates of 66 facial features of a frame.

Videos:

- *Full HD:* Structure is same as SD Resolution folder and video resolution is 1920x1080.
  - *SD Resolution:* For each person (exp. S1) there are videos wich have standard definition resolution, i.e., 720x576.
  - *S1*
    - *Mono:* Half profile video files (Sony XR200)
    - *Stereo:*
      - *Left:* Frontal left view (TD20)
      - *Right:* Frontal right view (TD20)
      - *SidebySide:* Side by side stereo records (TD20)

Figure 3.21: A subject's half profile (a), front stereo (b) images and speech data (c)

Table 3.7: A part of subject list

Subject Number	Subject Name	Subject Age	Gender	Nationality	Date of Recording	Consent for Publication?Yes/No	e of slang language?Yes/I
1	Berk Dumanhan	23	M	Turkish	21.06.2012	Yes	No
2	Hüseyin Samet Tan	24	M	Turkish	24.06.2012	Yes	No
3	Ömer Tura	26	M	Turkish	20.06.2012	Yes	No
4	Baturalp Şimşek	21	M	Turkish	04.07.2012	Yes	Yes
5	Mehmet Görmez	23	M	Turkish	20.16.2012	No	No
6	Doğukan Şentürer	22	M	Turkish	04.70.2012	Yes	Yes
7	Onur Karadeniz	21	M	Turkish	04.07.2012	Yes	No
8	Berk Deniz Yılmaz	22	M	Turkish	04.07.2012	Yes	Yes
9	Altuğ Kalelioğlu	23	M	Turkish	06.07.2012	Yes	Yes
10	Hazal Ege Güneyi	21	F	Turkish	09.07.2012	Yes	No

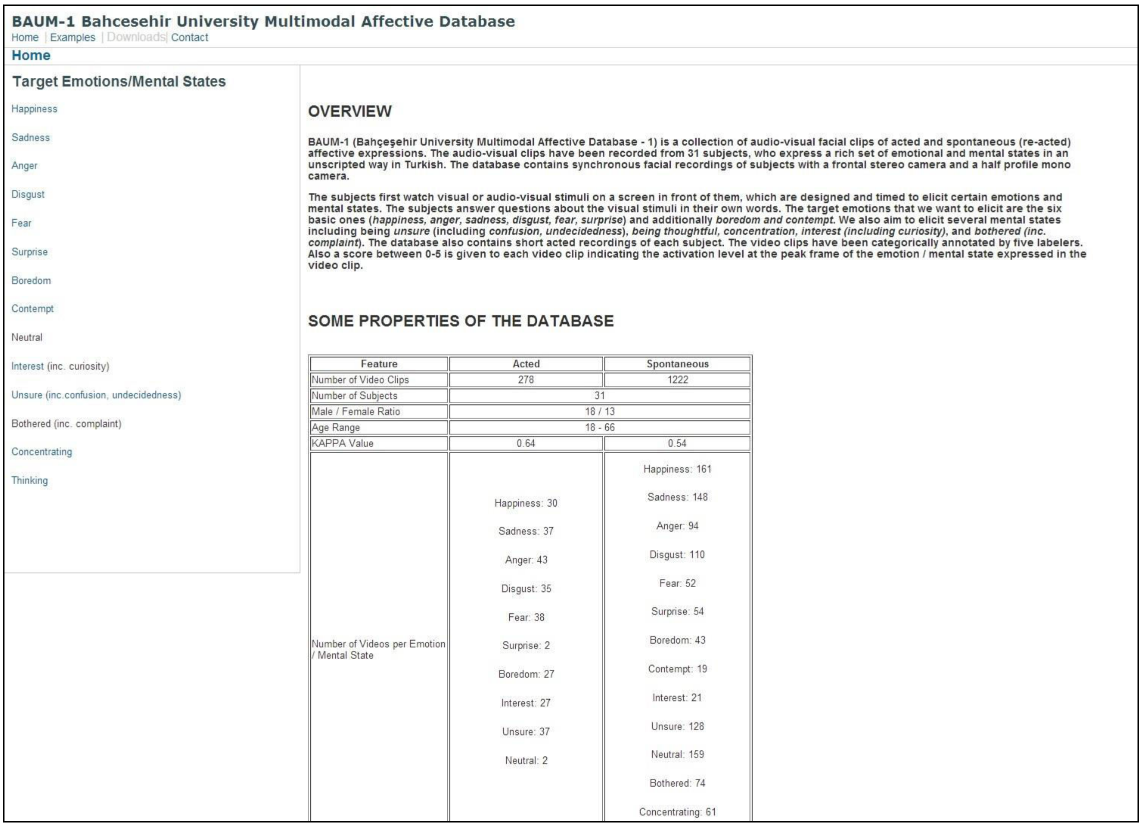
Table 3.8: Properties of BAUM-1 database

Feature	Acted Records (BAUM-1a)	Spontaneous Records (BAUM-1s)
Number of Videos	278	1222
Number of Subjects	31	
Male / Female ratio	18 M / 13 F	
Age Range	18-66	
KAPPA Value	0.64	0.54
Number of clips for each emotion and mental state	Happiness: 30 Sadness: 37 Anger: 43 Disgust: 35 Boredom: 27 Interest (Curiosity): 27 Fear: 38 Unsure (Confision): 37 Neutral: 2 Surprise: 2	Happiness: 161 Sadness: 148 Anger: 94 Disgust: 110 Boredom: 43 Interest (Curiosity): 21 Fear: 52 Unsure (Confusion): 128 Neutral: 159 Surprise: 54 Contempt: 19 Bothered: 74 Consantrated: 61 Thinking: 98

3.4 WEB SİTE OF BAUM-1 DATABASE

Our aim is to share the database with other researchers via a web site. The BAUM-1 web-Site has been designed by using Adobe Dreamviewer. The website includes example video and audio files, downloadable content, contact information and also helpful links to emotion definitions. The web site can be seen from: http://baum1.bahcesehir.edu.tr.

Figure 3.22: BAUM1 Web-Site



Feature	Acted	Spontaneous
Number of Video Clips	278	1222
Number of Subjects	31	
Male / Female Ratio	18 / 13	
Age Range	18 - 66	
KAPPA Value	0.64	0.54
Number of Videos per Emotion / Mental State	Happiness: 30 Sadness: 37 Anger: 43 Disgust: 35 Fear: 38 Surprise: 2 Boredom: 27 Interest: 27 Unsure: 37 Neutral: 2	Happiness: 161 Sadness: 148 Anger: 94 Disgust: 110 Fear: 52 Surprise: 54 Boredom: 43 Contempt: 19 Interest: 21 Unsure: 128 Neutral: 159 Bothered: 74 Concentrating: 61

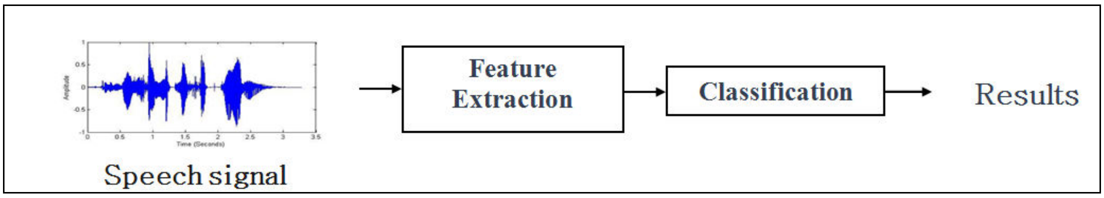
4. EMOTION RECOGNITION EXPERIMENTS ON BAUM-1

Below, we present preliminary emotion recognition experiments that we performed on BAUM-1 database. First, we will give results of the emotion recognition from speech experiments, which are done on the BAUM-1a (acted part) database. Then, we will present experimental results on peak frame selection from video sequences. Peak frames are the frames in a video at which the intensity of the emotional expression is maximum.

