

Multimodal Sentiment Analysis Using Audiovisual Format On Product Reviews

A project report submitted in partial fulfillment of the requirements for the award of degree

**BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING**

By

SAROJ FARHANA(N100312)

NAZMA BEGUM(N100330)

SIVA PARVATHI UPPALAPATI(N100731)

Under the Guidance of

Mrs.NAGARJUNA DEVI M.Tech

Assistant Professor in Department of Computer Science & Engineering



RGUKT-NUZVID

Nuzvid, Krishna, Andhra Pradesh - 521202.

April 2016



RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(A.P Government Act of 2008)

RGUKT – NUZVID

Nuzvid, Krishna, Andhra Pradesh - 521202

Ph:08656-235147; Telefax-08656-235150

CERTIFICATE OF COMPLETION

This is to certify that the work entitled, “**Multimodal Sentiment Analysis Using Audiovisual Format On Product Reviews**”, is bonafied work of **Saroj Farhana (N100312), NazmaBegum (N100330), SivaParvathi Uppalapati (N100731)**”, carried out under my guidance and supervision for the partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in the department of Computer Science and Engineering under RGUKT IIIT Nuzvid. This work is done during the academic session August 2015 – May 2016, under our guidance.

Mrs. NAGARJUNA DEVI

Project Supervisor
Assistant Professor Dept. of CSE
RGUKT IIIT Nuzvid

Mr. K K Singh

Head of Department
Assistant Professor Dept. of CSE
RGUKT IIIT Nuzvid



RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(A.P Government Act of 2008)

RGUKT – NUZVID

Nuzvid, Krishna, Andhra Pradesh - 521202

Ph:08656-235147; Telefax-08656-235150

CERTIFICATE OF EXAMINATION

This is to certify that we have examined the thesis entitled “**Multimodal Sentiment Analysis Using Audiovisual Format On Product Reviews**”, submitted by **SarojFarhana (N100312), NazmaBegum (N100330), SivaParvathi Uppalapati (N100731)** and hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance on partial fulfillment for the award of **Bachelor of Technology Degree** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusions drawn, as record in this thesis. It only signifies the significance of this thesis for the purpose for which it has been submitted.

Project Guide

Mrs. NAGARJUNA DEVI M.Tech

Assistant Professor Dept. of CSE

RGUKT - NUZVID

Project Examiner

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES



(A.P Government Act of 2008)

RGUKT – NUZVID

Nuzvid, Krishna, Andhra Pradesh - 521202

Ph:08656-235147; Telefax-08656-235150

DECLARATION

We, **SarojFarhana (N100312), NazmaBegum (N100330), SivaParvathi Uppalapati (N100731)** hereby declare that the project report entitled “**Multimodal Sentiment Analysis Using Audiovisual Format On Product Reviews** ” done by us under the guidance of **Mrs.NAGARJUNA DEVI M.Tech** is submitted for the partial fulfillment for the degree of Bachelor of Technologies in Computer Science and Engineering during the academic session December 2015 – April 2016 at RGUKT – Nuzvid.

We also declared that this project is result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references.

The results embodied in this project report have not been submitted to any other university of institute for the award of any degree or diploma.

Saroj Farhana(N100312)

Nazma Begum (N100330)

Siva Parvathi Uppalapati(N100731)

Place :Nuzvid

Date :

ACKNOWLEDGEMENT

We are deeply indebted to our guide **Mrs. Nagarjuna Devi** for her invaluable guidance, motivation and constant encouragement throughout the course of this work. Her help, advice and guidance have been a constant source of inspiration to us throughout the B.Tech course. We shall always cherish the time spent with her during the course of this work due to the invaluable knowledge gained in the field of reliability engineering.

We are grateful to **Prof.Raj Reddy (Chancellor, RGUKT)**, **Prof.Vijay Prakash(Vice Chancellor, RGUKT)** and **Mr.Venkata Dasu Veeranki(Director, RGUKT-NUZVID)** for their endeavor and extended support throughout our/my education in the institute.

