ÇANKAYA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

Project Report

Version 1

CENG 408

Innovative System Design and Development II

201917
Implementation of an Audio-Visual Emotional Recognition System

Merve DADAŞ – 201511016
Furkan KARADAŞ – 201511033
Uğur BAYEZİT – 201514013 (ECE)
Aydın ŞİŞMAN – 201514213 (ECE)

Advisors: *Assoc. Prof. Dr. Hadi Hakan MARAŞ &*
Dr. Lecturer Selma ÖZAYDIN

Table of Contents

İçindekiler

Abstract 4
Özet 4
1. Introduction 5
1.1 Contribution 5
2. Literature Search 6
2.1 Dataset 7
2.2 Preprocessing 7
2.3 Audio Preprocessing 7
2.4 Visual Preprocessing 7
2.5 Feature Extraction 8
2.5.1 Audio Part 8
2.5.2 Visual Part 9
2.6 Fusion 9
2.7 Model Training 10
2.8 Experiment Result 11
2.9 Conclusion 11
3 Summary 12
3.1 Summary of Conceptual Solution 12
3.2 Technology Used 12
4 Software Requirements Specification 12
4.1 Introduction 12
4.2 Overall Description 15
4.3 Requirements Specification 16
5 Software Design Description 21
5.1 Introduction 21
5.2 Design Overview 23
5.3 Architecture Design 23
5.4 Use Case Realization 29
5.5 Detection 29
6 Test Plan 30
6.1 Introduction 30
6.2 Features to be tested 31
6.3 Item Pass/Fail Criteria 31
6.4 References 31
6.5 Test Design Specification 31
7 Detailed Test Cases 35
7.1 GUI.SG 35
7.2 GUI.LG 36
7.3 AD.CNT.BT 36
7.4 AD.EV.BT 37
7.5 AD.EF.BT 37
7.6 AD.EA.BT 37
7.7 AD.LST.BT 38
7.8 AD.SHW.BT 38
7.9 AD.ADD.BT 38
7.10 AD.EDT.BT 39

7.11 AD.DEL.BT 39
7.12 AD.LSTUS.BT 40
7.13 AD.SHWUS.BT 40
7.14 AD.ADDUS.BT 40
7.15 AD.EDTUS.BT 41
7.16 AD.DELUS.BT 41
7.17 AD.SYSTR.BT 41
7.18 AD.ANLZ.BT 42
7.19 U.CNT.BT 42
7.20 U.EV.BT 43
7.21 U.EF.BT 43
7.22 U.EA.BT 43
7.23 U.LST.BT 44
7.24 U.SHW.BT 44
7.25 U.ADD.BT 44
8 Conclusions 45
9 Future Works 45
Acknowledgement 45
References 45

Abstract

People express emotions through differently. Utilization of both verbal and nonverbal communication channels allows to create a system in which the emotional state is expressed more clearly and therefore easier to understand. In this report describes the emotion recognition project consisting of audio and visual. Emotion recognition will be performed with features extracted from speech and video. In the development of the project, image processing, signal processing and artificial intelligence technologies were used. The audio and video were processed separately and then combined with the fusion algorithm.

Key words: Image processing, speech processing, signal processing, deep learning, machine learning, classification

Özet

İnsanlar duyguları farklı şekillerde ifade eder. Hem sözel hem de sözel olmayan iletişim kanallarının kullanılması, duygusal durumun daha net bir şekilde ifade edildiği ve bu nedenle anlaşılması daha kolay olan bir sistem yaratılmasını sağlar. Bu rapor sesli ve görsel olan duygu tanıma projesini açıklar. Konuşma, görüntü ve videodan elde edilen özellikler ile duygu tanıma gerçekleştirilecektir. Projenin geliştirilmesinde görüntü işleme, sinyal işleme ve yapay zekâ teknolojileri kullanılmıştır. Ses ve video ayrı olarak işlendi ve sonra füzyon algoritması ile birleştirildi.

Anahtar Kelimeler: Görüntü işleme, konuşma işleme, sinyal işleme, derin öğrenme, makine öğrenmesi, sınıflandırma

1. Introduction

Modern day security systems rely heavily on bio-informatics, like as speech, fingerprint, facial images and so on. Besides, determination of a user’s emotional state with facial and voice analysis plays a fundamental part in man-machine interaction (MMI) systems, since it employs non-verbal cues to estimate the user’s emotional state. Therefore, recognizing human emotion has been an attractive task for data scientists. On the other hand, there are many challenges in emotional data evaluation such as collection of proper datasets, definition of number of emotions to recognize, selection of the labelled data, etc. Due to the many challenging tasks under evaluation, MMI systems that utilize multimodal information about their users' current emotional state are interest of the computer vision and artificial intelligence communities.

In this study a software algorithm will be implemented for extracting emotion related features from image and speech signals. Then we will infer an emotional state by designing a rule-based decision algorithm. Open source software algorithms will be utilized to implement the recommended system. The project will be directed as an interdisciplinary study and will be carried out as a joint work with a group in the department of Computer Engineering. Emotional image recognition and emotional speech recognition systems will be designed separately. In this scope, ECE Department's students will implement emotional speech recognition part of the proposed system, and CENG Department's students will implement the emotional facial recognition part of the system. As an interdisciplinary study, image and speech related systems are combined with a decision algorithm. For this purpose, a theoretical study will be directed for decision algorithms for audio visual recognition systems. After implementation of the audio-visual emotional system, image and speech related features can be extracted from an input audio visual signal. Afterwards, emotional situation of a user will be estimated. The developed system will be tried on some English dataset and the performance of the system will be tested.

A software program will be designed and developed for an Audio-Visual Emotional Recognition System. The system will be able to receive and process not only a human voice but also his/her face image in the form of recorded signals and will present information about the emotional state as an output.

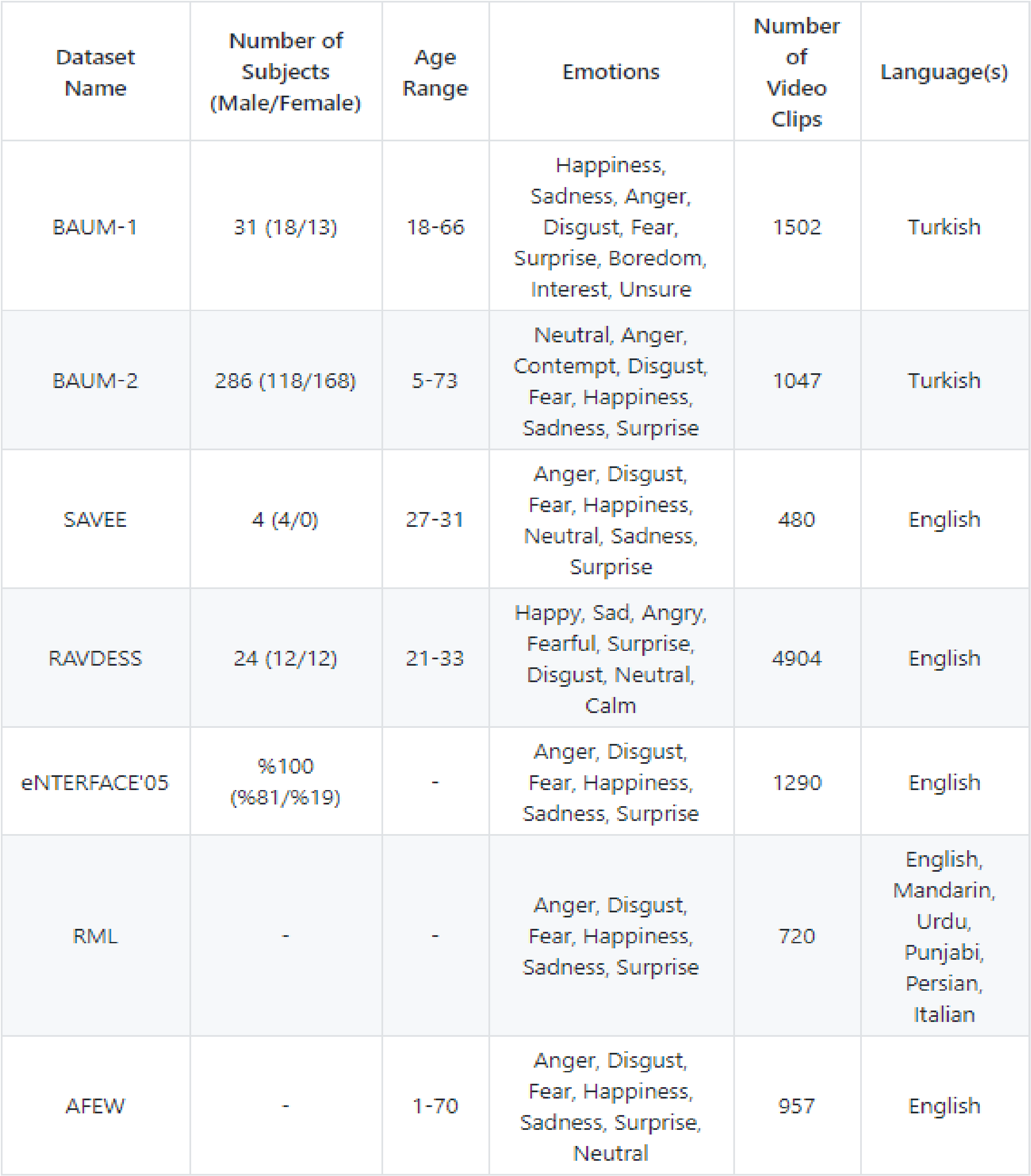
1.1 Contribution

This software system will be performed emotion recognition from audio, audio-visual video. With the easy-to-use user-interface of the system, the user can either record instant video/real time or upload an existing video to the system and perform emotion recognition. This system allows big corporate companies to measure customer satisfaction and perform the necessary analysis.

2. Literature Search

A literature review was conducted before the project was developed and details are given below. eBook Collection (EBSCOhost), Academic Search Complete, IEEE Xplore Digital Library were used for literature review and internet researched with filters.

Table I – Dataset Information



Dataset Name	Number of Subjects (Male/Female)	Age Range	Emotions	Number of Video Clips	Language(s)
BAUM-1	31 (18/13)	18-66	Happiness, Sadness, Anger, Disgust, Fear, Surprise, Boredom, Interest, Unsure	1502	Turkish
BAUM-2	286 (118/168)	5-73	Neutral, Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise	1047	Turkish
SAVEE	4 (4/0)	27-31	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise	480	English
RAVDESS	24 (12/12)	21-33	Happy, Sad, Angry, Fearful, Surprise, Disgust, Neutral, Calm	4904	English
eNTERFACE'05	%100 (%81/%19)	-	Anger, Disgust, Fear, Happiness, Sadness, Surprise	1290	English
RML	-	-	Anger, Disgust, Fear, Happiness, Sadness, Surprise	720	English, Mandarin, Urdu, Punjabi, Persian, Italian
AFEW	-	1-70	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	957	English

2.1 Dataset

To accomplish research on audio-visual affect recognition, appropriate datasets are needed. A large amount of research in audio-visual emotional recognition has been conducted with private datasets. However, a few audio-visual video datasets were made available publicly for the research community in recent years.

