ÇANKAYA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

Project Report
Version 1

CENG 407
Innovative System Design and Development I

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 &*
Asst. Prof. Selma ÖZAYDIN

Table of Contents

Abstract....2
Özet....2
1. Introduction....3
1.1 Contribution....3
2. Literature Search....4
2.1 Dataset....5
2.2 Preprocessing....5
2.2.1 Audio Preprocessing....5
2.2.2 Visual Preprocessing....6
2.3 Feature Extraction....6
2.3.1 Audio Part....6
2.3.2 Visual Part....7
2.4 Fusion....8
2.5 Model Training....9
2.6 Experiment Result....9
2.7 Conclusion....10
3. Summary....10
3.1 Summary of Conceptual Solution....10
3.2 Technology Used....10
4. Software Requirements Specification....11
4.1 Introduction....11
4.2 Overall Description....13
4.3 Requirements Specification....14
5. Software Design Description....20
5.1 Introduction....20
5.2 Design Overview....22
5.3 Architecture Design....22
5.4 Use Case Realization....28
5.5 Detection....28
6. Conclusions....29
7. Future Works....29
Acknowledgement....29
References....30

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, image 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, image 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.

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 databases 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. The test subjects first watch visual or audio-visual stimulation on a screen in front of them, which are designed to elicit determined emotions and mental states. They answer questions about visual stimulants [1]. The target emotions that generally intended to elicit are the six basic ones that are happiness, anger, sadness, disgust, fear, surprise. 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.

2.2.1 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 the center frequencies of the filters are distributed on a MEL scale. The 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 for this process. The VAD technique uses the short-time energy (STE) and short-time zero-crossing rate (STZCR) features, and these steps follow that the speech signal x(m) in the time domain is divided into n number of frames, and the STZCR is calculated from the weighted average from the number of times the speech signal changes sign within a time window, and then STE and STZCR 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].

2.2.2 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 into 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, the 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 the facial region and remove unwanted background information. Viola-Jones (VJ) algorithm was used for this process [4].

Egils Avots, Tomasz Sapinski et al. preprocessing part, the video is divided into frames for visual-based features, that the purpose in doing so select mainframes from a video. When select mainframes, it is used numerical frame difference that is said as the sum of the absolute difference between pixels. An image pair supplies a score of similarity for frames. The system averages the difference values for the last 10 processed frames. If the new frame has a different value that is less 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].

2.3 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.

2.3.1 Audio Part

M. S. Hossain and G. Muhammad proposed a 2D CNN architecture for speech signals. There are four convolution layers and three pooling layers. The last layer is a fully connected neural network with two hidden layers. A SoftMax function is applied to the output of the fully connected layer. 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), Kernel Regression (KR) and K-nearest neighbors (KNN) for feature extraction. Mel-frequency cepstral

coefficients (MFCC) were used to train the Hidden Markov Model (HMM). Recognized the four emotions, and six archetypal emotion classifications of power coefficients used 12 MEL-based speech signals to train the Markov model [3].

Egils Avots, Tomasz Sapinski et al. when extracting the feature of the audio part, focuses on the non-linguistic properties of the audio signal. Then, they have extracted Mel-frequency cepstral coefficients (MFCCs), which are calculated for a 400-millisecond sliding window with a step size of 200 millisecond. Thus, the property vector has obtained. One feature vector consists of 34 parameters, the first 21 represent the global audio features and the remaining 13 coefficients represent the local MFCC. For MFCC feature extraction, there are used parameters that are pre emphasis coefficient 0.97, 20 filter bank channels, 13 cepstral coefficients, 300 Hertz lower frequency limit and 3700 Hertz upper-frequency limit [6].

Wang and Guan et al. proposed employ pitch, intensity, and the first 13 MFCC features on audio feature extraction tasks. Mel-frequency Cepstral Coefficient (MFCC) is the most well-known spectral features since it is used to model the human auditory perception system. MFCC contains sound quality features that are formulas, spectral energy distribution, harmonics-noise ratio and so on [7].

2.3.2 Visual Part

M. S. Hossain and G. Muhammad proposed a 3D CNN architecture for video signals. There are eight convolution layers and five max-pooling layers. In the end, there are two fully connected layers. A SoftMax layer follows the fully connected layers. The stride of the filters is one. The input to the model is 16 keyframes (RGB) resized to 227 × 227. The 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 forehead, eyebrow, low eye, right cheek and left cheek area. Then defined a local source of coordinates for each frame. Then provided these by reducing spatial data collected from markers in each frame of the video to a 4-dimensional property vector per sentence [3].

Kah Phooi Seng, Li-Minn Ang, et al. have made feature extraction for a face with the approach found in the steps below. A combination of feature extraction techniques has designed using the Bi-directional Principal Component Analysis (BDPCA) and Least Square Linear Discriminant Analysis (LSLDA) to extract and discriminate the visual features amongst the six emotion classes. A new data fusion scheme called Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classification is proposed for the back-end processing in the visual path. Kernel and Laplacian mappings have first used to compute the similarities of the attribute-based and the relation-based (graph) information from the visual features in parallel. The kernels and Laplacians matrices have then optimized and merged to form an optimal Kernel-Laplacian matrix. Together with the mappings and optimized clustering, the radial basis function (RBF) neural network performs the classification [4].

Egils Avots, Tomasz Sapinski, et al. have made that facial images are labeled according to their emotions, before feature extraction and then, trained using Convolutional Neural Network (CNN - AlexNet Architecture) [5]. The transfer learning approach is used when training. The transfer learning approach is where a pre-trained network is used as a starting point to learn a new task. Then, images are randomly translated in X and Y directions in the range of −30 to 30 pixels to ensuring that CNN learns general features. For CNN tuning, according to recommended setup based on MATLAB documentation, most important parameters can be found that are Weight Learn Rate Factor = 20, Bias Learn Rate Factor = 20, Mini Batch Size = 10, Max Epochs = 10, Initial Learn Rate = 1e-4, Validation Frequency = 3, Validation Patience = Inf [6].

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

2.4 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 proposed ELM-based fusion. The ELM is based on a single hidden layer feed-forward network. In this proposed emotion recognition system, they used two ELMs successively for fusion. After the feature extraction is performed, the outputs obtained are given to the input of the ELM. ELM-1 has that the number of hidden layer neuron is 100, and ELM-2 is that 250. The output scores are converted into probabilities using the softmax function and then fed into the model training part [2].