We express gratitude to **Mr. Krishna Kumar Singh (HOD of CSE)** and other faculty members for being source of inspiration, and constant encouragement which helped us in completing the project successfully.

Our sincere thanks to batch mates of 2010CSE, who have made our stay at RGUKT – NUZVID , a memorable one.

Finally, we express gratitude to our parents for supporting us in every walk of our lives.

ABSTRACT

The main purpose of our project Multimodal sentiment analysis is to analyze the emotions, attitude and opinions from audio visual format. A company can improve the quality of its product and services by analyzing the reviews about the product. Sentiment analysis is widely used in managing customer relations. There are many textual reviews from which we cannot extract emotions by traditional sentimental analysis techniques. Some sentences in the textual reviews may derive deep emotions but do not contain any keyword to detect those emotions, so we use audiovisual reviews in order to detect emotions from the facial expressions and by analyzing the voice of the customer .

In our project we take audiovisual input and extract emotions from video and audio in parallel from audiovisual input, finally classify the overall review as positive , negative or neutral based on combining overall emotions detected from customers facial expressions and from his voice.

Keywords:- Analyzing reviews, facial expression, audiovisual.

TABLE OF CONTENTS

- TITLE PAGE
- CERTIFICATE OF COMPLETION
- CERTIFICATE OF EXAMINATION
- DECLARATION
- ACKNOWLEDGEMENT
- ABSTRACT
- LIST OF FIGURES

1. INTRODUCTION	1
2. LITERATURE REVIEW	
2.1 Keyword Spotting	3
2.2 Concept Based Approach	3
2.3 Sentiment Analysis	4
2.3.1 Document Level Sentiment Analysis	4
2.3.2 Sentence Level Sentiment Analysis	4
2.3.3 Entity and Aspect Based Sentiment Analysis	5
3. REQUIREMENTS	
3.1 Functional Requirements	6
3.2 Non-functional Requirements	6
3.3 Hardware Requirements	6
3.4 Software Requirements	6
3.5 Knowledge Requirements	7
4. AUDIOVISUAL FORMAT APPROACH	

4.1 Introduction	7
4.2 Types of Expressions	7
4.3 Proposed Work	8
4.4 Data Set Employed	9

5. IMPLEMENTATION

5.1 Analyzing Video Input	10
5.1.1 Face Detection	10
5.1.2 Feature Points Extraction	11
5.1.3 Calculating Distance	13
5.1.4 Estimating Probabilities	14
5.2 Analyzing Audio Input	14
5.2.1 Pitch	14
5.2.2 Intensity	15
5.2.3 Pause	15
5.2.4 Loudness	15
5.3 Fusion of both Audio and Video Features	15
5.4 Process Flow Chart	16

6. SOURCE CODE

6.1 Expressionrecog.m	17
6.2 bulddetector.m	19
6.3 detectFaceParts.m	19
6.4 Audioreview.m	20
6.5 Fusion.m	21
6.6 recoggui.m	22

7. TESTING

7.1 Steps To Perform Test	23
7.2 Fusion of Both Audio and Video Data	27
7.2.1 Audio Data	27
7.2.2 Video Data	29
7.2.3 Product Review polarity	31

8.CONCLUSION AND FUTURE IMPLEMENTATION	32
--	----

REFERENCES AND LINKS	33
----------------------	----

LIST OF FIGURES :

4.2 Showing different types of Expressions	8
5.1.1.1 Viola – Jones nose detection Algorithm	11
5.1.2 Shi&TomasiMin Eigen Value corner detector	13
5.1.3 Change in distance between feature points w.r.t to Base Image	14
5.4 Processing flow chart of multimodal sentiment analysis	16
7.1 Steps to perform Test	23
7.2 Testing on different Images	26
7.3 Intensity and pitch Graphs	27

1.INTRODUCTION

Sentiment Analysis can be many things, but in general it is a classification task. Sentiment Analysis has been one of the most targeted research topics on past decades. Given a document(e.g a review, a blog spot or a tweet),the goal is to automatically obtain its sentiment which is mostly considered as binary class problem(positive and negative) or is Multiclass problem (positive, negative and neutral).