The process to acquire the audio-visual videos follows similar steps for most of the datasets. These are contained synchronous facial recordings of subjects with a frontal stereo camera and a half profile mono camera [1]. The target emotions that generally intended to elicit are the six basic ones that are anger, sadness, happiness, disgust, surprise, fear. Besides be aimed to elicit several mental states that are confused, thinking, concentrating, interested, and complaining.

Table I presents a summary of the most used audio-visual emotion recognition datasets.

2.2 Preprocessing

The preprocessing process, which is important in many problems, has an important place in the detection of emotion recognition from audio-visual. In many studies that audio-visual emotional recognition is seen that visual preprocessing and audio preprocessing are examined as two separate cases. The main preprocessing techniques at the end of the research are as follows.

2.3 Audio Preprocessing

Mel spectrogram is used to be obtained in the preprocessing part of the audio part. The signals are divided into 40 milliseconds. The dividing frames are multiplied by the hamming window. A fast Fourier transform is applied. Then, a 25 bandpass filter is applied, and MEL scale. Then, logarithm function is applied to the filter outputs to suppress the dynamic range. Previous outputs are arranged to generate the MEL spectrogram of the signal [2].

Audio signals are preprocessed to reduce background noise. Voice Activity Detector Technique is used. Short-time zerocrossing rate (STZCR) and short-time energy (STE) features, and these steps follow that the speech signal x(m) in the time domain is divided into n frames, and the STZCR is calculated with the weighted average. After that, STZCR and STE are compared to determine whether the signal is present and finally the unwanted signals are discarded from the frames to be processed for feature extraction [4].

2.4 Visual Preprocessing

The visual part consists of videos. The process to obtain the visual preprocessing methods follows similar steps. All videos in the dataset are generally divided into the same number of frames. Algorithms for frame selection are applied. The face region in the frame is cropped. After this processing steps that frame is converted grayscale and then the frame is resized. Thus, preprocessing is performed for the visual part. The following are some of the algorithms and methods used in the preprocessing process.

The video is divided a certain number of frames, the histogram of each frame is calculated and then, the chi-square distance is used to find the difference between consecutive frames. Before the histograms are calculated, and face region in the data set is crop using the viola face recognition algorithm. If the face doesn't find in frame, this frame is ignored and continues the next frame. After the keyframe is selected, the frame is converted grayscale. The mean normalization, LBP and IDP also calculated per the keyframe and are selected keyframe. Thus, the frames to be used for feature extraction are determined [2].

Face detection and localization are should be performed before image processing to remove unwanted background information. Viola-Jones (VJ) algorithm was proposed for this process from Kah Phooi Seng, et al. [4].

Egils Avots, Tomasz Sapinski et al. preprocessing part, the video is divided into frames for visualbased features, that the purpose in doing so select mainframes from a video. When select mainframes, it is used difference of frame that is said as the sum of the difference between pixels. A pair of images provides the similarity ratio for frames. The system averages the difference for the last 10 frames. If the new frame has a different value that is smaller than average 1.5, the frame is skipped. This operation is made to skip frames automatically. The Viola-Jones Algorithm is applied to cut the face area, from the main selected mainframes [6].

2.5 Feature Extraction

Audio-visual emotional recognition has been studied from many perspectives, yielding multiple alternatives for feature extraction. It is examined in two groups that are audio preprocessing and visual preprocessing as that below.

2.5.1 Audio Part

M. S. Hossain and G. Muhammad proposed a 2D CNN architecture. They proposed 4 convolution layers and 3 pooling layers. Then, last layer is a fully connected neural network with two hidden layers. After that, they applied A SoftMax function the output. The output of the SoftMax is then fed into the fusion part [2].

Carl Busso, Zhigang et al. proposed a Maximum Likelihood Bayes classifier (MLB), K-nearest neighboorhood (KNN) and Kernel Regression (KR) for feature extraction. MFCC were used to train the Hidden Markov Model (HMM). They proposed, recognized the four emotions, and six archetypal emotion classifications of power coefficients used 12 MEL-based train the Markov model [3].

Egils Avots, Tomasz Sapinski et al. when extracting the feature of the audio part, centers on the non-linguistic properties of the audio signal. Then, they have extracted MFCCs, which are calculated for a 400-millisecond moving window with a step size of 200 millisecond.

Thus, the property vector has obtained. For MFCC, there are used parameters that are pre emphasis 20 filter bank channels, coefficient 0.97, 13 cepstral coefficients, 3700 Hertz upper-frequency limit, and 300 Hertz lower frequency limit [6].

Wang and Guan et al. proposed employ intensity, pitch, and the first 13 MFCC features on audio feature extraction [7].

2.5.2 Visual Part

M. S. Hossain and G. Muhammad proposed a 3D CNN architecture for video signals. They used 8 convolution layers, 5 pooling layers and, 2 fully connected layers. A softmax layer follows the fully connected layers and also include one filters. The given input to the model is 16 keyframes as an RGB color and then resized to 227 × 227. Output of the softmax is then fed into the fusion part [2].

Carl Busso, Zhigang et al. 10-dimensional feature vector has used to dynamic model HMM and then during feature extraction has split the data into five blocks that are the eyebrow, forehead, low eye, left cheek and right cheek area. Then defined a local source of coordinates for each frame. Then provided these by reducing data collected on each frame of the video to a 4-D property vector [3].

Li-Minn Ang, Kah Phooi Seng et al. have proposed made feature extraction for a face with the approach found in the steps below. A feature extraction technique has used Bi-directional Principal Component Analysis (BDPCA) and Least Square Linear Discriminant Analysis (LSLDA) to differentiate and extract the visual features among 6 emotion classes [4].

Egils Avots, Tomasz Sapinski, et al. have made that tagging facial images according to emotions before feature extraction and then, trained using Convolutional Neural Network (CNN - AlexNet Architecture) [5]. The transfer learning approach is used when training. Then, images are randomly transformed in X and Y directions in the range of -30 to 30 pixels to ensuring that CNN learns general features. For CNN hyper tuning, most significant parameters can be found that are Bias Learn Rate Factor is 20, Weight Learn Rate Factor is 20, Mini Batch is that 10 Initial Learn Rate is 1e-4, Max Epochs is 10, Validation Patience is Inf and, Validation Frequency is 3 [6].

Wang and Guan et al. used the Gabor filter bank of 8 orientations and 5 scales, and obtain Gabor coefficients for each facial image. Gabor coefficients are included in Local Binary Patterns (LBP) and Local Phase Quantization. From visual features also used Long ShortTerm Memory (LTSM) and CNN for video feature extraction [7].

2.6 Fusion

The data obtained from audio and video files are classified by the fusion algorithm. The fusion algorithms of the articles that we reviewed in this section are explained below.

M. S. Hossain and G. Muhammad have proposed ELM model for fusion. The ELM has a feed-forward network, and single hidden layer. In this proposed system, they used two ELM's. After feature extraction has implemented, the outputs have obtained are given to the input of the ELM. ELM-1 has 100 neurons, ELM-2 is that 250. The outputs have using the softmax function and then, fed into the model training part [2].

There are two approximations to fusion that are feature level and score level. In this study, score level approach has preferred. A sliding window has applied to the speech signal for the audio path. With this window, audio files are processed continuously, and emotions are defined. A similar method was applied in the visual path [4].

To achieve a single prediction result in the fusion section, Egils et al. made a value of 6 points corresponding to the accuracy they had previously classified from audio and visual. The highest of these possibilities represents the necessary emotional state, the label. Thus, the decision level algorithm is combined [6].

Zeng et al. proposed to use a Multi-stream Fused Hidden Markov Models (MFHMM) to applied model level fusion. The Multi-stream Fused Hidden Markov Models (MFHMM) has pieced together bi-modal knowledge from visual-audio flow in terms of the greatest mutual information criterion and entropy principle [7].

Lin et al. proposed emotion recognition for audio and visual flows are used an error weighted semicoupled HMM. Tripled Hidden Markov Models (THMM) is accepted to audio-visual emotional recognition [8].

2.7 Model Training

The outputs of the fusion are given as an input to the model that is trained according to these data and provides a classification of the video by emotional states. The techniques used in the articles are described are that below.

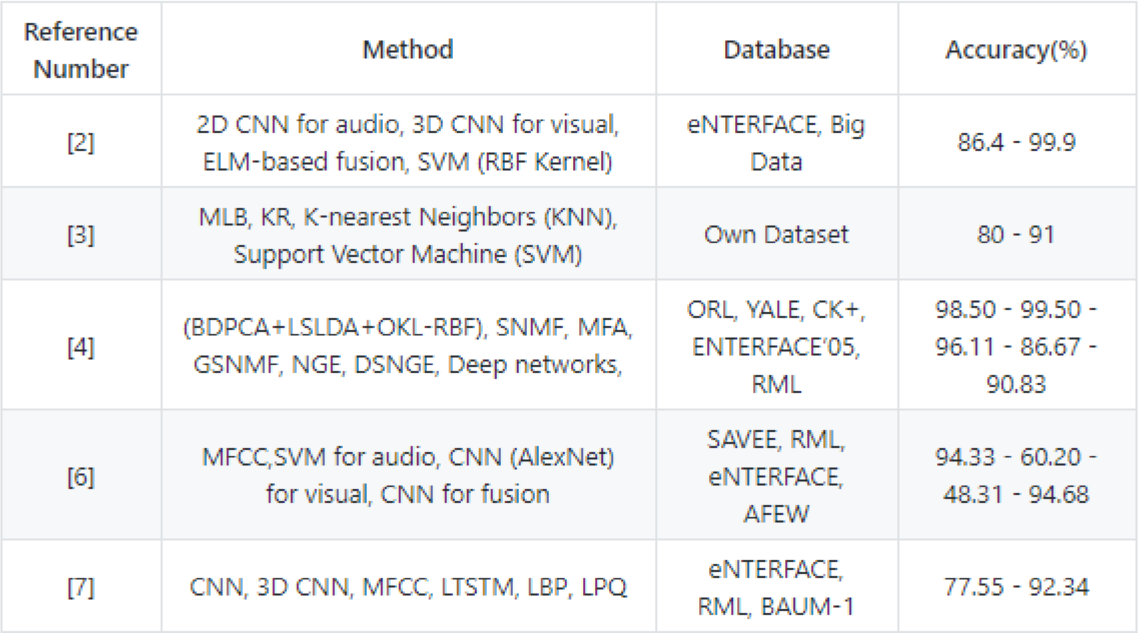
The outputs of the ELM is the input to the Support Vector Machine (SVM) for M. S. Hossain and G. Muhammad's proposed system. They tried two kernels for Support Vector Machine (SVM) that are Radial Basis Function (RBF) and Polynomial. Radial Basis Function (RBF) has accomplished good in the testing [2].