4.1 EMOTION RECOGNITION FROM SPEECH ON BAUM-1A

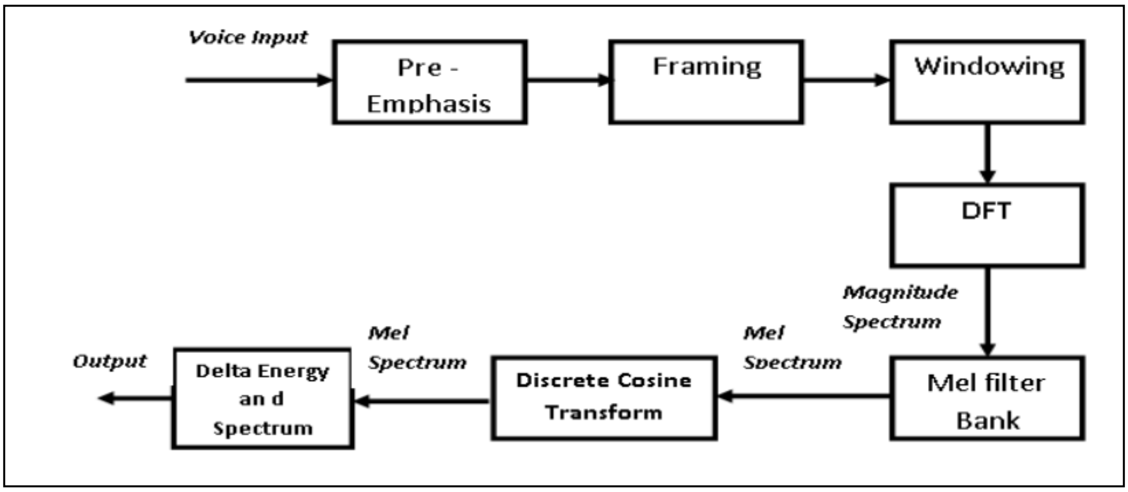
As the first emotion recognition experiment in BAUM-1 database, we tried to recognize the emotion from the speech information only. The overall flowchart of the procedure is given in Figure 4-1. The features that we used are the well-known MFCC (Mel-Frequency Cepstral Coefficients) [13] and RASTA-PLP [40] features are classified using Support Vector Machines (SVM) [41].

Figure 4.1: The procedure for emotion recognition from audio.



The mel-frequency cepstrum (MFC) is a representation of short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz [42]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A pitch is present on Mel Frequency Scale to capture important characteristics of phonetic in speech.

Figure 4.2: Steps of MFCC



MFCC consists of six computational steps. These are; Pre-emphasis, Framing, Hamming Windowing, Fast Fourier Transform, Mel Filter Bank Processing, Discrete Cosine Transform..

Step 1: Pre-emphasis

In this step, speech signal passes through a filter which emphasizes higher frequencies. This will increase energy of signal at higher frequency.

Step 2: Framing

In this step, the voice signal is divided into frames of N samples, adjacent frames are being seperated by M (M<N).

Step 3: Hamming Window

Hamming Window equation is given as:

N=number of samples in each frame

Y[N]=Output signal

X(n)=Input signal

W(n)=Hamming Window

Then the resulted signal is shown below;

$$Y(N) = X(n) \times W(n)$$

Step 4: Fast Fourier Transform

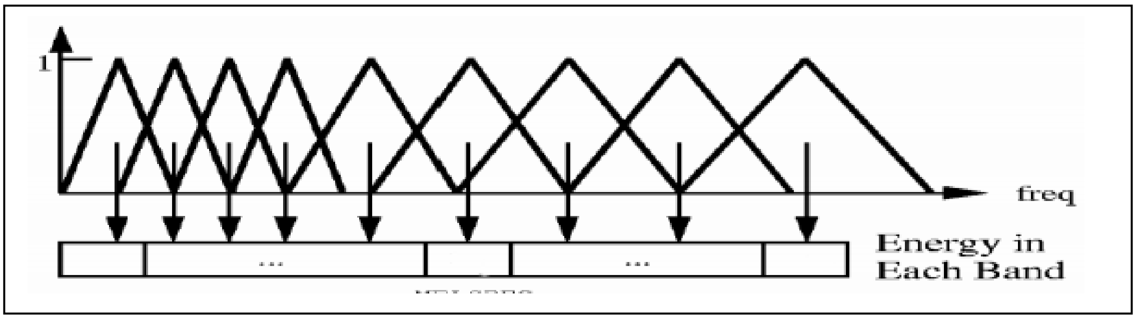
In this step, FFT is applied to convert each frame of N samples from time domain into frequency domain.

$$\mathrm{Y}(\omega) = \mathrm{FFT}\left[\mathrm{h(t)} * x(t)\right] = \mathrm{H}(\omega) * \mathrm{X}(\omega)$$

Step 5: Mel Filter Bank Processing

The bank of filters according to Mel Scale as shown in Fig4-3 is performed.

Figure 4-3: Mel Scale Filter Bank



This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [44], [45]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f:

$$F(Mel) = [2595 * \log 10[1 + f]700]$$

Step 6: Discrete Cosine Transform:

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conver sion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors.

On BAUM1a, firstle silence parts of the audio has been elliminated. Then MFCC (N=25ms, M=10ms) and RASTA PLP are calculated. 12 coefficient is used for MFCC and 13 coefficient is used for RASTA PLP. Also for each of them, first and second order derivatives are added to feature vector. Then these 9 statistics have been calculated: maximum, minimum, maximum location, minimum location, mean, variance, range, skewness, kurtosis. As a result, for each speech file, 675 dimensional vector which is combination of 12x3x9=324 dimensional MFCC, 13x3x9=351 dimensional RASTA PLP feature vectors, is obtained.

In the classification process, we used support vector machines with a kernel of degree 2. The results given below are subject independent since we used leave-one-subject-out (LOSO) cross validation. In Table 4.1, the confusion matrix for five emotions on BAUM-1a is given. We can see that sadness has the highest recognition rate (83.7 percent) and fear has the lowest recognition rate (55.2 percent). The average emotion recognition rate is 70.71 percent.

In Table 4.2, the confusion matrix for seven emotions/mental states are given, with the addition of boredom and interest. We can observe that again sadness has the highest recognition rate (81.08 percent) and boredom has the lowest recognition rate (40.7 percent). The average emotion recognition rate is 63.62 percent.

In order to compare the emotion recognition on BAUM-1a with another well-known database in the literature, we used the eNTERFACE database. This database has acted video recordings from 42 subjects in English. The emotions that exist in the database are the six basic emotions: anger, disgust, fear, happiness, sadness and surprise. The confusion matrix for emotion recognition from speech on eNTERFACE database is given in Table 4-3. We can observe that anger has the highest recognition rate (86.7 percent) and fear has the lowest recognition rate (62.3 percent). The average emotion recognition rate is 73.98 percent. This is in agreement with our average emotion recognition rate, since both databases are acted.

In order to test the emotion recognition performance across languages, we used eNTERFACE database for training and BAUM-1a for testing. The confusion matrix is given in Table 4-4. We can see that the recognition rates are quite low. Anger has the highest recognition rate (53.5 percent), and disgust has the lowest recognition rate (8.6 percent). The average emotion recognition rate is only 25.96 percent.