To date, most of the works in sentiment analysis have been carried out on natural language processing. Available dataset and resources for sentiment analysis are restricted to text-based sentiment analysis only. With the advent of social media, people are now extensively using the social media platform to express their opinions. People are increasingly making use of videos (e.g. YouTube Videos, VideoLectures),images (e.g., Flickr, Picasa, Face- book) and audios (e.g., podcasts) to air their opinions on social media platforms. Thus, it is highly crucial to mine opinions and identify sentiments from the diverse modalities.

So far the field multimodal analysis has not received much attention and no prior work has specifically addressed extraction of features and fusion of information extracted from different modalities like video and audio. Research in this field is rapidly growing and attracting the attention of both academic and industry alike. This combined with advances in signal processing and AI has led to development of advanced intelligent systems that intend to detect and process affective information contained in multimodal sources. The majority of such state-of-the-art-frameworks however, rely on processing signal modality i.e., text, audio and video. Further all of these systems are known to exhibit limitations in terms of meeting robustness, accuracy and overall performance requirements , which in turn greatly restrict usefulness of such systems in real world applications.

The aim of multi-sensor data fusion is to increase the accuracy and reliability of estimates. Many applications, e.g., navigation tools, have already demonstrated the potential of data fusion. This depicts the importance and feasibility of developing a multimodal framework that could cope with all two sensing modalities: audio and video in human-centric environments. The way humans communicate and express their emotions and sentiments can be expressed as multimodal. The audio and visual modalities are concurrently and cognitively exploited to enable effective extraction of the affective information conveyed during communication.

With significant increase in the popularity of social media like Facebook and Youtube many users tend to upload their opinions on products in video format. On the contrary, people wanting to buy the same product, browse through on-line reviews and make their decisions. Hence, the market is more interested in mining opinions from video data rather than text data. Video data may contain cues to identify sentiment of the opinion holder relating to the product. Audio data with in a video expresses the tone of the speaker and visual data conveys the facial expressions, which in turn help to understand the affective state of the users. The video data can be a good source for sentiment analysis but there are major challenges that

need to be overcome. For example expressiveness of the opinions vary from person to person. A person may express his or her opinions more vocally while others may express them more visually.

Hence, when a person express his opinions more vocal modulation the audio data may contain most of the clues for opinion mining. However, when a person is communicative through facial expressions, then most of the data required for opinion mining, would have been found in facial expressions. So, a generic modal need to be developed which can adapt itself for any user and can give a consistent result. Our multimodal sentiment classification modal is trained on robust data and the data contains the opinions of the many users.

2.LITERATURE REVIEW

Micro blogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on microblogs. One challenge is to build technology to detect and summarize an overall sentiment.

Sentiment analysis has been a burning topic for quite a few years. Recently it is used as an effective tool to understand the opinions of the public and also in various social media application. With the recent growth of social websites like Facebook, YouTube and Amazon gathering public opinion i.e., honest feedback is relatively effortless. Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level.

The traditional Sentiment analysis can be done in following ways

2.1 Keyword Spotting

This approach classifies the text on basis of presence of keywords like Happy, Sad, Afraid and Bored. This is a very naïve approach having drawback in following areas:

a) It can't reliably classify negated sentences.

E.g.:- "It was a **happy** moment" and "It wasn't a **happy** moment".

Both the sentences on the basis of keyword HAPPY will be classified as POSITIVE.

b) Sometimes meaning of a sentence conveys the emotion rather than any keyword in it.

E.g.:- "My husband decided to file for divorce and he wants to take custody of my children away from me".

2.2 Concept Based Approach

Sentiment analysis is done on the basis of web ontology. The system grasp the conceptual and affective information associated with natural language opinions. In this approach, keywords are not used blindly but it relies on implicit meanings associated with natural language. This approach heavily relies on knowledge base it uses. Textual reviews may involve ambiguous words, for e.g.:- *bomb*, which may lead to incorrect polarities assigned to reviews.