Carl Busso and Zhigang et al. used SVC (Support Vector Machine) with polynomial kernel functions (2nd order) for hyper parameters. Result of this SVC (Support Vector Machine) for emotion recognition, seen that SVC (Support Vector Machine) better than statistical classifiers. Then used the LOOCV (Leave-One-Out Cross-Validation) for the resampling method [3].

2.8 Experiment Result

In this paper, 5 articles were examined. [2][3][4][6][7] Algorithms, preprocessing methods, feature extraction and fusion part, used in audio-visual emotion recognition system are explained in the above sections. In this section, the accuracy rates obtained from those examined articles are shown.

Table II – Summary of Literature Review on Emotional Recognition from Audio-Visual Modality



Reference Number	Method	Database	Accuracy(%)
[2]	2D CNN for audio, 3D CNN for visual, ELM-based fusion, SVM (RBF Kernel)	eNTERFACE, Big Data	86.4 - 99.9
[3]	MLB, KR, K-nearest Neighbors (KNN), Support Vector Machine (SVM)	Own Dataset	80 - 91
[4]	(BDPCA+LSLDA+OKL-RBF), SNMF, MFA, GSNMF, NGE, DSNGE, Deep networks,	ORL, YALE, CK+, ENTERFACE'05, RML	98.50 - 99.50 - 96.11 - 86.67 - 90.83
[6]	MFCC,SVM for audio, CNN (AlexNet) for visual, CNN for fusion	SAVEE, RML, eNTERFACE, AFEW	94.33 - 60.20 - 48.31 - 94.68
[7]	CNN, 3D CNN, MFCC, LTSTM, LBP, LPQ	eNTERFACE, RML, BAUM-1	77.55 - 92.34

2.9 Conclusion

In this literature review, more than 8 articles have been reviewed and summarized in the sections above. In recent years, investigators have recommended a diversity of methods for audio-visual emotional recognition. While distinguishing human emotions leavings a challenging task, with the developments in the field of deep learning, error rates have decreased notably.

3 Summary

3.1 Summary of Conceptual Solution

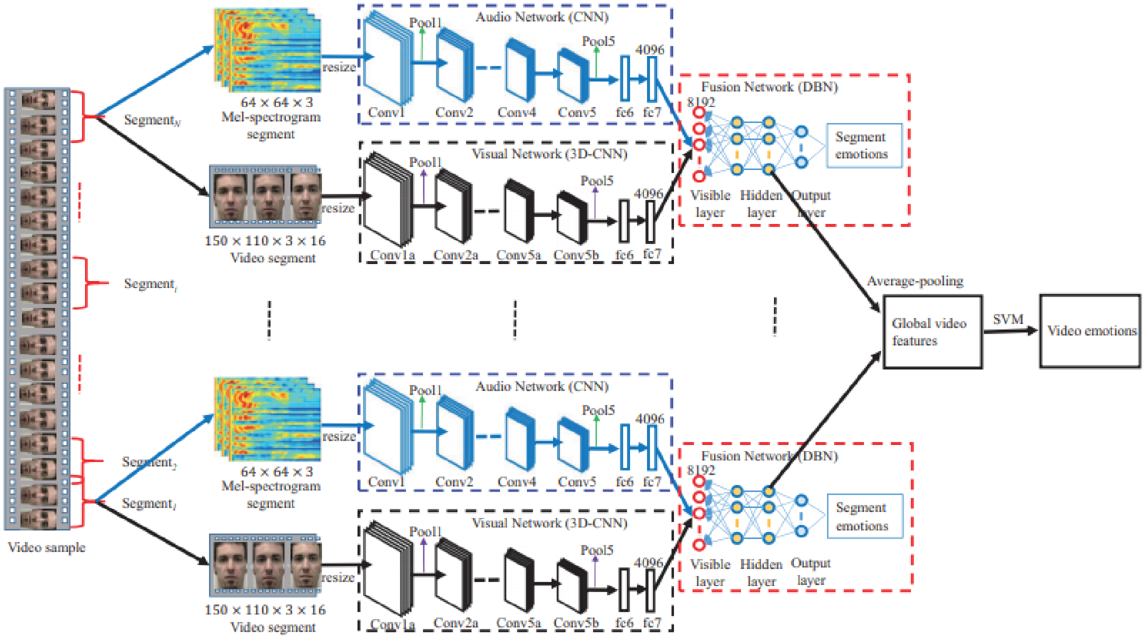


Figure I – The Structure of Deep Model for Audio-Visual Emotional Recognition

3.2 Technology Used

This software will communicate with Python to run image, audio and audio-visual processing functions, and the software will be developed with the Python language. The software will also update itself periodically to ensure high accuracy for optimal performance. The target platform will be Microsoft Windows, Linux and macOS and JetBrains PyCharm will be the development environment.

4 Software Requirements Specification

4.1 Introduction

4.1.1 Purpose

Aim of this document is explaining the system which is called audio-visual emotional recognition. This system goal to provide recognition of emotions from a speech, video and audio-visual.

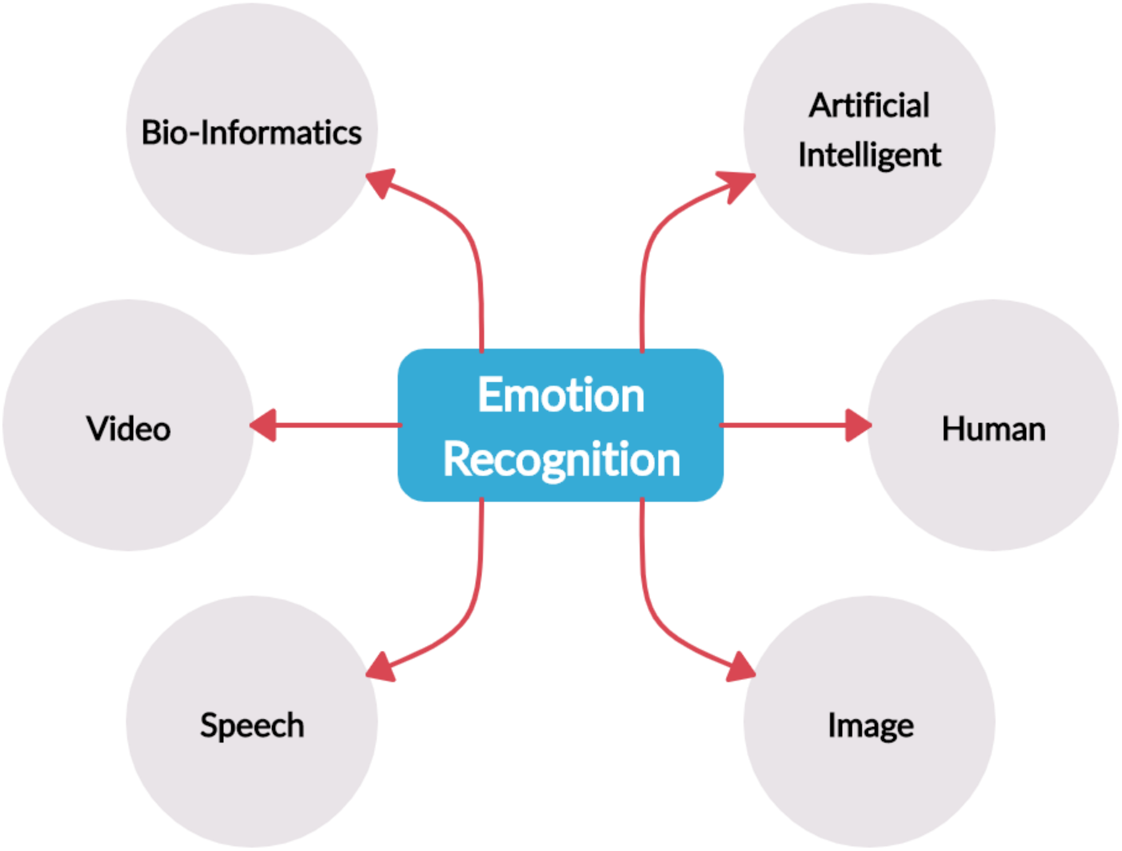


Figure II – Project Overview

The requirements of the project are detailed below. Also identifies the function and non-functional requirements with a use case diagram. All in all, this document is used for how users or admin interact with the system and understand how the mechanism works without any problems.

4.1.2 Scope of Project

Modern day security systems rely heavily on bio-informatics, like as speech, fingerprint, facial images and so on. Besides, determination of a user's emotional state with facial and voice analysis plays a fundamental part in human-machine interaction (HMI) systems, since it employs non-verbal cues to estimate the user's emotional state.

This software system will be performed emotion recognition from audio, video and audio-visual video. With the easy-to-use user-interface of the system, the user can either record instant video/real time or upload an existing video to the system and perform emotion recognition. This system allows big corporate companies to measure customer satisfaction and perform the necessary analysis. There must be an admin in the background that manages the system. The admin has job descriptions that are separate from the user. Section 4.3.2 is described in detail.

There are two actors in the system which are user and admin. First actor is user that can upload or record content also can add information of contents and finally, system gives emotional result of this content. On the other actor, is admin that has responsible the system and what performs maintenance the system and view all contents. Detailed information is described in 4.3.2 Functional Requirements.

4.1.3 Glossary

Table III - Glossary of SRS

Term	Definition
SRS	Software Requirements Specifications
Admin	Person who manage the system
User	Person who wants to know the situation of emotion
Mp4	A file format created by the Moving Picture Experts Group (MPEG) as a multimedia container format designed to store audio-visual data [9].
Waw	A file format for speech
Jpeg	Joint Photographic Experts Group. It’s a standard image format for having compressed image data [10].
Png	Portable Graphics Format. It is store uncompressed image format [11].
Usb	Universal Serial Bus
Content	Speech-Video

4.1.4 Overview of Document

The continiue to this document is ordered as follows: chapter 2 describe the definiton of the properties of project for users who use the system and read the document. The constraints and risks of this system are explained detaily. Chapter 3 is mostly written for project's developers and describes in technical terms the details of the requirements of this system. Also, functions used by the user to use the project software and the properties of these functions are explained detaily.

4.2 Overall Description

4.2.1 Product Perspective

An emotion recognition system can detect the emotion condition of a person either from his image or speech information. In this scope, an audio-visual emotion recognition system requires to evaluate the emotion of a person from his speech and image information together.

The software described in this SRS will be used to detect people's emotions. This project can be used in several areas that like to measure customer satisfaction in a marketing platform, help advertisers to sell products more effectively.

4.2.1.1 Development Methodology

While developing the project, we have decided to use Scrum. Scrum is an agile software development methodology. It is one of the project management methodologies and it is used to manage complex software processes. In performing this management, it split the whole and follows a method based on repetition. It provides that the target is achieved through regular feedback and planning. It has a structure that is flexible for needs and open to innovations. Communication and teamwork are very important [12].

The one most advantage of scrum is that reviewing each sprint before moving to another that testing is conducted throughout the process, so permits teams to change the scope of the project at whatever point.