Table 4.1: Confusion matrix for emotion recognition from speech on BAUM-1a database for five emotions. The left column indicates the actual emotions. Numbers are in percentages. Average rate=70.71 percent

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**69.8**	4.7	11.6	4.7	9.3
Disgust	11.4	**71.4**	5.7	2.8	8.5
Fear	13.1	2.6	**55.2**	15.7	13.2
Happiness	10	3.3	6.7	**73.3**	6.7
Sadness	2.7	5.4	8.1	0	**83.7**

Table 4.2: Confusion matrix for emotion recognition from speech on BAUM-1a database for seven emotions. The left column indicates the actual emotions. Numbers are in percentages. Average rate=91.08 percent

	Anger	Disgust	Fear	Happiness	Sadness	Boredom	Interest
Anger	**69.8**	4.7	11.6	4.7	9.3	0	0
Disgust	11.4	**68.6**	2.9	2.9	8.6	2.9	2.9
Fear	7.9	2.6	**52.6**	21.05	10.52	2.6	2.6
Happiness	6.7	3.3	6.7	**73.3**	6.7	3.3	0
Sadness	0	2.7	8.1	0	**81.08**	2.7	5.4
Boredom	22.2	0	7.4	3.7	22.2	**40.7**	3.7
Interest	7.4	3.7	7.4	7.4	11.1	3.7	**59.3**

Table 4.3: Confusion matrix for eNTERFACE database. The left column indicates the actual emotions. Numbers are in percentages. Average rate=73.98 percent.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**86.7**	0.9	2.3	2.8	1.9	2.3
Disgust	3.7	**72.6**	7.4	6.1	6.1	4.2
Fear	6.9	6.1	**62.3**	6.9	9.3	8.4
Happiness	4.2	2.4	3.8	**80.2**	3.8	5.7
Sadness	4.2	4.2	5.1	5.1	**73.5**	7.9
Surprise	4.7	3.3	7.4	6.5	12.6	**65.6**

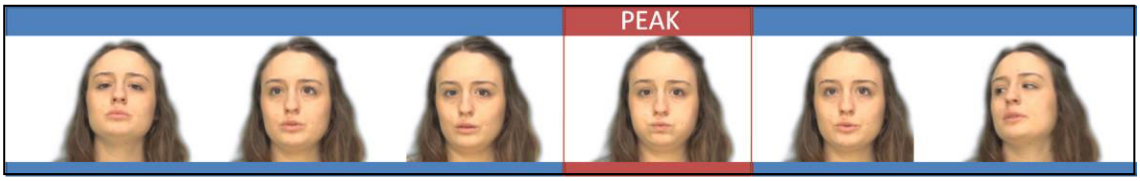
Table 4.4: Confusion matrix for emotion recognition from speech using five emotions when eNTERFACE database is used for training and BAUM-1a is used for testing. The left column indicates the actual emotions. Numbers are in percentages. Average rate = 25.96 percent.

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**53.5**	9.3	4.7	20.9	11.6
Disgust	8.6	**8.6**	20	37.1	25.7
Fear	52.6	5.3	**10.5**	26.3	5.3
Happiness	36.7	10	26.7	**16.7**	10
Sadness	13.5	13.5	24.3	8.1	**40.5**

4.2 Peak Frame Selection for Emotion Recognition from Video

When a video clip contains a certain emotion, and the goal is to recognize the emotion from facial expressions, some frames reflect the emotion better than other frames (see Figure 4-4) . Therefore, if we detect the peak frames first, then use them for emotion recognition from facial expressions, our emotion recognition rates will be higher as compared to using non-peak frames.

Figure 4.4: Example Peak Frame for boredom



Given a video reflecting a single emotion, our goal is to detect the frames in the video, which reflect the emotion with the maximum intensity, i.e. the peak frames. Below, we present a method (called as MAXDIST) to detect the peak frames based on a dissimilarity matrix computed using all the frames. The advantages of the method are: it does not need any prior training and the facial features used for peak frame detection can also be used for facial expression recognition.

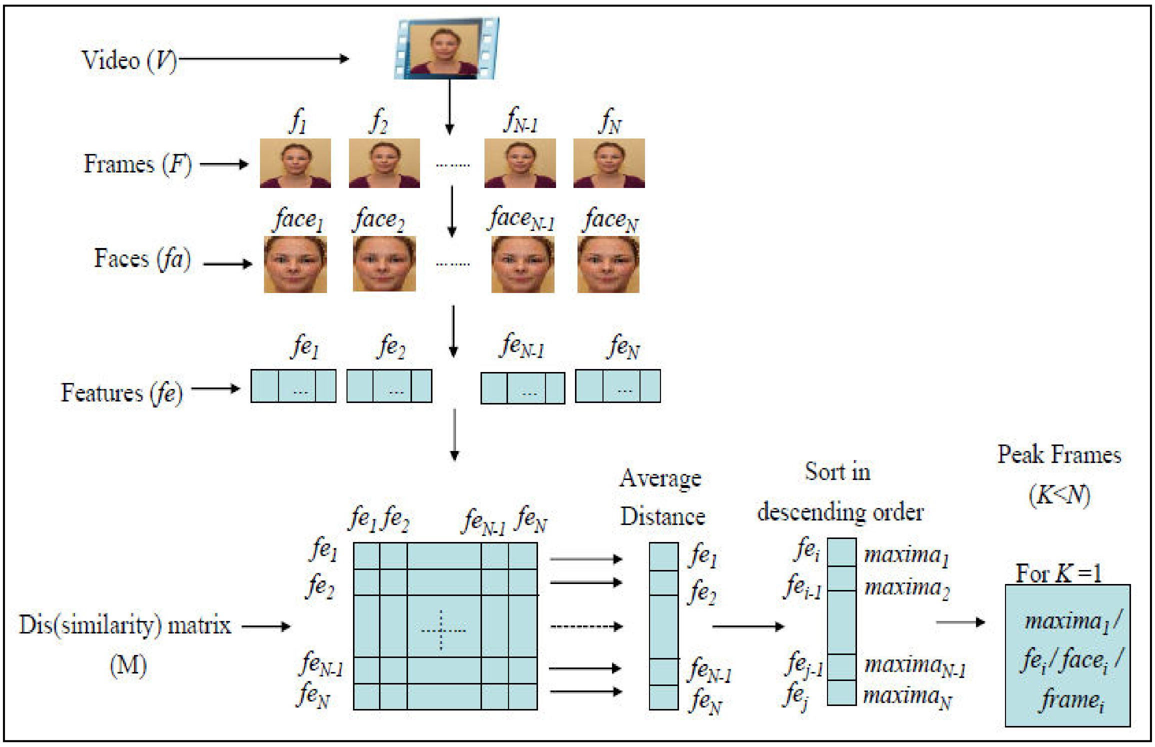
The first step of the method is to detect and align the face at each frame. From the tracking points, we already know the location of face. By using outer points of the face, it is possible to crop the face, so that unnecessary details about the background and the hair is eliminated, as shown in the third row of Figure 4-2. By applying neccessary transformations to each image, all frames are alligned while reference point is at nose and the distance between eyes is protected on each face. Then face feature vector of each vector is created. Here, we have used local phase quantization (LPQ) [43] features in our experiments. LPQ is calculated by quantizing phase of Discrete Fourier Transform (DFT) computed in local image windows.

After an LPQ based feature vector is calculated for each frame, a dissimilarity matrix M is generated for the video as shown in the last row of Figure 4-2. Each element of the dissimilarity matrix M(i,j) represents the Euclidean distances between feature vector of frame i and feature vector of frame j.

$$d(\mathrm{p},\mathrm{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The first row of matrix M then contains the distance of frame 1 from all the other frames in the sequence. The next step is to find the average of each row and order these averages in descending order. After sorting in descending order, the frame with maximum average feature vector distance to the others, is selected as peak frame. If the average distances are very close to each other, K peak frames can also be selected.