In such cases we can have videos which contain reviews about products and on the basis of those reviews we can assign the polarity to the product. Our project sentiment analysis on audiovisual format provides an opportunity to mine opinions and sentiment. The smiles, gazes, pauses and voice pitch are identified as relevant features. Many new areas such as facial expression, voice intensity, pauses, pitch etc. are used in opinion mining from audiovisual formats.

2.3 Sentiment Analysis :

With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls and focus groups in order to gather public opinions because there is an abundance of such information publicly available. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites. Each site typically contains a huge volume of opinion text that is not always easily deciphered in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are thus needed.

In general, sentiment analysis has been investigated mainly at three levels

1. Document Level
2. Sentence Level
3. Entity and Aspect Level

2.3.1 Document Level:

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment (Pang, Lee and Vaithyanathan, 2002; Turney, 2002). For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

2.3.2 Sentence Level:

The task at this level goes to the sentences and determines whether each sentence expressed as a positive, negative or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to *subjectivity classification* (Wiebe, Bruce and O'Hara, 1999), which distinguishes sentences (called *objective sentences*) that express factual information from sentences (called *subjective sentences*) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., “We bought the car last month and the windshield wiper has fallen off.” Researchers have also analyzed clauses (Wilson, Wiebe and Hwa, 2004), but the clause level is still not enough, e.g., “Apple is doing very well in this lousy economy.”

2.3.3 Entity and Aspect Level:

Both the document level and the sentence level analysis do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called *feature level (feature-based opinion mining and summarization)* (Hu and Liu, 2004). Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a *sentiment* (positive or negative) and a *target* (of opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. For example, although the sentence “*Although the service is not that great, I still love this restaurant*” clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the *restaurant* (emphasized), but negative about its *service* (not emphasized).

In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence “*The iPhone’s call quality is good, but its battery life is short*” evaluates two aspects, *call quality* and *battery life*, of *iPhone* (entity). The sentiment on iPhone’s *call quality* is positive, but the sentiment on its *battery life* is negative. The *call quality* and *battery life* of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analysis.

Both the document level and sentence level classifications are already highly challenging. The aspect-level is even more difficult. It consists of several sub-problems. To make things even more interesting and challenging, there are two types of opinions, i.e., regular opinions and comparative opinions (Jindal and Liu, 2006b). A regular opinion expresses a sentiment only on an particular entity or an aspect of the entity, e.g., “Coke tastes very good,” which expresses a positive sentiment on the aspect *taste* of Coke. A comparative opinion compares multiple entities based on some of their shared aspects, e.g., “Coke tastes better than Pepsi,” which compares Coke and Pepsi based on their tastes (an aspect) and expresses a preference for Coke.

But all this existing traditional sentimental analysis techniques or approaches based on text are not able to provide exact opinion of the consumers. There are some textual reviews which express deep emotions but do not contain any keyword to detect that emotion. To provide solution to this problem now at present there is a shift of textual sentimental classification to Multimodal sentimental classification which is a base to our project .

3.REQUIREMENTS:

3.1 FUNCTIONAL REQUIREMENTS:

3.1.1 YouTube Dataset:

Trusted product review video set is collected which act as major input in sentiment calculation.

3.2 NON-FUNCTIONAL REQUIREMENTS:

3.2.1 Efficiency:

As we have calculated polarities considering number of occurrences there is a chance of getting correct polarity.

3.2.2 Accuracy:

As far as concerned our project provides a better accuracy in regards with paper chosen.

3.3 HARDWARE REQUIREMENTS:

System with Windows OS and 4 GB RAM.

3.4 SOFTWARE REQUIREMENTS:

Matlab R2013 version

PRAAT Tool

Format Factory

3.5 KNOWLEDGE REQUIREMENTS:

Matlab Image processing knowledge

PRAAT tool usage manual knowledge

4. AUDIOVISUAL FORMAT APPROACH

4.1 Introduction:

Multimodal sentiment analysis is the analysis of emotions, attitude and opinion from audiovisual format. There are many textual reviews from which we cannot extract emotions by traditional sentiment analysis techniques. Some sentences in the textual reviews may derive deep emotions but do not contain any keyword to detect those emotions, so we used audiovisual reviews in order to detect emotions from the facial expressions of the customer.