4.2.2 User Characteristics

Video and audio files must be in a specific format. The video format must be mp4, the audio format must be waw. Video and audio files must have a maximum duration of 10 seconds.

4.2.3 Constraints

Video and audio files must be in a specific format. The video format must be mp4, the audio format must be waw. Video and audio files must have a maximum duration of 10 seconds.

4.2.4 Risks

For the software to run stable, the inputs must provide certain conditions. These conditions are listed below:

- Video quality,
- No shadow in video files,
- No background noises in speech and video file,
- Face should be visible,
- No hoarse voice.

4.3 Requirements Specification

4.3.1 External Interface Requirements

4.3.2 System Interfaces

This part explained in 4.3.2 Functional Requirements.

4.3.2.1 System Interfaces

Our software will be able to work actively on all platforms with python 3.6 installed. What the user can do in the interface is listed below:

- Can externally add files,
- Should be, contact information can be specified,
- Should be comment on the emotion of the video.

Unlike the user, the administrator will be able to make the features listed below.

- Test and train the system,
- Can comment on files uploaded by the user,
- Will be able to access and edit the information uploaded by the user,
- Can data statistics in uploaded files (female, male, age range, country, natio).

4.3.2.2 Hardware Interfaces

The computer to be used must have 1 USB port for video recordings. Besides, it must have 1 microphone input for voice recordings.

4.3.2.3 Software interfaces

The computer to be used must have the libraries attached to python. Some of these libraries are Librosa, OpenCV, Keras, Sklearn, etc.

4.3.2.4 Communications interfaces

There is an internet connection is required to run this software.

4.3.3 Functional Requirements

4.3.3.1 Profile Management Use Case

Use Case:

- Login
- Sign Up
- Validation
- Exit

Diagram:

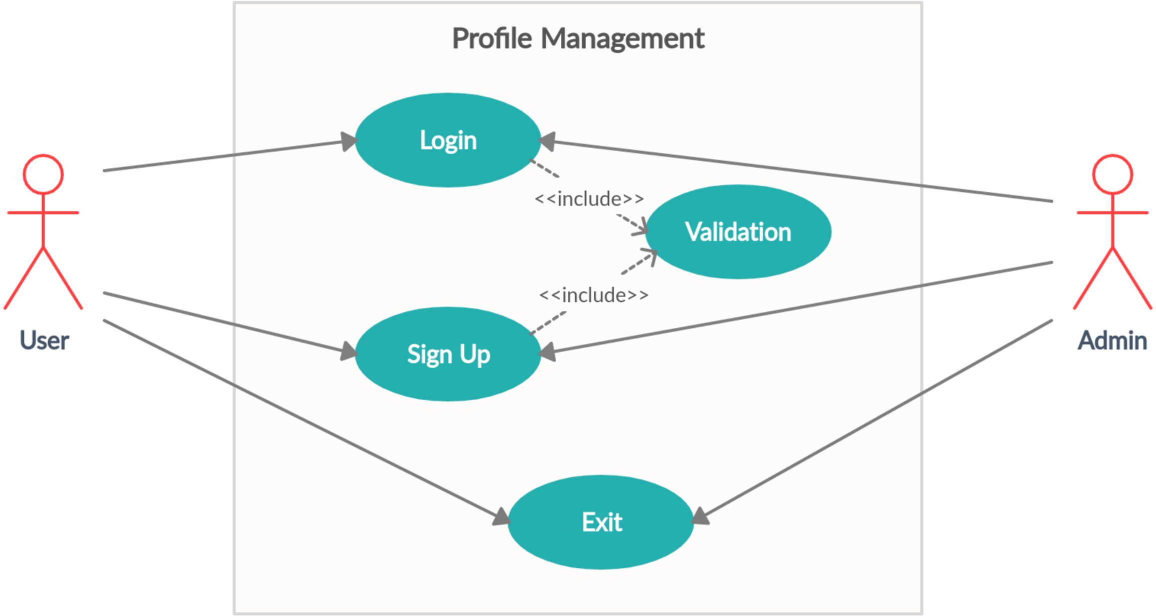


Figure III – Profile Management Use Case

Brief Description:

Figure II shows profile management use case diagram. When user and admin first entered within the system, they come across the authentication menu. Admin and user can use the functions that are Sign Up, Login and Exit.

Initial Step by Step Description:

- Users and admin must login the system.

  i. If the username and password is invalid that should re-login.
- Users and admin can exit from the system.

4.3.3.2 User Use Case

Use Case:

- Upload Content

- Get Result
- Content Information
- Add
- Show

Diagram:

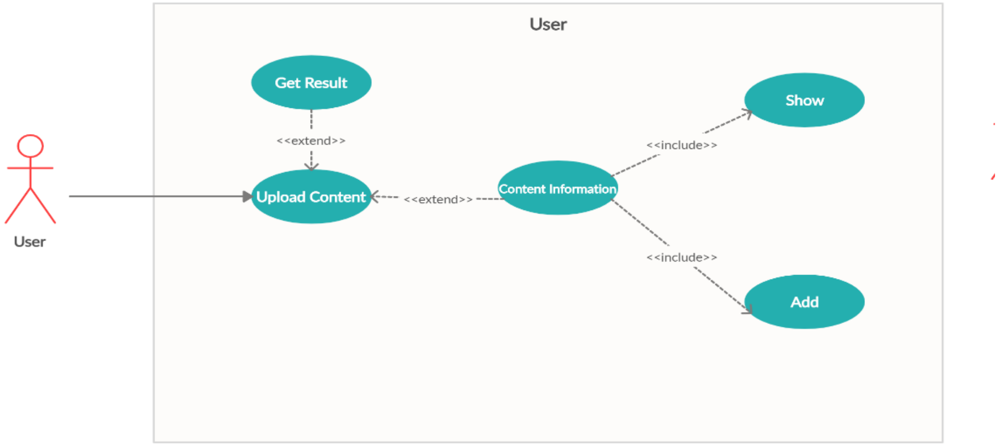


Figure IV – User Use Case

Brief Description:

In user diagram (Figure III) defines what type of action the user can perform on the system. User is able to use the following function: Upload Content, Get Result, also Add and Show in Content Informations and see all uploaded content in the system.

Initial Step by Step Description:

1. User selects Upload Content, the system will wait for you to upload files from your computer. Also, it accepts some format (mp4, waw).
2. After user have uploaded content, should enter, add and view, content's information.
3. Get Result; after user have uploaded or recorded content, the system will give you the result as an emotion.

4.3.3.3 Admin Use Case

Use Case:

- List Content

- System Train
- System Test
- Analyze
- Show Content

- Content Information
- Add
- Edit
- Delete

Diagram:

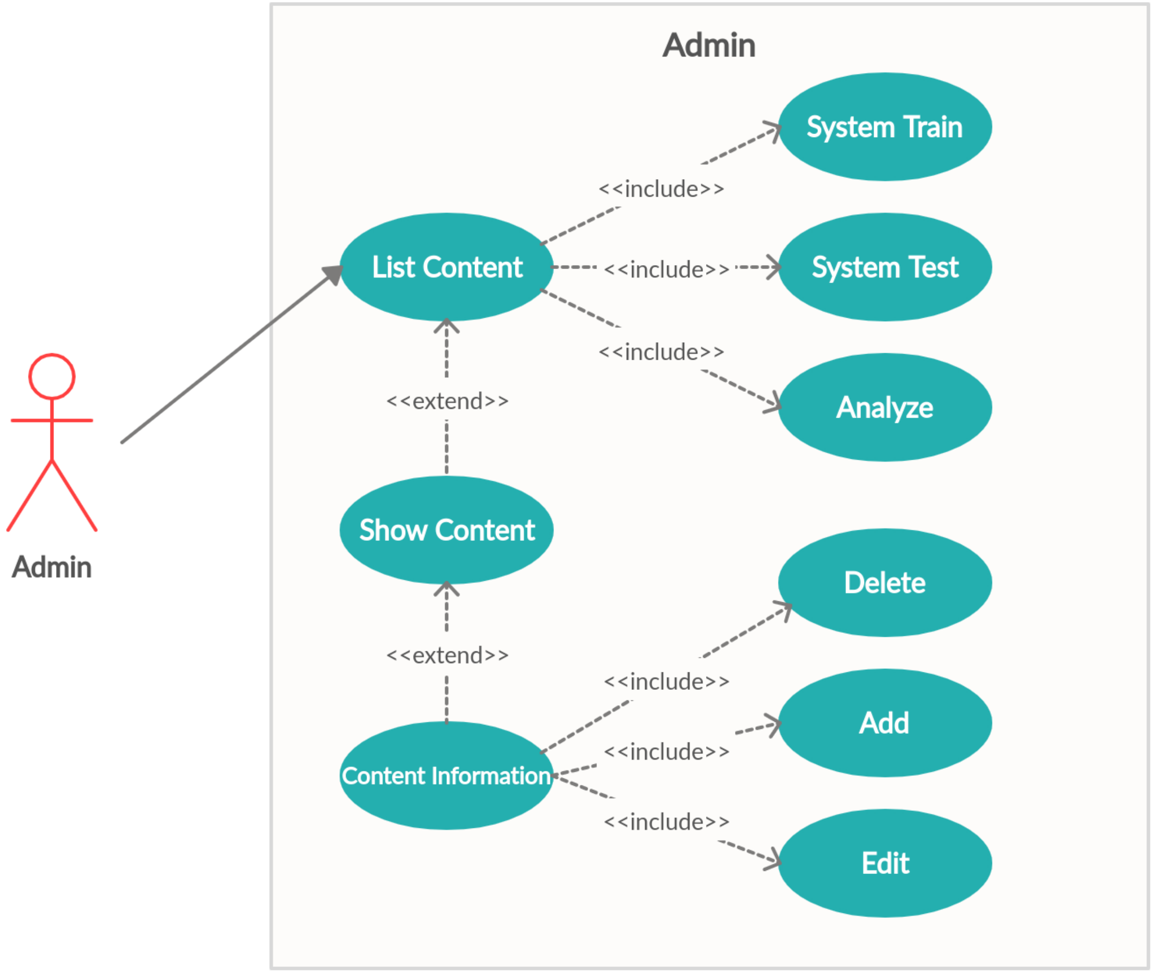


Figure V – Admin Use Case

Brief Description:

The admin is authorized to intervene in the system. Figure IV is admin use case diagram that explains admin's privileges.

Initial Step by Step Description:

1. Admin can see all uploaded content in the system.

2. Admin can train or test the system using these contents.
3. Admin can analyze data statistics in content. (E.g. Female/Male ratio, age range)
4. Admin can view content and edited, deleted and added content's information.

4.3.4 Performance Requirements

The minimum system requirements for the computer to be used are as follows:

- Processors: Intel® Core™ i3 processor or Amd Phenom X4
- Disk space: 1 GB
- Operating systems: Linux, macOS, and Windows 7 or later
- Python versions: 3.6.X or higher
- Included development tools: Anaconda
- Compatible tools: Microsoft Visual Studio, PyCharm, Spyder or VSCode

4.3.5 Software system attributes

4.3.5.1 Reliability

System reliability will improve as long as the video's sound quality is good and the person's face is clearly visible. Since the size and type of the file to be uploaded is limited, no system crashes will be allowed.