Figure 4.5: Steps of the MAXDIST Algorithm



The steps of the algorithm can be summarized as follows:

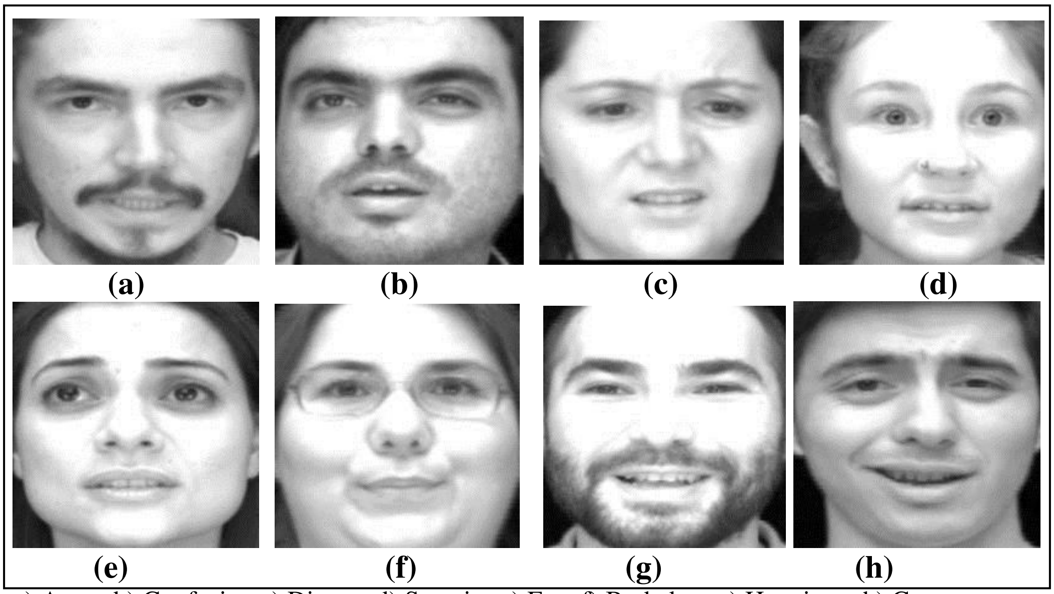
i Given a video sequence with N frames, generate the dissimilarity matrix, M, where each element M (i,j). i,j ∈ {1,2....N} denotes the distance between facial features vectors of frame i and frame j.

ii For each frame (i.e. row j), compute is average distance score with respect to the other (N-1) faces (frames). Order these values in descending order.

ii Choose K peak frames that have the largest average distance scores.

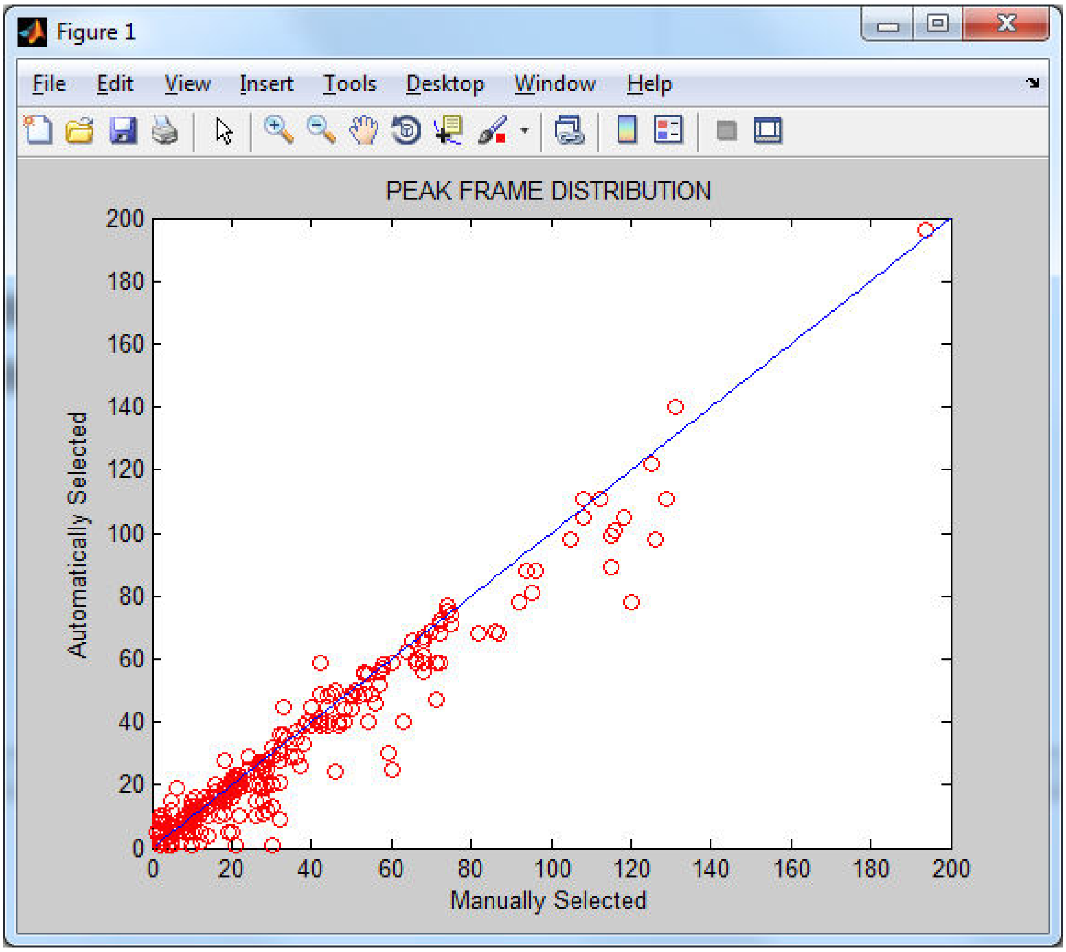
We tested the proposed method on BAUM1a (acted part of the database). In Figure 4.6, we give examples of peak frames selected from videos containing various emotions.

Figure 4.6: Example of automatically selected peak frames



a) Anger b) Confusion c) Disgust d) Surprise e) Fear f) Bodedom g) Happiness h) Contempt

Figure 4.7: Comparison of automatically and manually selected peak frames.



In order to test the MAXDIST method, we compare the peak frames selected manually and automatically. In the Figure 4-7 we give the manually selected peak frame numbers versus automatically detected peak frame numbers (red circles). The blue line represents the ideal situation where the manually selected and automatically detected frame numbers are the same. We can observe that most of the peak frames are located in the first 80 frames of a video and there are many red circles are close to the blue line as well as deviations. The mean of the error is 5.2362 frames and standard deviation of the error is 6.5662 frames. Since the neighbor frames of a peak frame look very similar and the average total number of frames in all video is 109.58; the maximum error ratio is around 10%. Also the correlation coefficient of these two data is 0.9691 which is very close to 1.

$$\rho_{\mathrm{X,Y}} = \mathrm{corr}(\mathrm{X,Y}) = \frac{\mathrm{cov}(\mathrm{X,Y})}{\sigma_{\mathrm{X}}\sigma_{Y}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_{\mathrm{X}}\sigma_{Y}}$$

5. CONCLUSIONS AND FUTURE WORK

In this thesis, we collected an audio-visual spontaneous emotional Turkish database named as BAUM-1 (**BA**hçeşehir **U**niversity **M**ultimodal Database 1). There are 31 subjects (18 male, 13 female) in the database and age range of subjects is 18-66, although most subjects are in the range 20-30. The collection of spontaneous recordings is called as BAUM-1s. In this part of database, subjects watched a carefully designed stimuli video that evokes various emotions and mental states and expressed their own feelings with their own sentences. The database also contains acted recordings called as BAUM-1a. In this part of database, subjects were given a scenario and uttered given sentences with the target emotions.

BAUM1 contains about 1500 video and audio clips in different formats in HD and SD resolutions, which have total length of over 25 hours. The database is novel as it contains spontaneous recordings of mental states as well as emotions. To the best of our knowledge, this is the first spontaneous audio-visual database recorded in Turkish. We share the database with the research community via a web site and we hope that the database will be useful for researchers who work in affective computing.

We have done preliminary emotion recognition from speech experiments in the acted part of the database. Audio-visual affective and mental state recognition methods on the database is ongoing.

REFERENCES

Books

Carletta, 1996, Assesing agreement on classification tasks: The kappa statistic, *Computational Linguistics,* pp 249-254.

Cowie, 2003, Emotional speech: towards a new generations of databases, *Speech Communication,* pp 33-60.

Periodicals

Bozkurt, 2003, Formant position based weighted spectral featuresfor emotion recognition, *Sciene Direct*, Speech Communication 53 (2011) 1186–1197

Busso, 2008, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database”, *Journal of Language Resources and Evaluation,* Vol.42, No.4, pp. 335-359.

Chen, 1998, Emotion recognition from audiovisual information in Proc. *IEEE Workshop on Multimedia Signal Processing*, (Los Angeles, CA, USA), pp. 83–88.