4.2 Types Of Expressions:

Various features such as the location of the eyes, eyebrows and mouth are used for the analysis of facial expressions.

There are three types of expressions namely:

1. Micro expressions:

These are the expressions with the duration of 1/50 to 1/25seconds. These expressions are generated when a person feels less confidence.





2. Macro expressions:

These are the expressions with the duration of 0.5 to 4 seconds. These expressions are generated when a person is confident.

3. Subtle expressions:

These expressions are generated when the intensity of emotion is not strong. In order to detect facial expressions the distance between several points like ends of eyebrows, tip of nose, ends of lips and tip of cheeks is calculated. The change in facial expressions is detected by detecting the changes in the distance between these points. Voice signals are used in order to increase the accuracy of emotion recognition called multimodal emotion recognition.

Fig1: Showing different Type of Expressions

Emotions	View	Description
<i>Happiness</i>		Showing interest, engagement with activity [3].
<i>Sadness</i>		No commitment to the activity, boredom.
<i>Confusion</i>		Difficulty in understanding things.
<i>Neutral</i>		No effects in facial expressions.

4.3 PROPOSED WORK:

We studied several statements that cannot be analyzed using traditional sentimental analysis techniques either due to ambiguity or multiple meaning of words. Some statements convey deep emotions but don't contain words to express them.

The list of studied statements is:

1. She thought she ditched me but I am free now.
2. I met my best friend by an accident.
3. I avoided an accident.
4. I was sad because I had no shoes until I met a man who had no feet.
5. My husband decided to file for divorce and he wants to take custody of our children from me.

In order to overcome these ambiguities in sentiment classification we decided to switch to multimodal sentiment analysis. Multimodal sentiment analysis involve more than one signal i.e., voice and video signals in order to enhance the accuracy in classifying emotions. In our project first we have analyzed video input detected emotion and then computed the emotions from audio input and fused both the inputs in order to obtain overall polarity.

4.4 DATA SET EMPLOYED:

Trusted Youtube product review videos are collected and used as an input to determine the polarity sentiment. Videos in the data set were about different topics(for instance politics, electronic product reviews, etc.). The videos were found using following keywords : opinion, review, best, best perfume, toothpaste, war, job, business, cosmetics product reviews ,baby products etc .

The videos were converted to mp4 format a standard size of 360_480. The length of the videos were vary from 2 to 5 min. All videos were preprocessed to avoid the issues of introductory and multiple topics. Many videos on Youtube contained an introductory sequence where title was shown, sometimes accompanied by visual animation. To address this issue first 30 sec was removed from each video and then used for emotion detection.

5.IMPLEMENTATION

5.1 ANALYZING VIDEO INPUT:

Preprocessed videos are received as input and then applied matlab code to convert the video frames to Image frames at 30 Hz frame rate. Emotion from individual Image frame is detected in following steps

5.1.1 Face Detection:

For Detecting the face in image frame we use Viola-Jones Face detection algorithm which detect the frontal face appearing in the image.

5.1.1.1 Viola–Jones Object Detection Framework

The Viola–Jones object detection frame work is the first object detection framework to provide competitive object detection rates in real-time proposed in 2001 by Paul Viola and Michael Jones. It was motivated primarily by the problem of face detection, although it can be trained to detect a variety of object classes. This algorithm is implemented in Open CV as `cvHaarDetectObjects()`. Viola Jones detector become famous due to its open source implementation in the OpenCV library. In order to find an object of an unknown size is usually adopted to work this field that possesses a high efficiency and accuracy to locate the face region in an image. The Viola - Jones method contains three techniques:

1. Integral image for feature extraction the Haar-like features is rectangular type that is obtained by integral image
2. Adaboost is a machine-learning method for face detection, The word “boosted” means that the classifiers at every stage of the cascade are complex themselves and they are built out of basic classifiers using one of four boosting techniques (weighted voting). The Adaboost algorithm is a learning process that is a weak classification and then uses the weight value to learn and construct as a strong classification.
3. Cascade classifier used to combine many features efficiently. The word “cascade” in the classifier name means that the resultant classifier consists of several simpler classifiers (stages) that are applied subsequently to a region of interest until at some stage the candidate is rejected or all the stages are passed. Finally, the model can obtain the non-face region and face region after cascading each of strong classifiers.