4.3.5.2 Availability

The system will work on all operating systems.

4.3.5.3 Security

In order to improve the software, we will be stored input data to the system and will use these data to develop this system. This data will be used to increase stability. Therefore, before receiving the data from the user, a pre-acceptance text will be indicated that the data will only be used for system improvement.

4.3.5.4 Maintainability

In order to increase the stability of the software, the training and test files of the software will be updated once a month by the administrator.

4.3.5.5 Ease of Us

Since the developed application is a user-oriented project, it should provide simple usage to the user. Therefore, the interface we will prepare will be understandable and user-oriented.

5 Software Design Description

5.1 Introduction

5.1.1 Purpose

The purpose of this Software Design Document (SDD) is explaining the system which is called audio-visual emotional recognition. This system goal to provide recognition of emotions from a speech, and video.

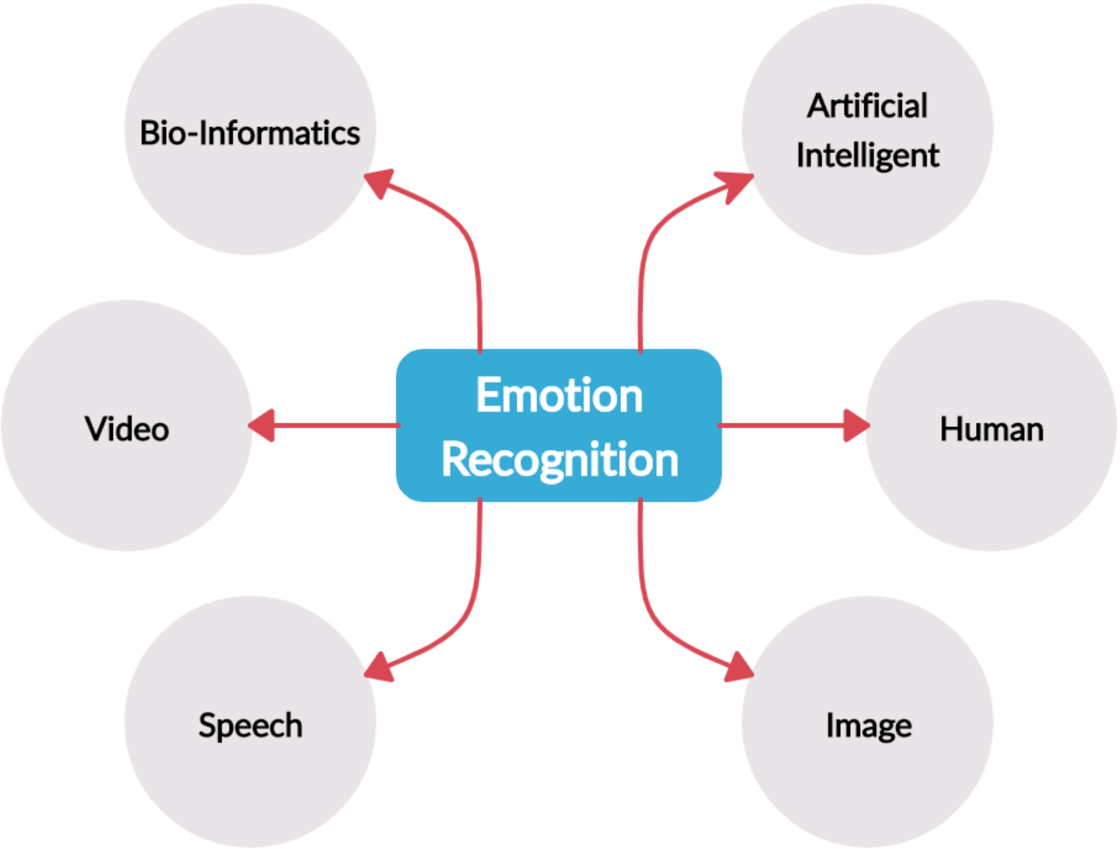


Figure VI – Project Overview

The aim of our project is to measure customer satisfaction by identifying feelings in any corporate or non-corporate company. It also provides some specific analysis results and leads to ideas that can move the company forward.

Besides, this document is written at a level that can be understood when read by an engineer who has been involved in any project does not have to be a computer engineer.

5.1.2 Scope

Modern day security systems rely heavily on bio-informatics, like as speech, fingerprint, facial images and so on. Besides, determination of a user's emotional state with facial and voice analysis plays a fundamental part in human-machine interaction (HMI) systems, since it employs nonverbal cues to estimate the user's emotional state.

This software system will be performed emotion recognition from audio, video and audio-visual video. With the easy-to-use user-interface of the system, the user can either record instant video/real time or upload an existing video to the system and perform emotion recognition. This system allows big corporate companies to measure customer satisfaction and perform the necessary analysis. There must be an admin in the background that manages the system. The admin has job descriptions that are separate from the user. Section 5.3.2 is described in detail.

There are two actors in the system which are user and admin. First actor is user that can upload or record content also can add information of contents and finally, system gives emotional result of this content. On the other actor, is admin that has responsible the system and what performs maintenance the system and view all contents. Detailed information is described in 5.3.2 Functional Requirements.

5.1.3 Glossary

Table IV - Glossary of SDD

Term	Definition
SDD	Software Design Document
Admin	Person who manage the system
User	Person who wants to know the situation of emotion
Content	Speech-Video

5.1.4 Overview of document

The rest of this document is organized as follows: Chapter 2 is written to provide an overview of the design of the project and to guide engineers on how to implement the system. Chapter 3 describes the realization of the use case. Finally, chapter 4 describes how to perform the emotion detection process and the fusion algorithm.

5.1.5 Motivation

We are engineering senior students who are excited, love research, and enjoy learning and producing new things. Our team is composed of two Computer Engineering students and two Electronic and Communication Engineering students. we think that working in a multidisciplinary project has a lot to teach. We are interested in speech processing, image processing and artificial intelligence fields. We aimed to develop ourselves more by choosing our project to cover these issues. Actually, learning new things. That is all our motivation.

5.2 Design Overview

5.2.1 Description of a Problem

Our problem in this project is to emotions and their definitions. We want to define emotions both speech and image separately. Then combine it with the fusion algorithm and improve the before developed methods.

5.2.2 Technologies Used

This software will communicate with Python to run image, audio and audio-visual processing functions, and the software will be developed with the Python language. The software will also update itself periodically to ensure high accuracy for optimal performance. The target platform will be Microsoft Windows, Linux and macOS and JetBrains PyCharm will be the development environment.

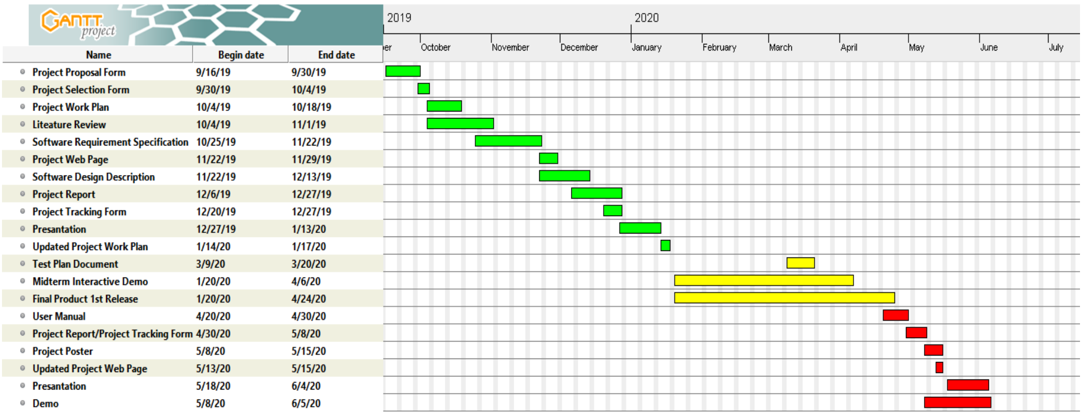
5.3 Architecture Design

5.3.1 Design Approach

While developing the project, we have decided to use Scrum which is an agile software development methodology. Scrum; is one of the project management methodologies and it is used to manage complex software processes. In performing this management, it split the whole and follows a method based on repetition. It provides that the target is achieved through regular feedback and planning. It has a structure that is flexible for needs and open to innovations. Communication and teamwork are very important [12].

The one most advantage of scrum is that reviewing each sprint before moving to another that testing is conducted throughout the process, so permits teams to change the scope of the project at whatever point and libraries.

In Gantt chart shown in Figure VII, represents working phases and durations of our senior project. By using Gannt chart we divided our tasks into small pieces, and we visualize flow of our project.



Name	Begin date	End date
Project Proposal Form	9/16/19	9/30/19
Project Selection Form	9/30/19	10/4/19
Project Work Plan	10/4/19	10/18/19
Liteature Review	10/4/19	11/1/19
Software Requirement Specification	10/25/19	11/22/19
Project Web Page	11/22/19	11/29/19
Software Design Description	11/22/19	12/13/19
Project Report	12/6/19	12/27/19
Project Tracking Form	12/20/19	12/27/19
Presantation	12/27/19	1/13/20
Updated Project Work Plan	1/14/20	1/17/20
Test Plan Document	3/9/20	3/20/20
Midterm Interactive Demo	1/20/20	4/6/20
Final Product 1st Release	1/20/20	4/24/20
User Manual	4/20/20	4/30/20
Project Report/Project Tracking Form	4/30/20	5/8/20
Project Poster	5/8/20	5/15/20
Updated Project Web Page	5/13/20	5/15/20
Presantation	5/18/20	6/4/20
Demo	5/8/20	6/5/20