Chen, 2000, Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction, *PhD dissertation*, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering,

Clavel, Fear-type emotion recognition for future audio-based surveillance systems, *Science Direct*, Speech Communication 50 (2008) 487–503

Dhall, 2011, Emotion Recognition Using PHOG and LPQ Features*, Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011 IEEE International Conference

Dumas, 2009, Emotional Expression Recognition using Support Vector Machines, Universty of California, San Diago

Ekman, 1978, Facial Action Coding System: Investigator’s Guide. *Palo Alto, CA: Consulting Psychologists Press*.

Emilia I. Barakova “*Emotional Recognition in Robots in a Social Game for Autistic Children”*, Eindhoven University of Technology

Gilke, 2012 , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues, Madhavi Pednekar, MFCC-based Vocal Emotion Recognition Using ANN, International Conference on Electronics Engineering and Informatics

Grimm, 2008, “The vera am mittag German audio-visual emotional speech database”, Proceedings of ICME08.

Gutierrez, 2012, Associations Between Facial Emotion Recognition, Cognition and Alexithymia in Patients with Schizophrenia, *Comparison of Photograhic and Virtual Reality PresentationsAnnual Review of Cybertherapy and Telemedicine*, , pp. 88-92

Hermansky,1994, Rasta Processing of Speech*, IEEE Transactions on Speech and Audio Processing*, Vol.2, No.4, October 1994

Jungum, 2009, Emotions in Pervasive Computing Environments, *IJCSI International Journal of Computer Science Issues*, Vol. 6, No. 1

Kaliouby, 2005, Mind-reading machines: automated inference of complex mental states, Cambridge, March 25th

Khulage, July 2013 , Analysis of speech under stress using linear techniques and non-linear techniques for emotion recognition system, *International Journal of Engineering Science and InnovativeTechnology (IJESIT)*, Volume 2, Issue 4

Kostoulas, The Effect of Emotional Speech on a Smart-Home Application, University of Patras, Greece

Martin, 2006, The eNTERFACE'05 Audio-Visual Emotion Database, *Proceedings of the 22nd International Conference on Data Engineering Workshops.*

Mase, 1991, Recognition of facial expression from optical flow*, IEICE Transactions*,vol. E74, pp. 3474–3483, October 1991.

McKeown, The Semaine Corpus of Emotionally Coloured Character Interactions, ICME08.

Miyasato, 1997, Facial emotion recognition using multi- modal information, in Proc. IEEE Int. Conf. on Information, *Communications and Signal Processing* (ICICS'97), (Singapore), pp. 397–401.

Muda, 2010, Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, *Journal of Computing*, Volume 2, Issue 3, March 2010, ISNN 2151-9617

Price, Design an automatic speech recognition system using maltab, University of Maryland Estern Shore Princess Anne.

Shen, 2009, Affective e-Learning: Using Emotional Data to Improve Learning in Pervasive Learning Environment. *Educational Technology & Society*, 12 (2), 176–189.

Garrison, 1913, An introduction to the history of medicine. Philadelphia & London: W. B. Saunders. p. 571.

Sneddon, 2007, The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data, *ACII 2007*, LNCS 4738, pp. 488–500

Wang (2009). Affective e-Learning: Using Emotional Data to Improve Learning in Pervasive Learning Environment. *Educational Technology & Society*, 12 (2), 176–189.

Black, 1995, Tracking and recognizing rigid and non-rigid facial motionsusing local parametric models of image motion, in Proc. *International Conf. Computer Vision*, (Cambridge, USA), pp. 374–381.

Wang, 2010, 'Non-rigid Face Tracking with Enforced Convexity and Local Appearance Consistency Constraint , *International Journal of Image and Vision Computing* (IVC).

Zeng, 2005, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39-58.

Other Publications

Anvil Labeling Software: http://www.anvil-software.de/

eNTERFACE'05 database: http://www.enterface.net/results/

Feeltrace Labeling Software:
http://emotion-research.net/toolbox/toolboxlabellingtool.2006-09-29.8782805660/

Humaine Database: http://humaine-db.sspnet.eu/

IAPS, International Affective Picture System: http://csea.phhp.ufl.edu/Media.html

IEMOCAP database: http://sail.usc.edu/iemocap/

Semaine Veritabanı*:* www.**semaine**-project.eu/

AutoIt, 2005, Jonathan Bennett & **AutoIt** Team http://www.autoitscript.com/site/autoit/

Avid Studio, 2011, Avid Technology Inc: http://www.avid.com/US/

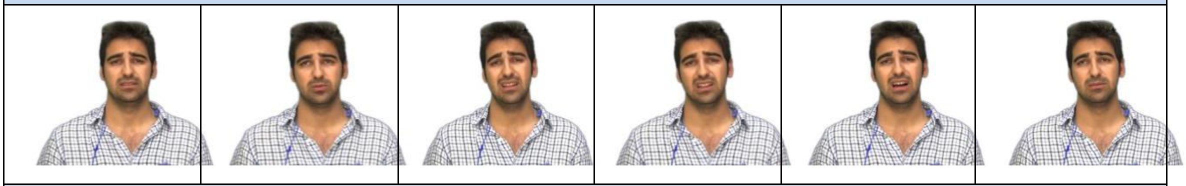
Mind Reading Software, 2003, Jessica Kingsley Publishers: http://www.jkp.com/

Sony Vegas, 2010, Sony Creative Software: http://www.sonycreativesoftware.com

APPENDICES

Appendix A.1 Examples from BAUM1

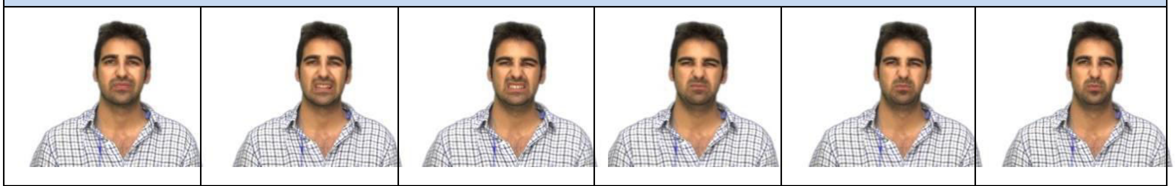
Subject 1 / Video 1 / Anger (3.57)



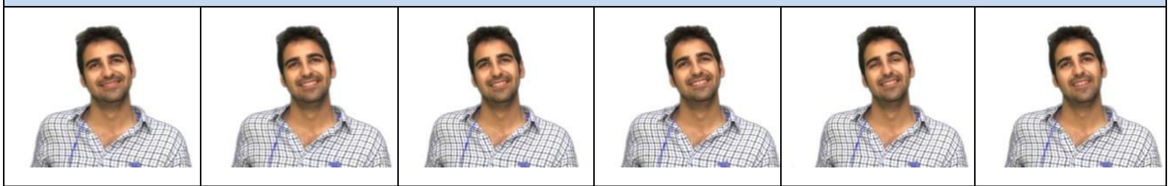
Subject 1 / Video 3 / Boredom (2.4)

Subject 1 / Video 8 / Concentrated (2.92)

Subject 1 / Video 4 / Curiosity (3.09)



Subject 1 / Video 2 / Disgust (4.13)



Subject 1 / Video 10 / Happiness (3.15)

Subject 2 / Video 5 / Anger (3.88)

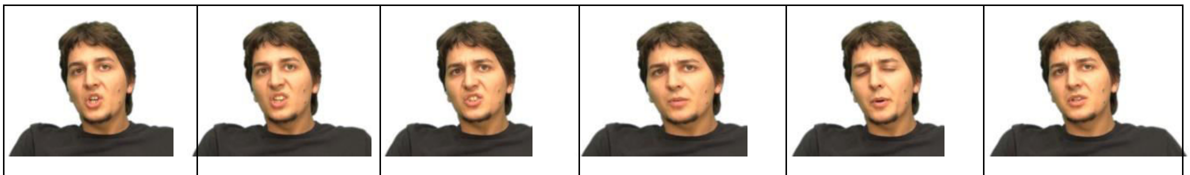
Subject 2 / Video 6 / Disgust (4.66)