There are two approaches to fusion. These approaches are feature-level fusion approaches and score-level fusion approaches. In this study, the score-level approach was preferred. For the audio path, a sliding window is applied to the speech signal. 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 employ a Multi-stream Fused Hidden Markov Models (MFHMM) to implement model-level fusion. MFHMM combines bimodal information from audio and visual streams in terms of the maximum entropy principle and the maximum mutual information criterion [7].

Lin et al. emotion recognition for audio and visual streams are employed an error weighted semi-coupled Hidden Markov Models (HMM). Tripled Hidden Markov Models (THMM) is adopted to perform audio-visual emotion recognition [8].

2.5 Model Training

The probability distribution of the outputs of the fusion part is the input to the model. The model 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 probability distribution of the outputs of the ELM fusion is the input to the SVM for M. S. Hossain and G. Muhammad's proposed system. They tried two kernels for SVM that are Radial Basis Function (RBF) and Polynomial. The RBF kernel has performed better in the experiments. RBF the optimization parameter of the SVM was set to 1 and the kernel parameter was 1.5 [2].

Carl Busso and Zhigang et al. used a support vector machine classifier (SVC) with 2nd order polynomial kernel functions. Result of this SVC for emotion recognition, seen that SVC better than statistical classifiers. Then used the leave-one-out cross-validation for the resampling method [3].

2.6 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

2.7 Conclusion

In this literature review, more than 8 articles have been reviewed and summarized in the sections above. Over the last years, researchers have proposed a diversity of methods for audio-visual emotional recognition. While distinguishing human emotions remains a challenging task, error rates have dropped significantly by the reason of progress in deep learning.

3. Summary

3.1 Summary of Conceptual Solution

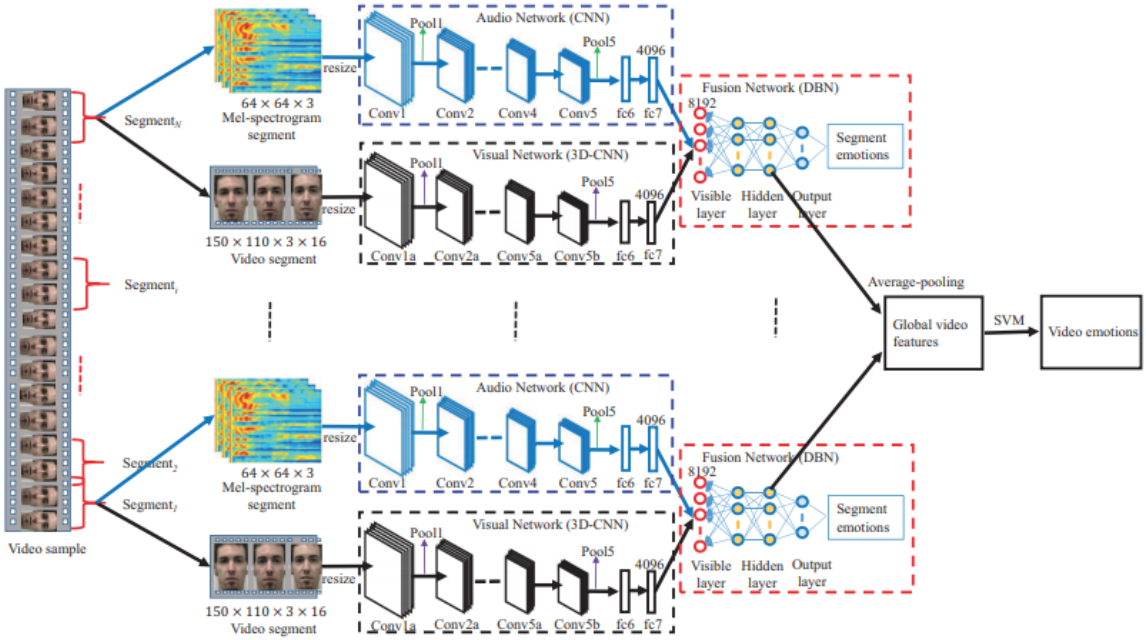

Figure I – The Structure of Deep Model for Audio-Visual Emotional Recognition

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

The purpose of this System Requirement Specification (SRS) document is explaining the system which is called audio-visual emotional recognition. This system goal to provide recognition of emotions from an image, speech, and video.