Figure VII – Project Work Plan

5.3.1.1 Class Diagram

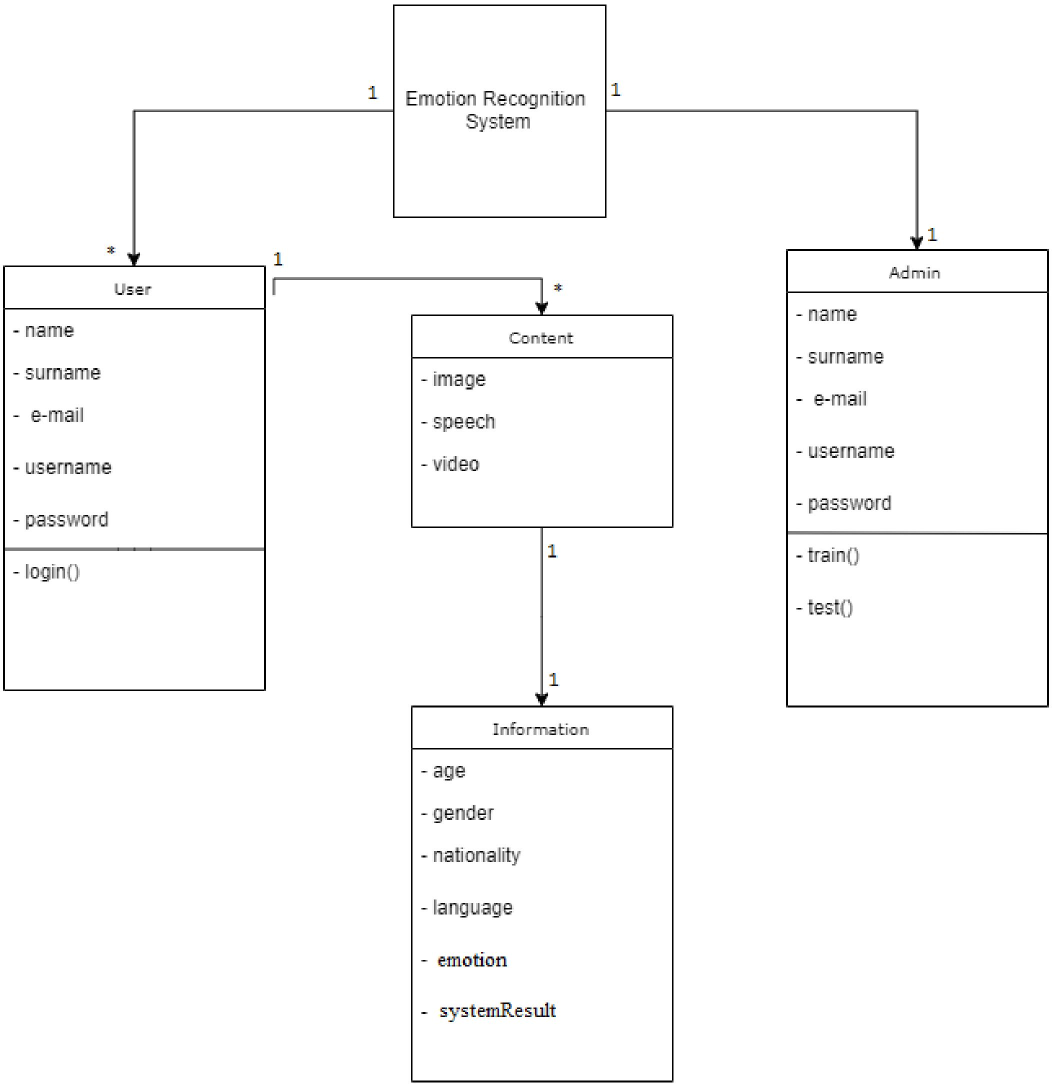


Figure VIII – Class Diagram

5.3.1.2 Activity Diagram

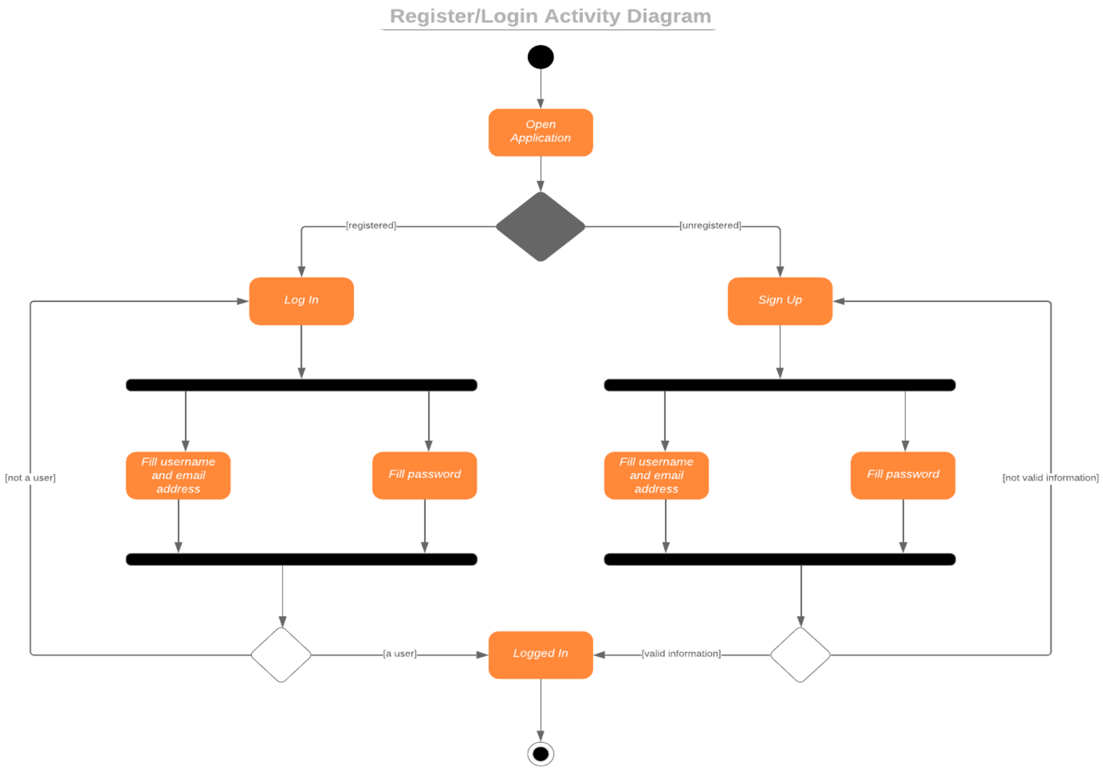


Figure IX – Register and Login Activity Diagram

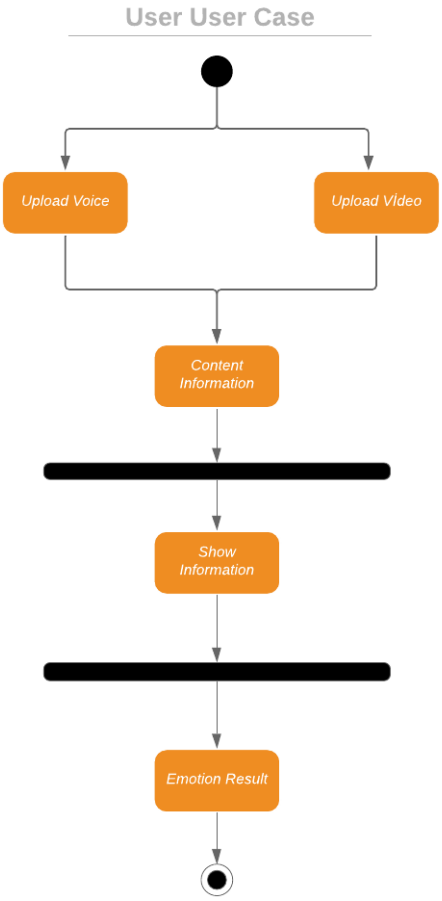


Figure X – User Activity Diagram

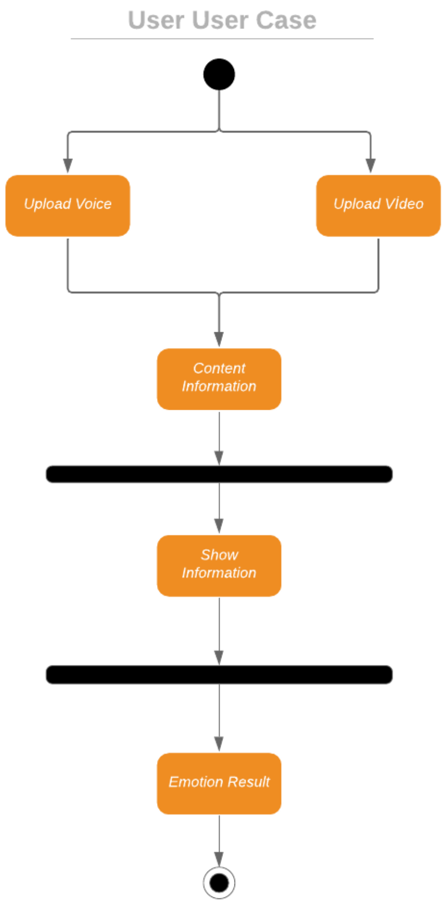


Figure XI – Admin Activity Diagram

5.3.1.3 Database Diagram

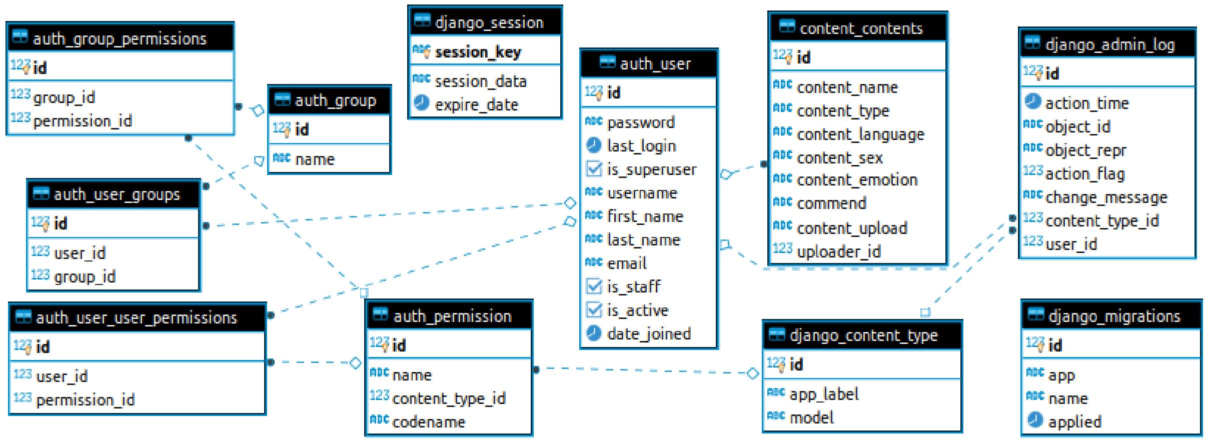


Figure XII – Database Diagram

5.3.2 Architecture Design of an Audio-Visual Emotional Recognition

5.3.2.1 Profile Management

Summary: When user and admin first entered within the system, they come across the authentication menu. Admin and user can use the functions that are Sign Up, Login and Exit.

Initial Step by Step Description:

1. Users and admin must login the system.
2. If the username and password are invalid that should re-login. Admin and user can exit from the system.

5.3.2.2 Admin Menu

Summary: The admin is authorized to intervene in the system.

Initial Step by Step Description:

1. Admin can see all uploaded content in the system.
2. Admin can train or test the system using these contents.
3. Admin can analyze data statistics in content. (E.g. Female/Male ratio, age range)
4. Admin can view content and edited, deleted and added content's information.

5.3.2.3 User Menu

Summary: User is able to use the following function: Upload Content, Get Result, also Show, Edit, Add and Delete in Content Informations.