Subject 2 / Video 2 / Happiness (2.51)

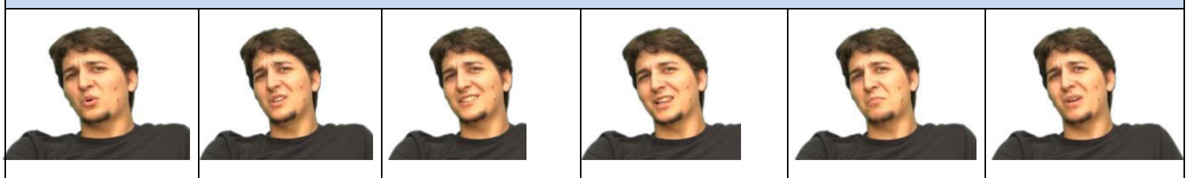
Subject 2 / Video 19 / Neutral (3.14)

Subject 2 / Video 14 / Confusion (3.34)

Subject 2 / Video 27 / Sad (3.04)



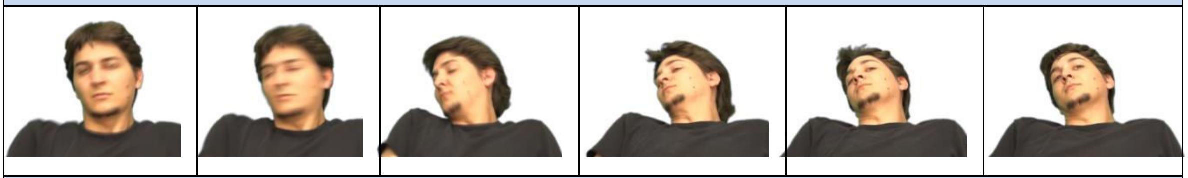
Subject 4 / Video 5 / Anger (3.88)



Subject 4 / Video 34 / Bothered (2.81)

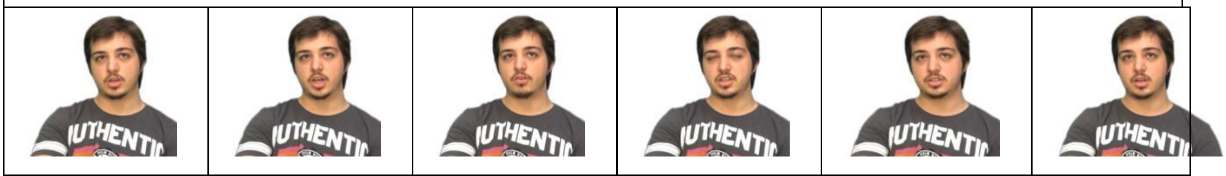
Subject 4 / Video 20 / Contempt (4.13)

Subject 4 / Video 6 / Disgust (4.66)

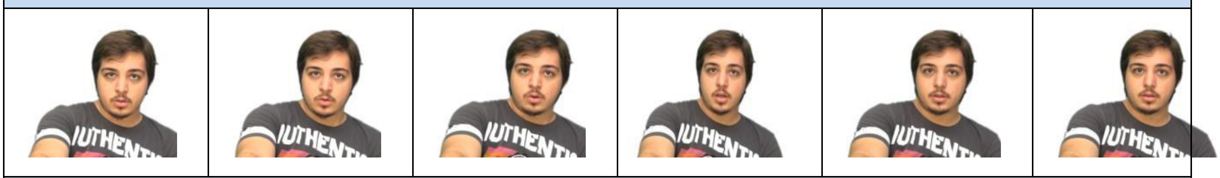


Subject 4 / Video 43 / Fear (4.70)

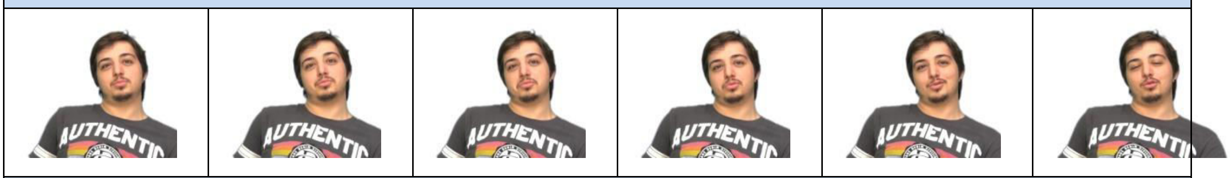
Subject 4 / Video 17 / Happiness (3.99)



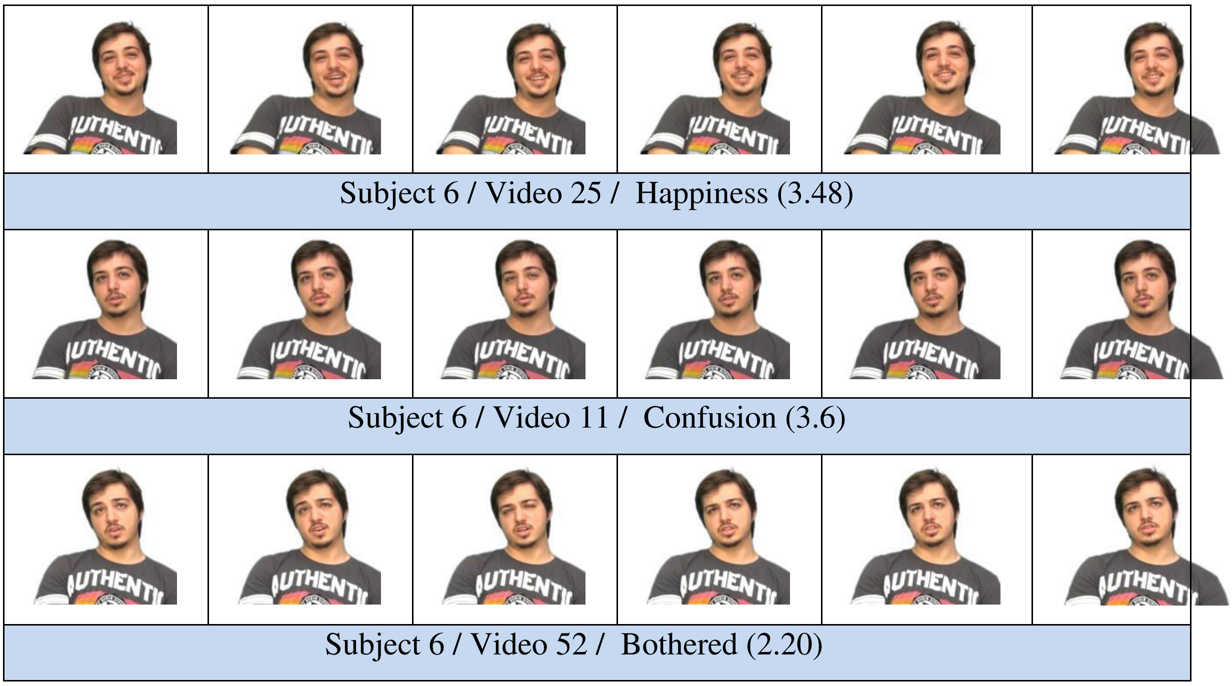
Subject 6 / Video 5 / Anger (2.41)