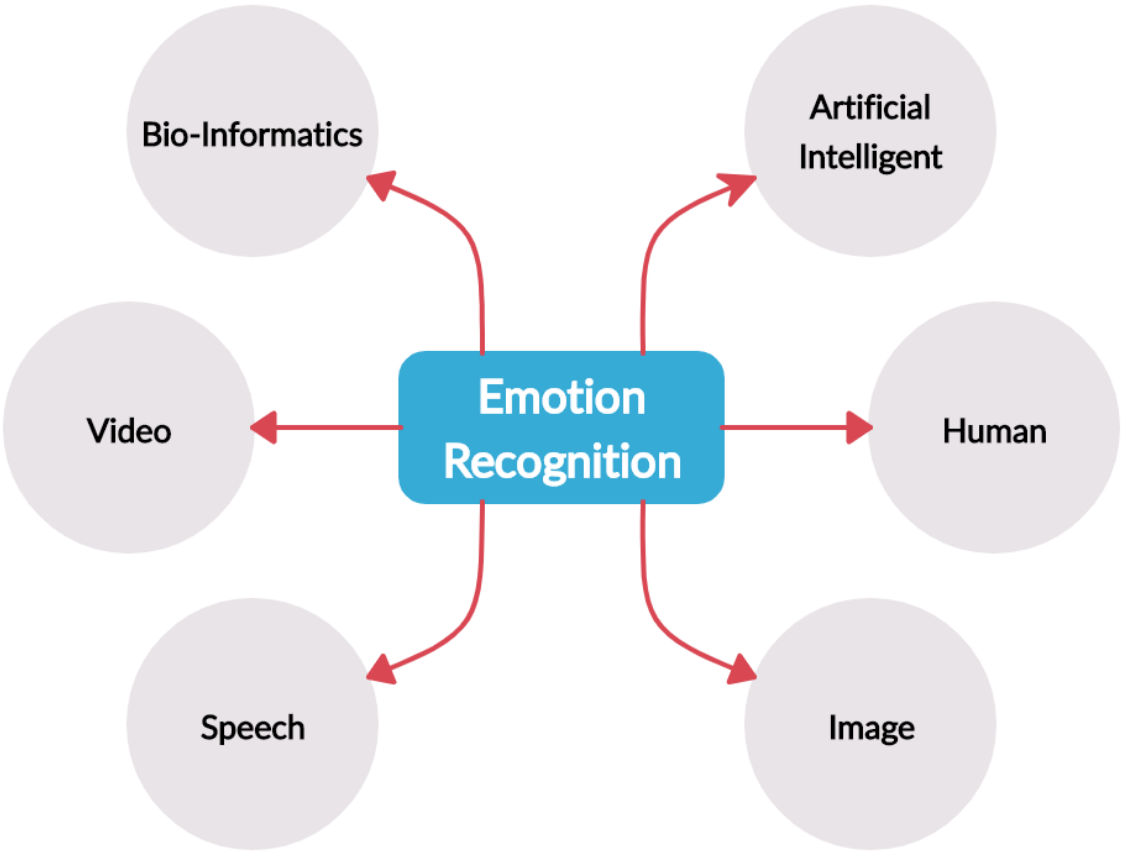

Figure II – Project Overview

This document includes detailed information about the requirements of the project. It 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 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 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 containing lossy and compressed image data [10].
Png	Portable Graphics Format. It is the most frequently used uncompressed raster image format on the internet [11].
Usb	Universal Serial Bus
Content	Image-Speech-Video

4.1.4 Overview of Document

The rest of this document is organized as follows: Section 2 explains the description of the project and properties for users who use the system and read the document. The constraints and risks of this system are mentioned. Section 3 is mainly written for developers of this system and describes in technical terms the details of the requirements of this system. The functions used by the user in order to use the project software and the tasks of these functions are described.

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 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.

4.2.2 User Characteristics

The person who will use this software must have basic computer knowledge. The user must read the user manual and apply it.

4.2.3 Constraints

Video, audio and image files must be in a specific format. The video format must be mp4, the audio format must be waw and image format must be png or jpeg. Video and audio files must have a maximum duration of 10 seconds. Image files must have at least 640x480 formats.

4.2.4 Risks

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

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

4.3 Requirements Specification

4.3.1 External Interface Requirements

4.3.1.1 System Interfaces

This part explained in 3.2 Functional Requirements.

4.3.1.2 User 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 shoot videos,
- Can externally add files,
- Optionally, contact information can be specified,
- Optionally, can 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 the information uploaded by the user,
- Can data statistics in uploaded files (female, male, age range, country, natio).

4.3.1.3 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.

4.3.1.4 Software interfaces

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

4.3.1.5 Communications interfaces

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

4.3.2 Functional Requirements

4.3.2.1 Profile Management Use Case

Use Case:

- Login
- Sign Up
- Validation
- Exit

Diagram:

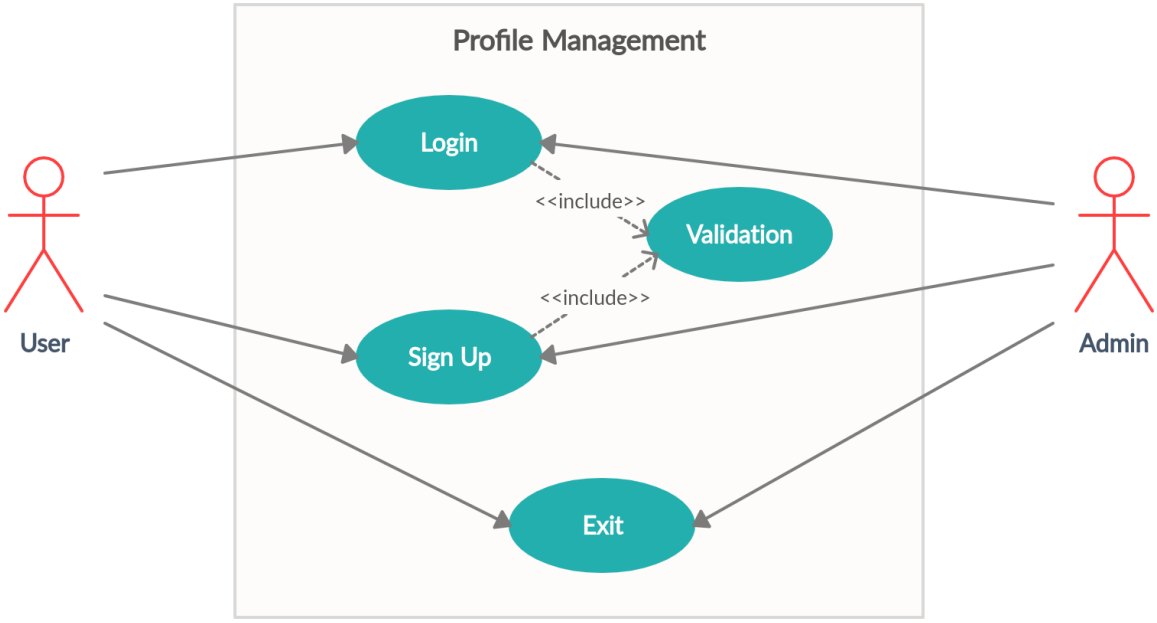

Figure III – Profile Management Use Case

Brief Description:

Figure III 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:

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

4.3.2.2 User Use Case

Use Case:

- Record Content
- Upload Content
- Get Result
- Content Information
- Show
- Add
- Edit
- Delete

Diagram:

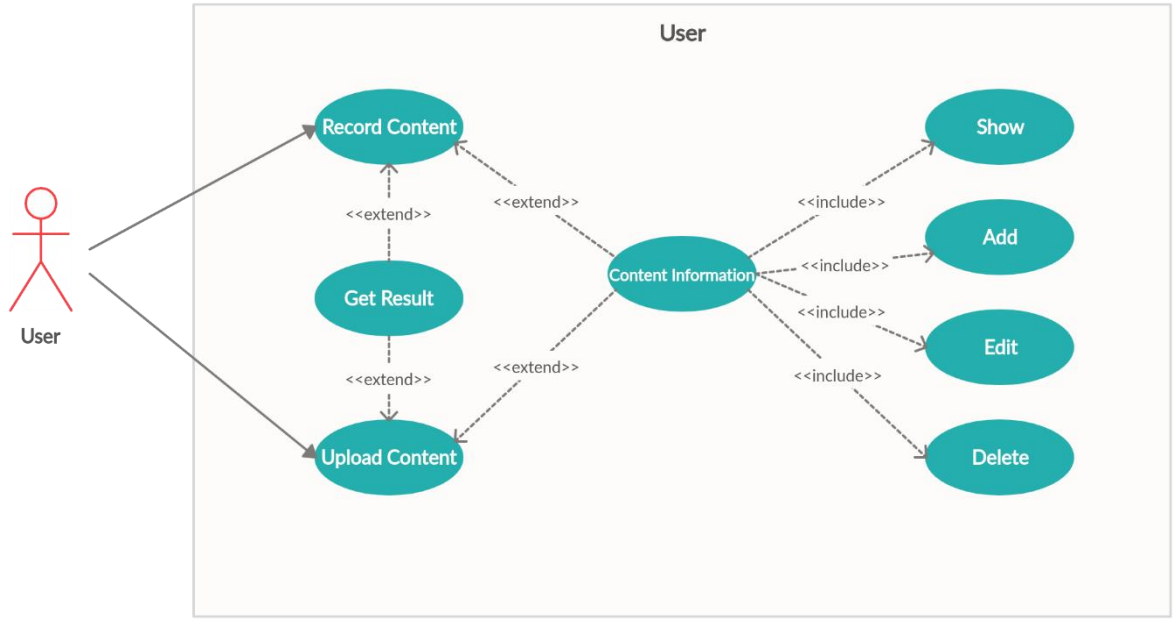

Figure IV – User Use Case

Brief Description:

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

Initial Step by Step Description:

1. If user selects Record Content, asks the user what kind of video/image/sound file to save.
    i. If user selects video, the system will be activated camera and microphone.
    ii. If user selects image, the system will be activated just camera.
    iii. 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, png/jpeg, waw).
3. After user have uploaded or recorded content, optionally can enter, view, delete or edit content's information.
4. Get Result; after user have uploaded or recorded content, the system will give you the result as an emotion.

4.3.2.3 Admin Use Case

Use Case:

- List Content
- System Train
- System Test
- Analyze
- Show Content
- Content Information
- Add
- Edit
- Delete

Diagram:

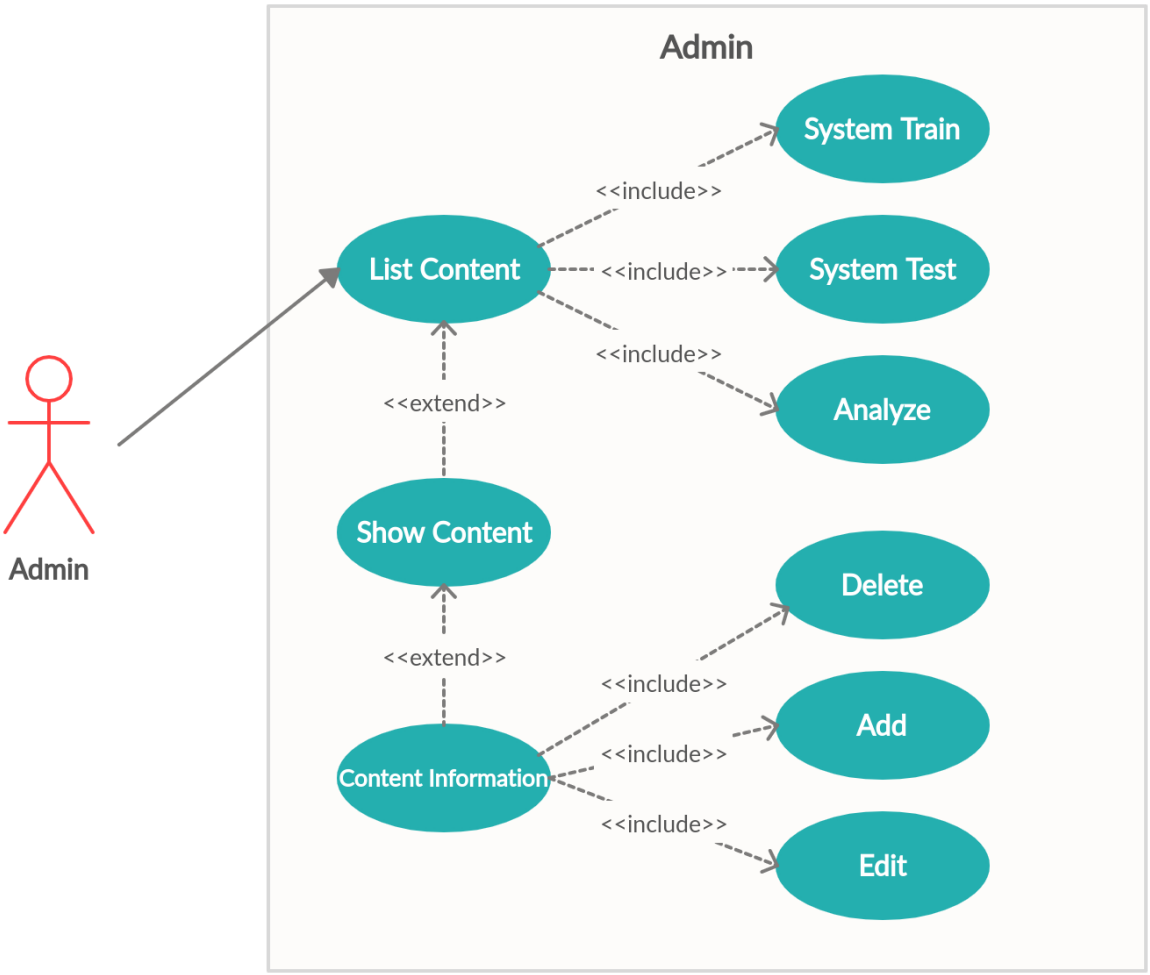

Figure V – Admin Use Case

Brief Description:

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

Initial Step by Step Description:

1. Admin can see all uploaded or recorded 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.

4.3.3 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: Windows 7 or later, macOS, and Linux
- Python versions: 3.6.X or higher
- Included development tools: Anaconda
- Compatible tools: Microsoft Visual Studio, PyCharm, Spyder or VSCode

4.3.4 Software system attributes

4.3.4.1 Reliability

System reliability will improve as long as the video's image/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.