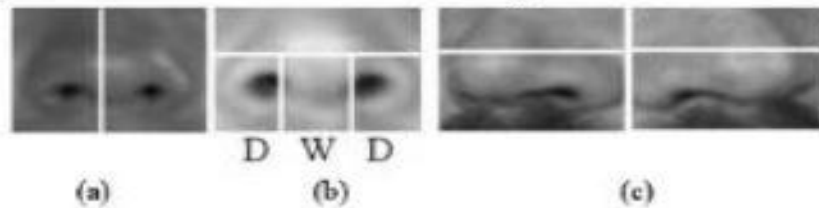
Viola-Jones Eye Detection Algorithm :

Eyes are detected based on the hypothesis that they are darker than other part of the face, finding eye analogue segments searching small patches in the input image that are roughly as large as an eye and are darker than their neighborhoods. A pair of potential eye regions is considered as eyes if it satisfies some constraints based on anthropological characteristics of human eyes. To discard regions corresponding to eyebrows, the model uses the fact that the center part of an eye region is darker than other parts. Then a simple

histogram analysis of the region is done for selecting eye regions since an eye region should exhibit two peaks while an eyebrow region shows only one.

Viola-Jones Nose Detection Algorithm:

Nose have three different local characteristics used as follows ;Similarity of both sides: The left and right sides of nose are similar in a front-view face as shown in Fig. (1a), this property of similarity can be measured using Euclidean distance between both sides.(ii) Dark-White-Dark (DWD) property: the lower part of nose region is characterized by two dark nostrils and a light sub region due to the reflection of light on the nose as shown in Fig



The two nostrils regions are less than the average of middle lighter sub region containing nose tip. (iii) The variation in lower/upper parts property: When the face is rotated some degrees for these two properties are despaired and the only clear property is the variation between lower part and upper part as shown in "Figure 2c". This variation can be measured by the variance in each part. Based on this analysis, a search is done for a certain region among the ten highest regions detected due to reflection of light at nose tip. Nose tip region is very bright as compared to other part of nose region. Due to the presence of very bright region, this region appears as black in binary image obtained in previous step because binary image and is generated by marking only dark pixels as white. To locate this region the algorithm find large connected black region in central region of localized nose image.

Viola-Jones Mouth Detection Algorithm

Detection and Extraction features from the mouth region; this model is composed of weak classifiers, based on a decision stump, which uses Haar features to encode mouth details. Experimental results show that the algorithm is Face image division based on physical approximation of location of eyes, nose and mouth on face and can find out the mouth region rapidly. It is useful in a wide range; moreover, it is effectual for complex background such as public mouth detection.

5.1.2 Feature Points Extraction:

After detecting the Face, Eyes, Nose and Mouth regions using Viola-Jones Detection Algorithm, to extract the strongest feature points and corner points within the bounded region of eyes, nose and mouth. There are various methods to extract the mouth corner points (interest points) such as:

1. Susan corner detector:

S.M. Smith and M. Brady introduced SUSAN operator, and find the interest point by comparing nucleus pixel to all pixel within circular mask, center pixel have intensity greater

than all the pixels in the mask like center of gravity rules and if we increase threshold at certain point it detects edges instead of corner.

2. Rosten & Drummond corner detector:

Rosten and Drummond detects the corner through local intensity comparisons of the pixels and is fastest corner detector take less time to all the other methods.

3. Harris & Stephens corner detector:

The Harris corner detector apply a selection criteria for detecting the interest points. For each pixel, a score is determined and if this score is greater than the certain value, the pixel is considered as a corner.