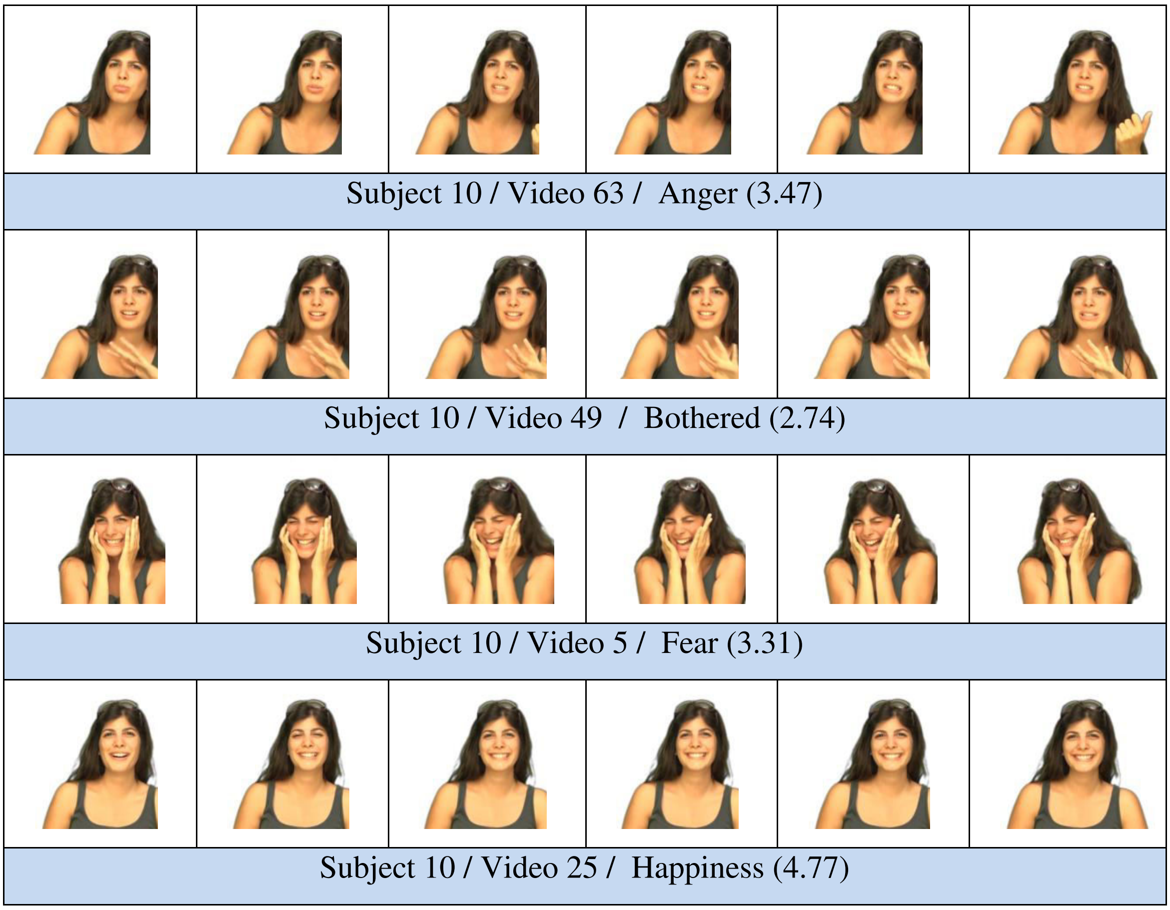


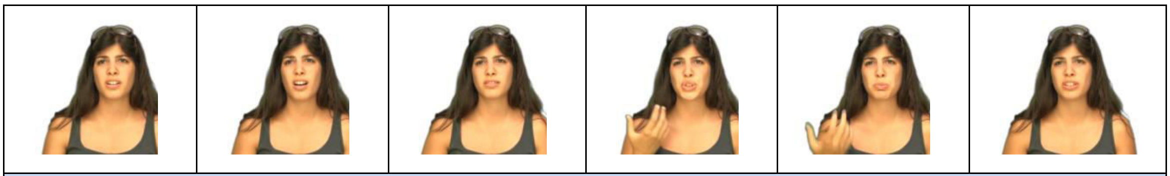
Subject 6 / Video 24 / Boredom (4.01)



Subject 6 / Video 26 / Contempt (2.10)


Subject 6 / Video 25 / Happiness (3.48)
Subject 6 / Video 11 / Confusion (3.6)
Subject 6 / Video 52 / Bothered (2.20)


Subject 10 / Video 63 / Anger (3.47)
Subject 10 / Video 49 / Bothered (2.74)
Subject 10 / Video 5 / Fear (3.31)
Subject 10 / Video 25 / Happiness (4.77)

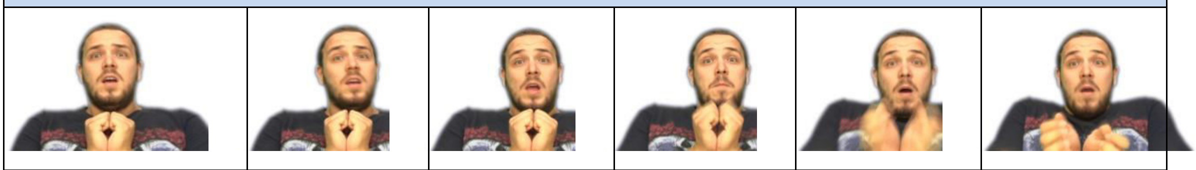


Subject 10 / Video 61 / Sadness (2.14)

Subject 10 / Video 27 / Thinking (4.05)

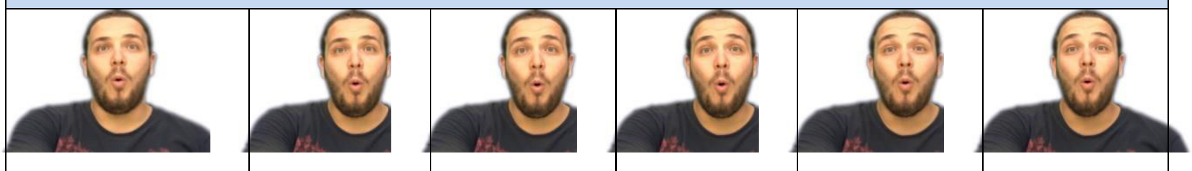
Subject 8 / Video 7 / Anger (4.92)

Subject 8 / Video 9 / Disgust (4.23)



Subject 8 / Video 4 / Fear (4.53)

Subject 8 / Video 22 / Happiness (4.90)



Subject 8 / Video 43 / Surprise (4.84)

Subject 8 / Video 42 / Thinking (4.09)

Appendix A.2 Araştırmaya Katılım Onay Formu

İnsan-Bilgisayar Etkileşimi için Konuşma ve Yüz İfadelerinden Spontan Duygu Tanıma

BahçeşehirÜniversitesi Elektrik-Elektronik Mühendisliği bölümünde Doç. Dr. Çiğdem Eroğlu Erdem tarafından yürütülmekte olan, TÜBİTAK tarafından desteklenen bir araştırma projesine katkıda bulunmak üzere davet edilmiş bulunmaktasınız. Bu çalışmaya katkınız tamamen gönüllülük esasına dayanmaktadır. Çalışmaya katılmaya onay vermeden once, aşağıda yazan bilgileri dikkatle okuyunuz ve anlamadığınız yerler hakkında sorular sorunuz.

- **ÇALIŞMANIN AMACI**

Şu anda yapmakta olduğumuz çalışmanın amacı, insanların spontan duygu ifadelerini görsel ve işitsel olarak kayıt altına almaktır. Bu kayıtlar daha sonra araştırmacılar ile paylaşılacak ve otomatik duygu tanıma yöntemlerinin geliştirilmesinde kullanılacaktır.

- **YÖNTEM**

Eğer bu çalışmaya katılmayı kabul ederseniz aşağıdakileri yapmanız istenecektir:

- Karşınızdaki ekranda size resimler ve videolar izleyeceksiniz. Lütfen dikkatinizi tamamen izlediğiniz görüntüye verin ve kendinizi o görüntüde rol alıyormuş gibi düşünün. Görüntüler ve videolar, izleyenlerde birbirinden farklı duygular uyandırmak amacıyla seçilmişlerdir.
- Bir resim ya da videonun izleme süresinin sonuna doğru, o görüntü hakkındaki duygu ve düşüncelerinizi anlatmanız istenecektir.
- Bir görüntüyü **anlatırken** mümkünse ekranın üzerinde yer alan **stereo kameraya bakmaya gayret ediniz.**
- Kayıt süresince sandalyenizi oynatmayınız, öne ve arkaya fazla eğilmeyiniz, ve ellerinizi yüzünüzden uzak tutmaya çalışınız.
- Görüntü hakkındaki duygu ve düşüncelerinizi anlatırken, neler hissettiğinizi mümkün olduğunca yüz ifadelerinize ve ses tonunuza yansıtarak, fakat abartmadan ya da azaltmadan ifade ediniz. Kullandığınız kelimeler çok da önemli değildir.
- Kayıt sırasında yüzünüz ve omuzlarınız farklı açılardan kaydedilecektir.