4.3.4.2 Availability

The system will work on all operating systems.

4.3.4.3 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.

4.3.4.4 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.

4.3.4.5 Ease of Use

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 an image, speech, and video.

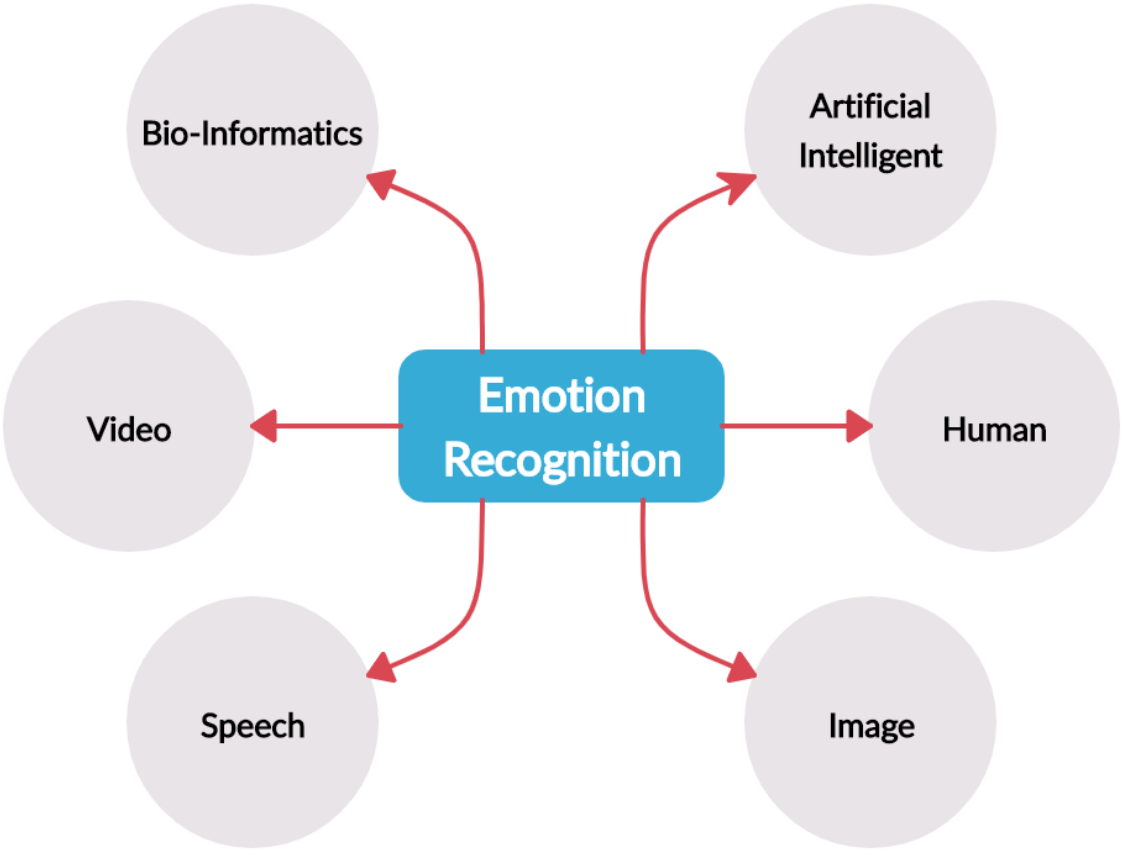

Figure VI – Project Overview

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 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 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 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	Image-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.