Initial Step by Step Description:

1. If user selects asks the user what kind of video/sound file to save.
    i. If user selects video, the system will be activated camera and microphone.
    ii. If user selects sound, the system will be activated just microphone. After file type is selected, the system will start recording.
2. If user selects Upload Content, the system will wait for you to upload files from your computer. Also, it accepts some format (mp4, waw).
3. After user has uploaded or recorded content, optionally can enter, view, delete or edit content's information.
4. If user selects Get Result; after user has uploaded content, the system will give you the result as an emotion.

5.4 Use Case Realization

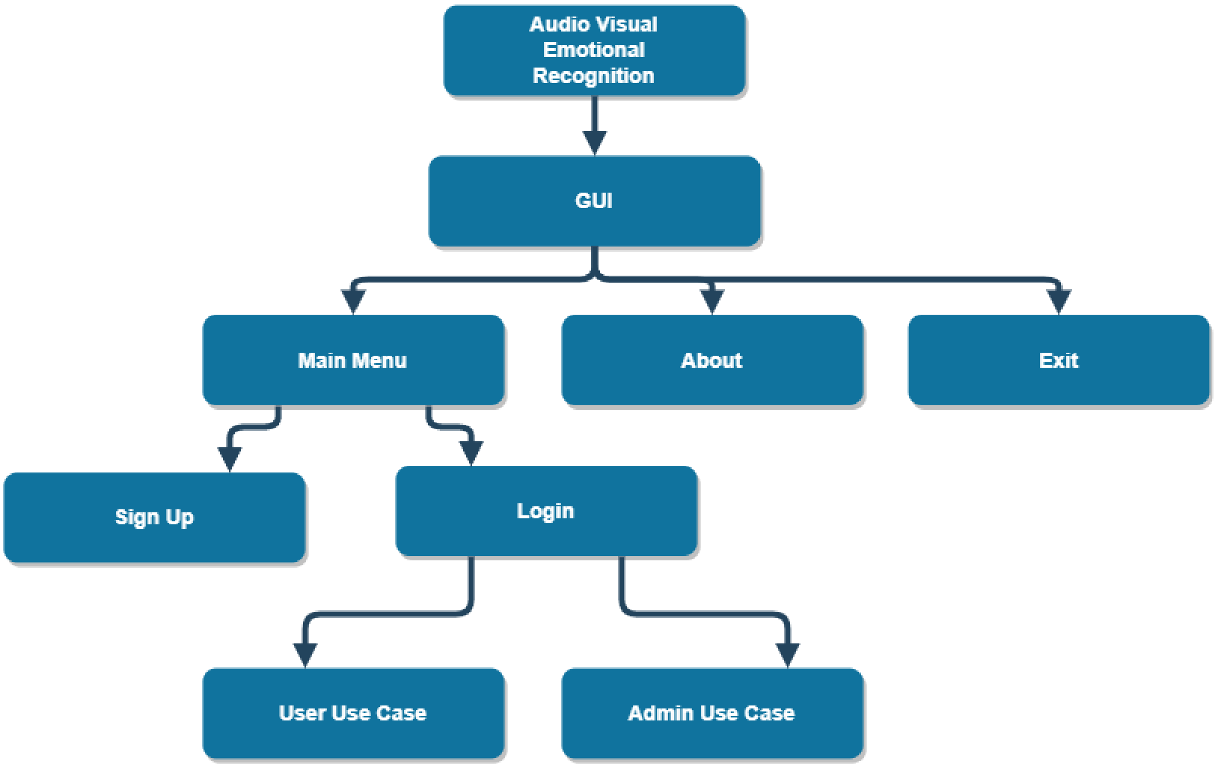


Figure XIII – Project Components

5.4.1 External Interface Requirements

All systems of the project are shown in the block diagram in the Figure XIII.

5.4.1.1 Project Components

5.4.1.1.1 GUI Design

GUI is designed to allow users and administrators to easily use the audio-visual emotional recognition software developed. GUI consists of three main headings. These headings are Main Menu, About, Exit. There are 2 sub-titles in the Main Menu. These sub-titles, Sign Up and Login. The user or administrator can easily register and use the system. The user and administrator have different uses. The user can upload audio and video and learn the emotion from these files. The administrator has access to all attached files as well as all of these. About title contains information about the software and GUI. The exit title is used to exit the GUI.

5.5 Detection

In this project, audio and video acquisition technique was used to create Audio-Visual emotion recognition. First, as shown in Figure XIV, the audio and video data from the video are preprocessed separately. Then the data from both the image and the speech goes through feature extraction. The features obtained as a result of this process are given to the fusion algorithm. The data from the Fusion algorithm shows us the emotion after the model training, we selected.

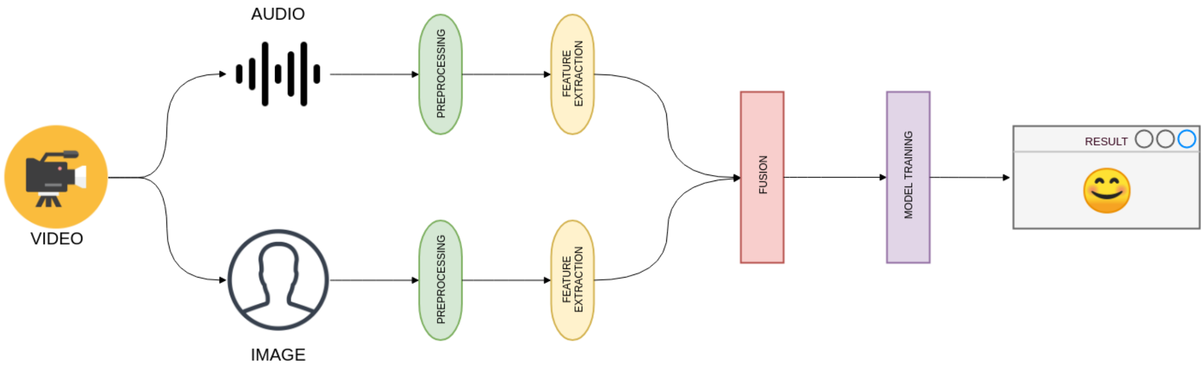


Figure XIV – System Overview

6 Test Plan

6.1 Introduction

6.1.1 Version Control

Version No	Description of Changes	Date
1.0	First Version	February 3, 2020
1.1	Second Version	March 10, 2020

6.1.2 Overview

We have two modes in our system which is admin mode and user mode. These modes and all interfaces that before mentioned in SRS and SDD Document will be tested.

6.1.3 Scope

This document provides a brief explanation about what will be our test cases and when we are testing and how testing our system.

6.1.4 Terminology

Acronym	Definition
GUI	Graphical User Interface
AM	Admin Mode

UM	User Mode

6.2 Features to be tested

This section lists and gives a brief description of all the major features to be tested. For each major feature there will be a Test Design Specification added at the end of this document.

6.2.1 Graphical User Interface

Basically contains, all buttons, loading and viewing content place that features will make it easier for people to use the application.

6.2.2 Admin Mode

Basically contains, test and train buttons, list of contents and their views and can editable its information places.

6.2.3 User Mode

Basically contains, loading content and its information add buttons. Also contains emotion result a panel.

6.3 Item Pass/Fail Criteria

6.3.1 Exit Criteria

- 100% of the test cases are executed.
- 95% of the test cases passed.
- All High and Medium Priority test cases passed.

6.4 References

[1] CENG408_Group17_SRS_V1.0, December 12, 2020

[2] CENG408_Group17_SDD_V1.0, December 13, 2020

6.5 Test Design Specification

6.5.1 Graphical User Interface

6.5.2 Sub Features to be tested

Sign Up Button (GUI.SG): All users (include admin) have to sign up to the system. After selecting “Sign Up” button, sign up interface will be shown.

Login Button (GUI.LG): All users (include admin) have to login to the system Selecting "Login" button, login interface will be shown.

6.5.3 Test Cases

TC ID	Priority	Scenario Description
GUI.SG	High	Select Sign Up button. After selection registered in system.
GUI.LG	High	Select log in button. After selection log in to the system.

6.5.4 Admin Mode

Upload Content Button (AD.CNT.BT):

After uploading the content to the content loading area in the main interface, all users (include admin) must click the "Upload" button.

Get Emotion Visual Button (AD.EV.BT):

Admin who wants to learn the emotional state of the uploaded content (only video, not audio) has to click on the "Get Result Visual" button.

Get Emotion Fusion Button (AD.EF.BT):

Admin who wants to learn the emotional state of the uploaded content (video with audio) has to click on the "Get Result Fusion" button.

Get Emotion Audio Button (AD.EA.BT):

Admin who wants to learn the emotional state of the uploaded content (only audio) has to click on the "Get Result Audio" button.

List All Content Button (AD.LST.BT):

The admin who wants to see the content he has uploaded should click the "List All Contents" button. Admin can see all of their contents after listing all the contents.

Show Content Button (AD.SHW.BT):

Admin can see the information of the content.

Add Content Information Button (AD.ADD.BT):

Admin can add the information of the content.

Edit Content Information Button (AD.EDT.BT):

Admin can edit the information of the content.

Delete Content Information Button (AD.DEL.BT):

Admin can delete the information of the content.

List All User Button (AD.LSTUS.BT):

The admin who wants to see the all users in the system, should click the "List All Users" button. Admin can see all of their information after listing all the users.

Show Users Information Button (AD.SHWUS.BT):

Admin can see the information of the user.

Add User Information Button (AD.ADDUS.BT):

Admin can add the information of the user.

Edit User Information Button (AD.EDTUS.BT):

Admin can edit the information of the user.

Delete User Information Button (AD.DELUS.BT):

Admin can delete the information of the user.

System Train Button (AD.SYSTR.BT):

Admin can train the system at any time.

System Analyze Button (AD.ANLZ.BT):

Admin can analyze the system at any time.

6.5.5 Test Cases

TC ID	Requirements	Priority	Scenario Description
AD.CNT.BT	6.1-6.2	High	Select Upload Content button. After selection uploaded content in system.
AD.EV.BT	6.1-6.2	High	Select Get Emotion Visual button. After selection, system give the emotion state according to visual algortihm.
AD.EF.BT	6.1-6.2	High	Select Get Emotion Fusion button. After selection, system give the emotion state according to fusion algorithm.
AD.EA.BT	6.1-6.2	High	Select Get Emotion Audio button. After, system give the emotion state according to audio algorithm.
AD.LST.BT	6.1-6.2	High	Select List all content button. After selection, all contents list in screen.
AD.SHW.BT	6.1-6.2	High	Select Show content button. After selection, content' content shows in screen.
AD.ADD.BT	6.1-6.2	High	Select Add Content Information button. After selection, content information add in the system.
AD.EDT.BT	6.1-6.2	High	Select Edit Content Information button. After selection, content information edit in the system.
AD.DEL.BT	6.1-6.2	High	Select Delete Content Information button. After selection, content information delete in the system.