- **OLASI RAHATSIZLIKLAR**

  - Bazı görüntüleri izlerken rahatsızlık hissedebilirsiniz. Önemli olan sizden anlatmanız istendiğinde, olumlu ya da olumsuz duygu ve düşüncelerinizi, içtenlikle ifade etmenizdir.

- Eğer çok fazla rahatsızlık hissederseniz kayıtlara istediğiniz an son verebilirsiniz. Böyle bir durumda lütfen Onur Önder'e hemen bilgi veriniz.

- **BU ÇALIŞMANIN OLASI TOPLUMSAL YARARLARI**

Otomatik duygu tanımanın insan-bilgisayar etkileşimi, güvenlik, sağlık ve e-eğitim gibi pek çok alanda uygulaması vardır. Bu tür uygulamalar için, kişinin fiziksel (yorgun, enerjik vb.), duygusal (üzgün, kızgın, mutlu vb.) ve zihinsel (dikkatli, kafası karışık vb.) durumunu kestirip, en uygun tepkiyi veren sistemler geliştirilmesi gereklidir.

- **GİZLİLİK**

Lütfen aşağıdaki ikisoru için uygun kutuyu işaretleyin.

Evet

Hayır

	Evet	Hayır
1. Bana ait kayıtların diğer araştırmacılarla paylaşılmasına onay veriyorum.	☐	☐
2. Bana ait görüntülerin bilimsel yayınlarda yer almasına onay veriyorum.	☐	☐

- **KATILIM VE SONA ERDİRME**

Bu çalışmada yer alıp almamaya karar veriniz. Eğer gönüllü olarak yer almaya karar verirseniz, herhangi bir zamanda hiçbir sorumluluk almadan kayıtları sona erdirme hakkınız vardır. İstemediğiniz soruları yanıtlamama hakkınız vardır.

Yukarıda anlatılanları anladım. Bütün sorularıma yeterince yanıt verildi ve bu çalışmada yer almayı kabul ediyorum. Bu formun bir kopyası bana da verildi.

____
Ad - Soyad

______________________________ ____________________
İmza Tarih

______________________________ ____________________
Tanık İmzası Tarih

Appendix A.3 MATLAB CODES

<u>wordcount.m</u>

```
function counts=wordcount(a)


input=a;
% find the unique elements in the input
uniqueNames=unique(input)';
% use string comparison ignoring the case
occurrences=strcmpi(input(:,ones(1,length(uniqueNames))),uniqueNames(ones(length(i
nput),1),:));
% count the occurences
counts=sum(occurrences,1);

for i=1:length(counts)
   disp([uniqueNames{i} ': ' num2str(counts(i))])
end
```

<u>kappa.m</u>

```
%% 5 OBSERVER, 14 CLASS KAPPA CALCULATOR
%% Reading Excell
[A,B]=xlsread('kap.xlsx');

Annot1=B(:,1);
Annot2=B(:,2);
Annot3=B(:,3);
Annot4=B(:,4);
Annot5=B(:,5);
%% Word Frequencies
[a b]=size(Annot5)
count1=wordcount(Annot1)/a;
count2=wordcount(Annot2)/a;
count3=wordcount(Annot3)/a;
count4=wordcount(Annot4)/a;
count5=wordcount(Annot5)/a;
%% How many agreements that we have?
x12=0; x13=0; x14=0; x15=0; x23=0; x24=0; x25=0; x34=0; x35=0; x45=0;

for i=1:a;
   if size(Annot1{i})==size(Annot2{i})
   if Annot1{i}==Annot2{i}
```

```
    x12=x12+1;
end
end

if size(Annot1{i})==size(Annot3{i})
if Annot1{i}==Annot3{i}
    x13=x13+1;
end
end

 if size(Annot1{i})==size(Annot4{i})
if Annot1{i}==Annot4{i}
    x14=x14+1;
end
 end

 if size(Annot1{i})==size(Annot5{i})
if Annot1{i}==Annot5{i}
    x15=x15+1;
end
 end

 if size(Annot2{i})==size(Annot3{i})
if Annot2{i}==Annot3{i}
    x23=x23+1;
end
 end

 if size(Annot2{i})==size(Annot4{i})
if Annot2{i}==Annot4{i}
    x24=x24+1;
end
 end

 if size(Annot2{i})==size(Annot5{i})
if Annot2{i}==Annot5{i}
    x25=x25+1;
end
 end

 if size(Annot3{i})==size(Annot4{i})
if Annot3{i}==Annot4{i}
```

```
        x34=x34+1;
    end
     end

     if size(Annot3{i})==size(Annot5{i})
    if Annot3{i}==Annot5{i}
        x35=x35+1;
    end
     end

      if size(Annot4{i})==size(Annot5{i})
    if Annot4{i}==Annot5{i}
        x45=x45+1;
    end
     end
end
%% PA: Observed Percentage Agrement, PE=Random Agreement
pa12=x12/a;          %Number of agreements / total
pe12=sum(count1.*count2); %Multiply word freqs elementally and sum them up to find
PE
k12=(pa12-pe12)/(1-pe12);

pa13=x13/a;
pe13=sum(count1.*count3);
k13=(pa13-pe13)/(1-pe13);

pa14=x14/a;
pe14=sum(count1.*count4);
k14=(pa14-pe14)/(1-pe14);

pa15=x15/a;
pe15=sum(count1.*count5);
k15=(pa15-pe15)/(1-pe15);

pa23=x23/a;
pe23=sum(count2.*count3);
k23=(pa23-pe23)/(1-pe23);

pa24=x24/a;
pe24=sum(count2.*count4);
k24=(pa24-pe24)/(1-pe24);
```

```
pa25=x25/a;
pe25=sum(count2.*count5);
k25=(pa25-pe25)/(1-pe25);

pa34=x34/a;
pe34=sum(count3.*count4);
k34=(pa34-pe34)/(1-pe34);

pa35=x35/a;
pe35=sum(count3.*count5);
k35=(pa35-pe35)/(1-pe35);

pa45=x45/a;
pe45=sum(count4.*count5);
k45=(pa45-pe45)/(1-pe45);

            KAPPA=(k12+k13+k14+k15+k23+k24+k25+k34+k35+k45)/10
```

CURRICULUM VITAE

FULL NAME: Onur Önder

ADDRESS: Kuştepe Mahallesi Yavrukuş Sokak No: 1/1 d:3 Silyanoğlu Apartmanı Şişli İstanbul

BIRTH PLACE / YEAR: İzmir / 1988

LANGUAGE: Turkish (native), English

HIGH SCHOOL: İzmir Kız Anadolu Lisesi, 2006

BS: Electrical & Electronics Engineering, Ege University, 2010

MS: Electrical & Electronics Engineering, Bahçeşehir University, 2014

Full support by both Bahçeşehir University (TA) & TÜBİTAK

NAME OF INSTITUTE: Natural and Applied Sciences

NAME OF PROGRAM: Electrical & Electronical Engineering

WORK EXPERIENCE: September 2010 – February 2013
Bahçeşehir University, Teaching Assistant
October 2013 – ongoing
Homeprof, Ultrametrik, Innovation Engineer

ARTICLES IN CONFERENCES

O. Önder, C. E. Erdem, M. Irak, "**A Re-acted Audio-Visual Emotional Database**", *Affective Computing for Mobile HCI Workshop*, İstanbul, Turkey, Sep. 2012.

O.Önder, S.Zhalehpour, C.E. Erdem, "**A Turkish Audio-Visual Emotional Database**", *SIU 2013* (Best Application Paper Finalist), Cyprus, April 2013.