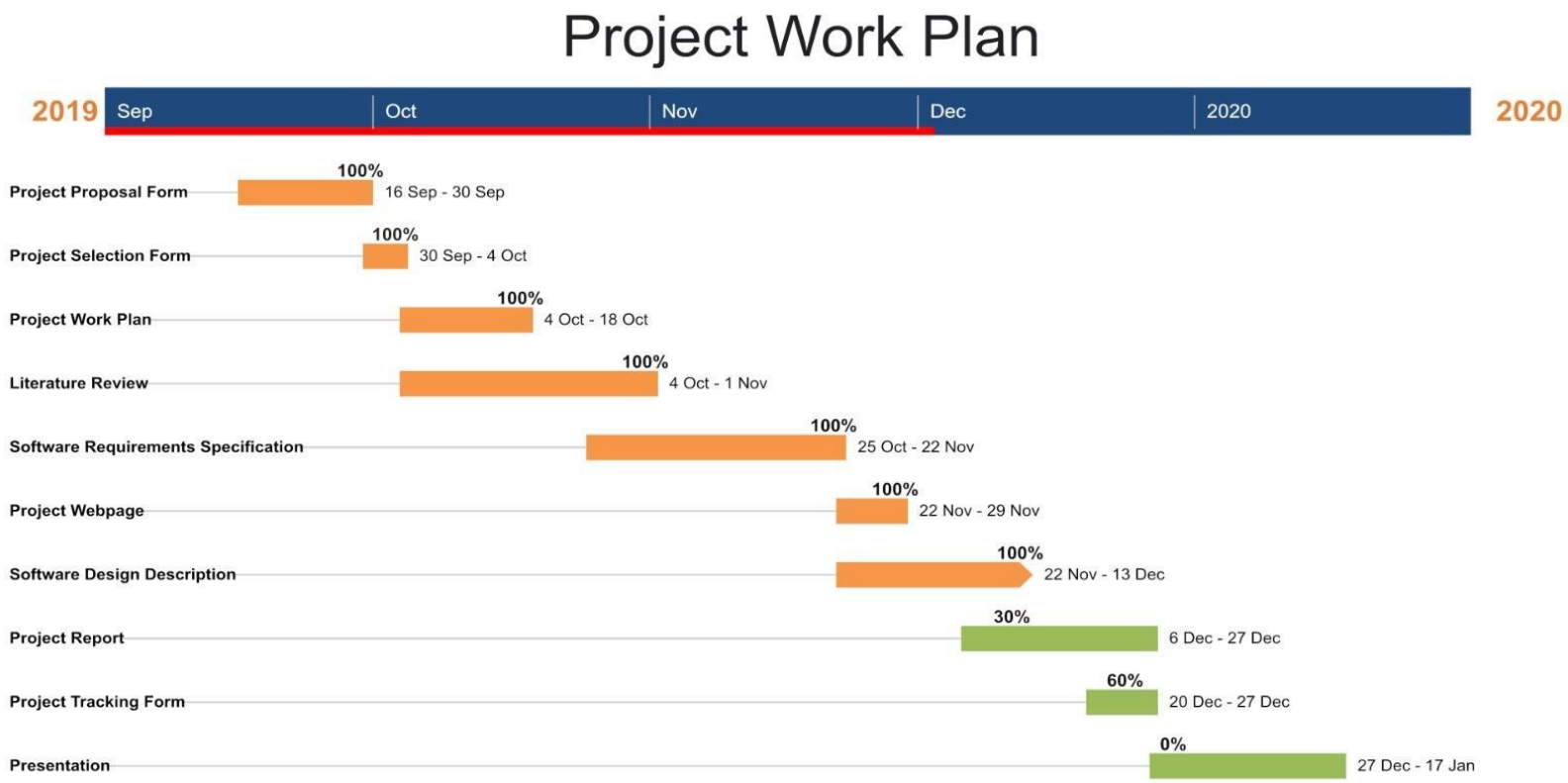

Figure VII – Project Work Plan

5.3.1.1 Class Diagram

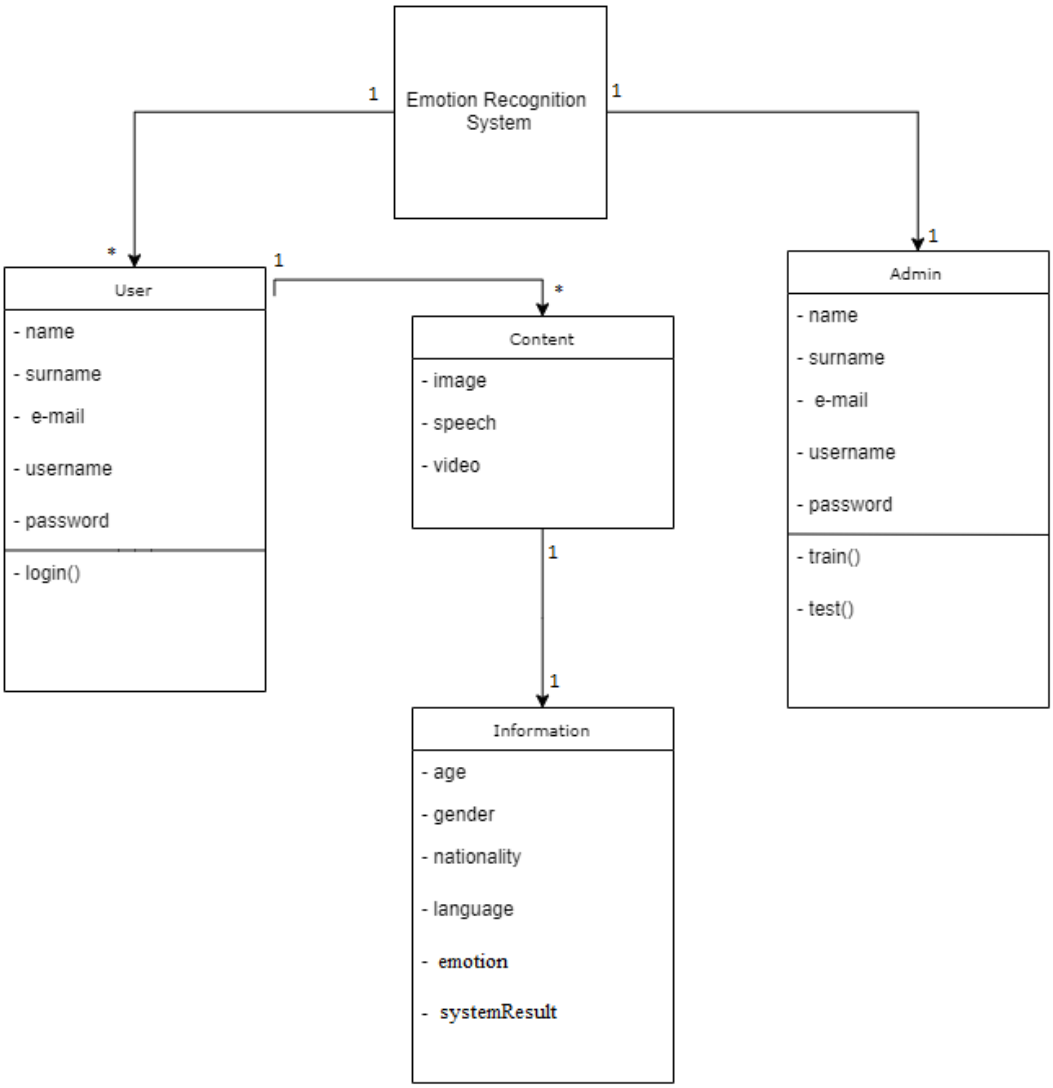

Figure VIII – Class Diagram

5.3.1.2 Activity Diagram

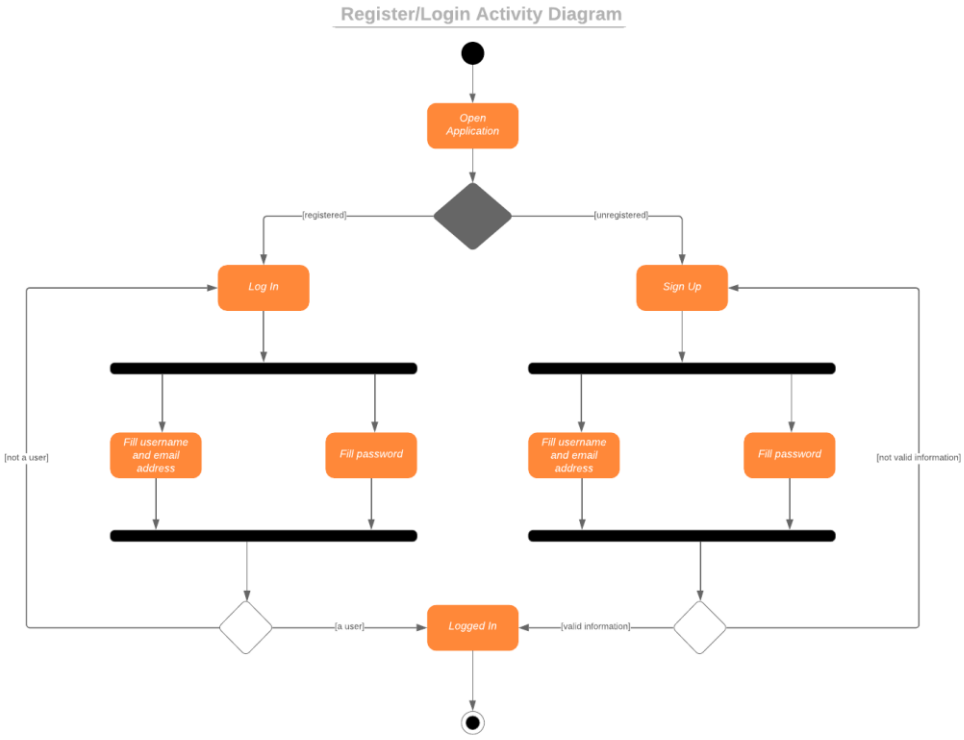

Figure IX – Register and Login Activity Diagram

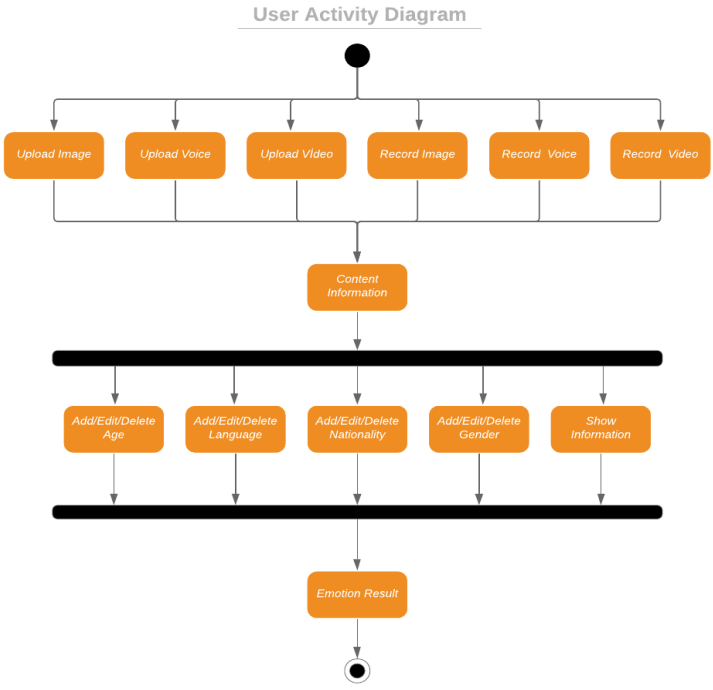

Figure X – User Activity Diagram

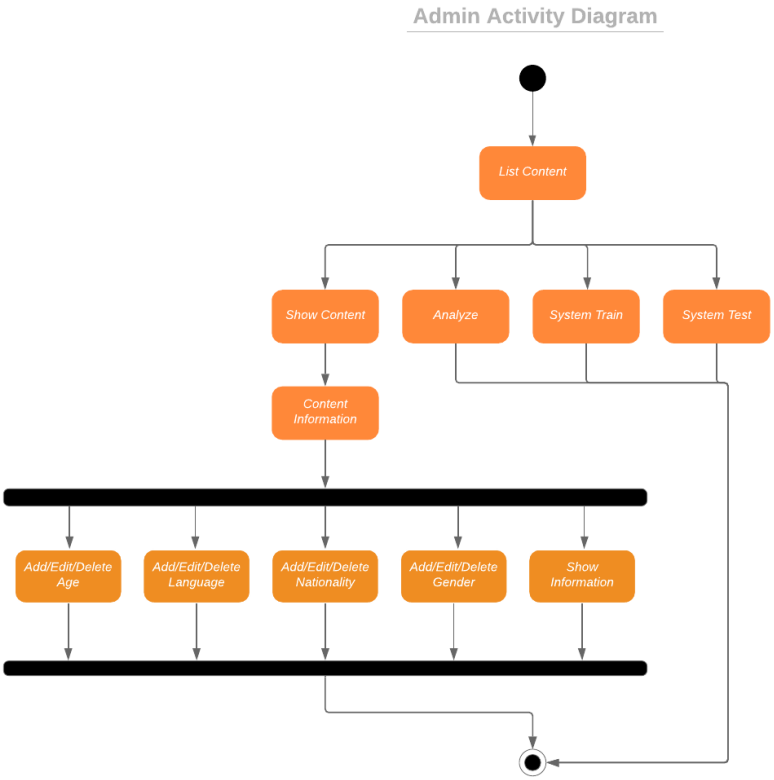

Figure XI – Admin Activity Diagram