4. D. Shi & Tomasi Min Eigen Value corner detector:

We detect the interest points by this method and is based on the Harris & Stephens corner detection. The basic difference between these methods are that, the corner is detected in Shi and Tomasi method by calculating minimum of two eigen values of the matrix instead of calculating the score from the function F. Complete algorithm describe in following manner

$$F(u, v) = \sum_l \sum_m w(l, m) [i(l + u, m + v) - i(l, m)]^2$$

Where F is Sum of squared differences between the original and moved window, u - l is direction window displacement, v - m is direction window displacement, w(l, m) is Weighting function of the window, either a gaussian or a window of ones, i(l + u, m + v) is intensity of the moved window, l × m is window size.

Taylor series approximation of i(l + u, m + v) - i(l, m) are:

$$F(u, v) = \sum_l \sum_m w(l, m) [i(l, m) + u i_l + v i_m - i(l, m)]^2$$

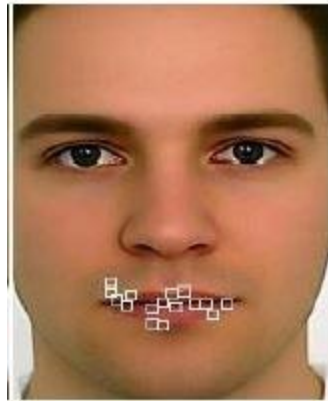
Now, matrix form of this approximation is -

$$F(u, v) = \sum_l \sum_m [uv] w(l, m) \begin{bmatrix} i_l^2 & i_l i_m \\ i_l i_m & i_m^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Then structure tensor form of final matrix is-

$$S = w(l, m) \begin{bmatrix} i_l^2 & i_l i_m \\ i_l i_m & i_m^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Now calculating the two eigen values (e1, e2) of S and take one which is minimum and consider as the corner point or interest point- $E = \min(e1, e2)$. For example figure below show feature points detected using Shi & Thomasi corner detector.



5.1.3 CALUCLATING DISTANCES:

The distance between various feature points is calculated in the baseline input image (d_i). Now the distance between the points in the test image (l) is calculated and compared with the baseline image distance using Euclidean distance method. A threshold value (μ) is used to classify this change in distance. Based on this difference between distances w.r.t. to threshold, the change indicator (C_i) is calculated as follows

$$c_i = \begin{cases} 1 & \text{if } d_i - d_l < \mu \\ -1 & \text{if } d_i - d_l > \mu \\ 0 & \text{if } d_i - d_l \text{ is very large} \end{cases}$$