AD.LSTUS.BT	6.1-6.2	High	Select List all user button. After selection, all users list in screen.
AD.SHWUS.BT	6.1-6.2	High	Select Show users Information button.After selection, user informations show in the system.
AD.ADDUS.BT	6.1-6.2	High	Select Add User Information button. After selection, user information add in the system.
AD.EDTUS.BT	6.1-6.2	High	Select Edit User Information button. After selection, user information edit in the system.
AD.DELUS.BT	6.1-6.2	High	Select Delete User Information button. After selection, user information delete in the system.
AD.SYSTR.BT	6.1-6.2	High	Select System Train button. After selection, System train start with new data.
AD.ANLZ.BT	6.1-6.2	High	Select System Analysis button. After selection, system analsis shows in the screen from using users information.

6.6 User Mode

6.6.1 Sub features to be tested

- **Upload Content Button (U.CNT.BT):**

After uploading the content to the content loading area in the main interface, all users (include admin) must click the "Upload" button.

- **Get Emotion Visual Button (U.EV.BT):**

The user who wants to learn the emotional state of the uploaded content (only video, not audio) has to click on the "Get Result Visual" button.

- **Get Emotion Fusion Button (U.EF.BT):**

The user who wants to learn the emotional state of the uploaded content (video with audio) has to click on the "Get Result Fusion" button.

- **Get Emotion Audio Button (U.EA.BT):**

The user who wants to learn the emotional state of the uploaded content (only audio) has to click on the "Get Result Audio" button.

- **List All Content Button (U.LST.BT):**

The user who wants to see the content he has uploaded should click the "List All Contents" button. Admin can see all of their contents after listing all the contents.

- **Show Content Button (U.SHW.BT):**

The user can see the information of the content.

- **Add Content Information Button (U.ADD.BT):**

The user can add the information of the content.

6.6.2 Sub features to be tested

Here list all the related test cases for these features:

TC ID	Requirements	Priority	Scenario Description
U.CNT.BT	6.1-6.2	High	Select Upload Content button. After selection uploaded content in system.
U.EV.BT	6.1-6.2	High	Select Get Emotion Visual button. After selection, system give the emotion state according to visual algortihm.
U.EF.BT	6.1-6.2	High	Select Get Emotion Fusion button. After selection, system give the emotion state according to fusion algorithm.
U.EA.BT	6.1-6.2	High	Select Get Emotion Audio button. After, system give the emotion state according to audio algorithm.
U.LST.BT	6.1-6.2	High	Select List all content button. After selection, all contents list in screen.
U.SHW.BT	6.1-6.2	High	Select Show content button. After selection, content' content shows in screen.
U.ADD.BT	6.1-6.2	High	Select Add Content Information button. After selection, content information add in the system.

7 Detailed Test Cases

7.1 GUI.SG

TC ID	GUI.SG
Purpose	Register to the system.
Requirements	-

Priority	H
Estimated Time Needed	2 min
Dependency	Run the application
Setup	Sign in to the system.
Cleanup	Close page

7.2 GUI.LG

TC ID	GUI.LG
Purpose	Sign in to the system
Requirements	6.1
Priority	H
Estimated Time Needed	2 min
Dependency	User or admin should registered in the system.
Setup	Sign in to system.
Cleanup	Close page

7.3 AD.CNT.BT

TC ID	AD.CNT.BT
Purpose	Adding new content
Requirements	6.1-6.2
Priority	H
Estimated Time Needed	2 min
Dependency	Admin should sign in the system.
Setup	Uploaded content
Cleanup	Close page

7.4 AD.EV.BT

TC ID	AD.EV.BT
Purpose	Estimate to emotion for visual data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	Admin should sign in the system.
Setup	Estimated emotion for visual data
Cleanup	Close Page

7.5 AD.EF.BT

TC ID	AD.EF.BT
Purpose	Estimate to emotion for fusion data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	Admin should sign in the system.
Setup	Estimated emotion for fusion data
Cleanup	Close Page

7.6 AD.EA.BT

TC ID	AD.EA.BT
Purpose	Estimate to emotion for audio data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	Admin should sign in the system.

Setup	Estimated emotion for audio data
Cleanup	Close Page

7.7 AD.LST.BT

TC ID	AD.LST.BT
Purpose	List all contents
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	2 min
Dependency	Admin should sign in the system.
Setup	Listed all contents
Cleanup	Close page

7.8 AD.SHW.BT

TC ID	AD.SHW.BT
Purpose	Showing content' content
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	2 min
Dependency	Admin should sign in the system.
Setup	Show content' content
Cleanup	Close Page

7.9 AD.ADD.BT

TC ID	AD.ADD.BT
Purpose	Adding Content Information
Requirements	6.1-6.2

Priority	L
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.
Setup	Added content information
Cleanup	Close Page

7.10 AD.EDT.BT

TC ID	AD.EDT.BT
Purpose	Editing Content Information
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.
Setup	Edited Content Information
Cleanup	Close Page

7.11 AD.DEL.BT

TC ID	AD.DEL.BT
Purpose	Deleting Content Information
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	
Dependency	Admin should sign in the system.
Setup	Deleted Content Information
Cleanup	Close Page

7.12 AD.LSTUS.BT

TC ID	AD.LSTUS.BT
Purpose	List all users
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	2 min
Dependency	Admin should sign in the system.
Setup	Listed all users
Cleanup	Close Page

7.13 AD.SHWUS.BT

TC ID	AD.SHWUS.BT
Purpose	Show user's information
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.
Setup	Listed all information for selected user.
Cleanup	Close Page

7.14 AD.ADDUS.BT

TC ID	AD.ADDUS.BT
Purpose	Adding user informations
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.

Setup	Adding to defining information for user
Cleanup	Close Page

7.15 AD.EDTUS.BT

TC ID	AD.EDTUS.BT
Purpose	Editing user informations
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.
Setup	Editing to defining information for user
Cleanup	Close Page

7.16 AD.DELUS.BT

TC ID	AD.DELUS.BT
Purpose	Deleting user' information
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	1 min
Dependency	Admin should sign in the system.
Setup	Delete any information for user' information
Cleanup	Close page

7.17 AD.SYSTR.BT

TC ID	AD.SYSTR.BT
Purpose	Training system
Requirements	6.1-6.2

Priority	H
Estimated Time Needed	?
Dependency	New data
Setup	Trained System with new data
Cleanup	Close page

7.18 AD.ANLZ.BT

TC ID	AD.ANLZ.BT
Purpose	Analyzing System
Requirements	6.1-6.2
Priority	H
Estimated Time Needed	5 min
Dependency	-
Setup	Analyzed all user/content information
Cleanup	Close page

7.19 U.CNT.BT

TC ID	U.CNT.BT
Purpose	Uploading Content
Requirements	6.1-6.2
Priority	H
Estimated Time Needed	2 min
Dependency	User should login to the system and uploading properly content type
Setup	Uploaded content
Cleanup	Close page

7.20 U.EV.BT

TC ID	U.EV.BT
Purpose	Estimate to emotion for visual data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	User should sign in the system.
Setup	Estimated emotion for visual data
Cleanup	Close Page

7.21 U.EF.BT

TC ID	U.EF.BT
Purpose	Estimate to emotion for fusion data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	User should sign in the system.
Setup	Estimated emotion for fusion data
Cleanup	Close Page

7.22 U.EA.BT

TC ID	U.EA.BT
Purpose	Estimate to emotion for audio data
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	5 min
Dependency	User should sign in the system.

Setup	Estimated emotion for audio data
Cleanup	Close Page

7.23 U.LST.BT

TC ID	U.LST.BT
Purpose	List all contents
Requirements	6.1-6.2
Priority	L
Estimated Time Needed	2 min
Dependency	User should sign in the system.
Setup	Listed all contents
Cleanup	Close page

7.24 U.SHW.BT

TC ID	U.SHW.BT
Purpose	Showing content' content
Requirements	6.1-6.2
Priority	M
Estimated Time Needed	2 min
Dependency	User should sign in the system.
Setup	Show content' content
Cleanup	Close Page

7.25 U.ADD.BT

TC ID	U.ADD.BT
Purpose	Adding Content Information
Requirements	6.1-6.2

Priority	L
Estimated Time Needed	1 min
Dependency	User should sign in the system.
Setup	Added content information
Cleanup	Close Page

8 Conclusions

At the end of the Ceng407 & Ceng408 course, the development stages of a software project, literature review and its importance, moreover, writing Software Requirements Specification and Software Design documents, were mastered on a real-world project. In this process, Literature Review, Software Requirements Specification, Software Design and Test Plan documents was written. Details of the studies performed in chapters 2, 4, 5 and 6 can be examined. In addition, all future works stated at the end of the ceng407 course project report have been completed. Looking back, the sound and images were processed separately and then combined with the fusion algorithm. With deep learning algorithms, models had trained and a model that can be found in accurate predictions had created. It was predicted that there might be problems in the fusion part, as a matter of fact, but many attempts were overcome. The user interface has been created and has reached a level that will be available for use. In this process, a notice was written excluding the fusion part and sent to the International Conference on Image Processing (ICIP) 2020 conference.

9 Future Works

We hope to be admitted to the conference and discuss the issue there so that the project can be further developed technically. We also plan to write a full paper that includes the fusion section later.

Acknowledgement

We are grateful for guidance we have received from *Assoc. Prof. Dr. Hadi Hakan MARAŞ and Dr. Lecturer Selma ÖZAYDIN*. The help we received from them was a great asset to improve this project and ourselves.

References

[1] Onur Önder, Sara Zhalehpour, and Çiğdem Eroğlu Erdem. A Turkish Audio-Visual Emotional Database. In 2013 21st Signal Processing and Communications Applications Conference (SIU), PAGES 1-2, April 2013.

[2] M. S. Hossain and G. Muhammad. Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data. Information Fusion, VOL. 49, PP. 69-78, September 2019.

[3] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. Proceedings of the 6th International Conference on Multimodal Interfaces, October 2014

[4] Kah Phooi Seng, Li-Minn Ang, Chien Shing Ooi. A Combined Rule-Based & Machine Learning Audio-Visual Emotion Recognition Approach. IEEE Transactions on Affective Computing, VOL. 9, NO. 1, January-March 2018

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

[6] Egils Avots, Tomasz Sapinski, Maie Bachmann, Dorota Kaminska. Audiovisual Emotion Recognition in Wild. Machine Vision and Applications, VOL. 30, ISSUE 5, PAGES 975-985, 19 July 2018

[7] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition. IEEE Transactions on Circuits and Systems for Video Technology, VOL. 28, ISSUE 10, October 2018

[8] Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei. Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition. IEEE Transactions on Multimedia, VOL. 14, ISSUE 1, February 2012

[9] What is Mp4?, 2019. [Online].
Available: https://www.techopedia.com/definition/10713/mp4 [Accessed 21 November 2019]

[10] What is a Jpeg File?, 2019. [Online].
Available: https://www.paintshoppro.com/en/pages/jpeg-file/ [Accessed 21 November 2019]

[11] What is a Png File?, 2019. [Online]. Available: https://www.paintshoppro.com/en/pages/png-file/ [Accessed 21 November 2019]

[12] What is Scrum?, 2019. [Online]. Available: https://www.scrum.org/resources/what-is-scrum [Accessed 21 November 2019]