5.3.1.3 Database Diagram

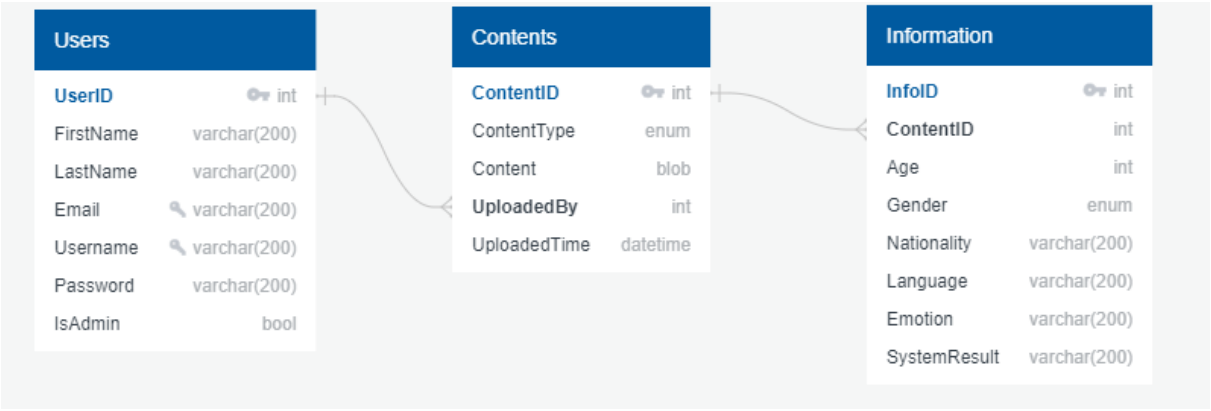

Figure XII – Database Diagram

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: Users and admin must login the system. 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 or recorded 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: Record Content, Upload Content, Get Result, also Show, Edit, Add and Delete in Content Informations.

Initial Step by Step Description:

1. If user selects Record Content, asks the user what kind of video/image/sound file to save.
    i. If user selects video, the system will be activated camera and microphone.
    ii. If user selects image, the system will be activated just camera.
    iii. 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, png/jpeg, 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 or recorded content, the system will give you the result as an emotion.

5.4 Use Case Realization

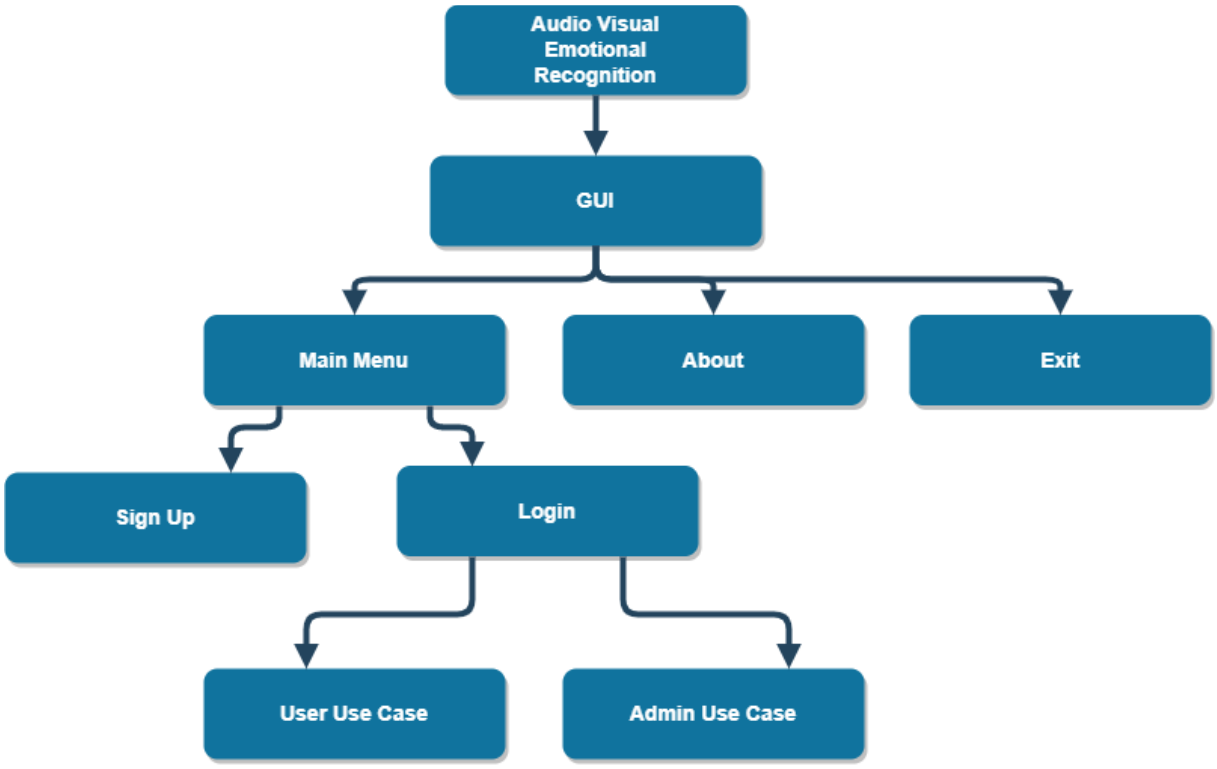

Figure XIII – Project Components

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

a. 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, image, 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.

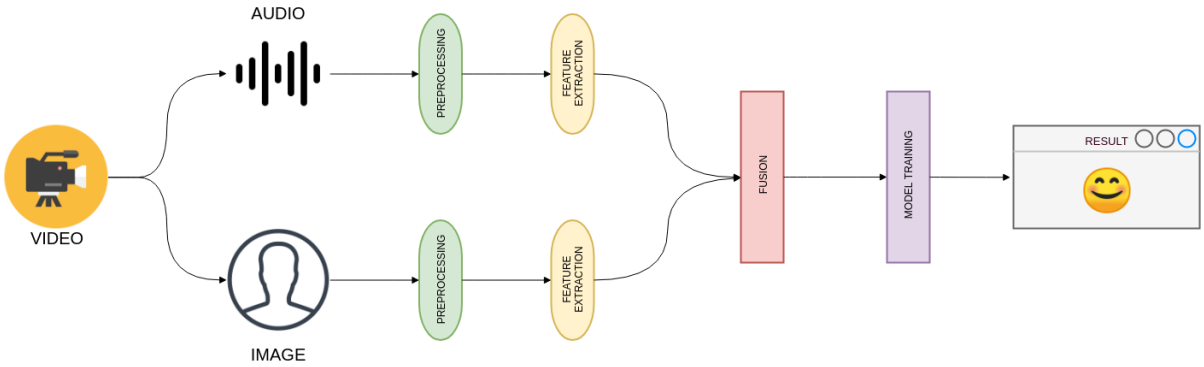

Figure XIV – System Overview

6. Conclusions

At the end of the Ceng407 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 and Software Design documents was written. Details of the studies performed in chapters 2, 4 and 5 can be examined.

7. Future Works

In the continuation of the project, implementation will be started as mentioned in our documents. The audio and video files will be processed separately and then combined with the fusion algorithm. After pre-processing our data will be trained with deep learning algorithms and then classified with machine learning techniques. Obviously, the problems may be in the fusion part. However, we believe that we can overcome this stage by trying many different algorithms. Finally, we will combine the user-interface of the analysis program will work properly and will be able to measure customer satisfaction properly. We hope that our system, which will be repeatedly trained and tested by a specific person (admin) at certain time intervals, so will improve every day.

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]