Classification can be defined as

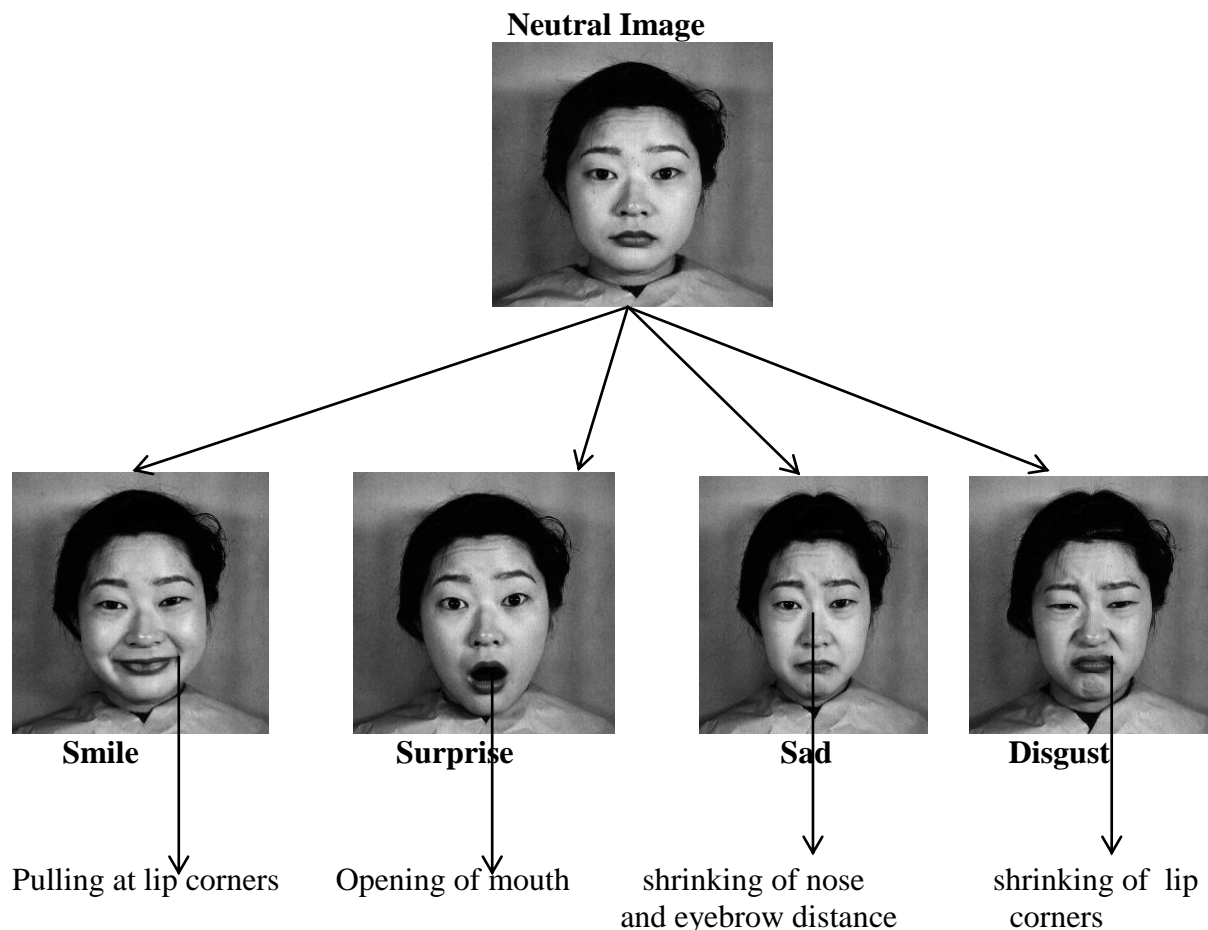
Classification of Emotions

Class	Description
Positive	If there is much change in the distance between feature points.
Negative	If there is less change in the distance between feature points.
Neutral	If there is no change in the distance between feature points.

5.1.4 ESTIMATING PROBABILITIES:

The probability (P_a) of classifying emotions into various classes like positive (smile, surprise), negative (sadness) and neutral (no change in emotions) depends on the value of (C_i) i.e., change estimated in feature points.

Fig1: Change in distance between feature points w.r.t to Base Image



5.2 ANALYZING AUDIO INPUT:

To analyze audio input first the product review videos are transcribed to audio format using Format Factory software. Analysis of audio input involves the analysis of four features that are pitch, pause, voice intensity and loudness. These four features are analyzed using PRAAT software tool as follows

5.2.1 PITCH:

It measures the variation in voice during entire video. Pitch is automatically computed using PRAAT tool.

5.2.2 INTENSITY:

It is the measure of sound power of spoken words. Intensity is automatically computed using PRAAT tool.

5.2.3 PAUSE:

It is the number of audio frames counted as silent. A threshold value is used to identify the samples with or without speech.

5.2.4 LOUDNESS :

Strength of voice is measured as loudness. Loudness in voice is controlled by the amplitude of sound waves and measured using PRAAT tool.

5.3 FUSION OF BOTH AUDIO AND VIDEO FEATURES

A statement is detected by detecting a pause in the audio frames; the polarity to statement is assigned on the basis of both extracted audio and video features.

1. Positive videos are characterized by increased number of smiles and an increased number of pauses.
2. Negative videos are characterized by higher voice intensity and sadness over face.
3. Neutral videos are monotonous in nature and involve no change in audio and video features.

The overall polarity assigned to product is the polarity assigned to maximum number of statements in the audiovisual review i.e., positive, negative or neutral. A count is performed in order to detect polarity occurred maximum number of times.

5.4 PROCESS FLOW CHART :

An audiovisual review is given as input to the system and the output from the system is the polarity assigned the audiovisual review. Both audio and video signals are analyzed in parallel with the same frame rate i.e., 30 Hz with the sliding window of 50 milliseconds.

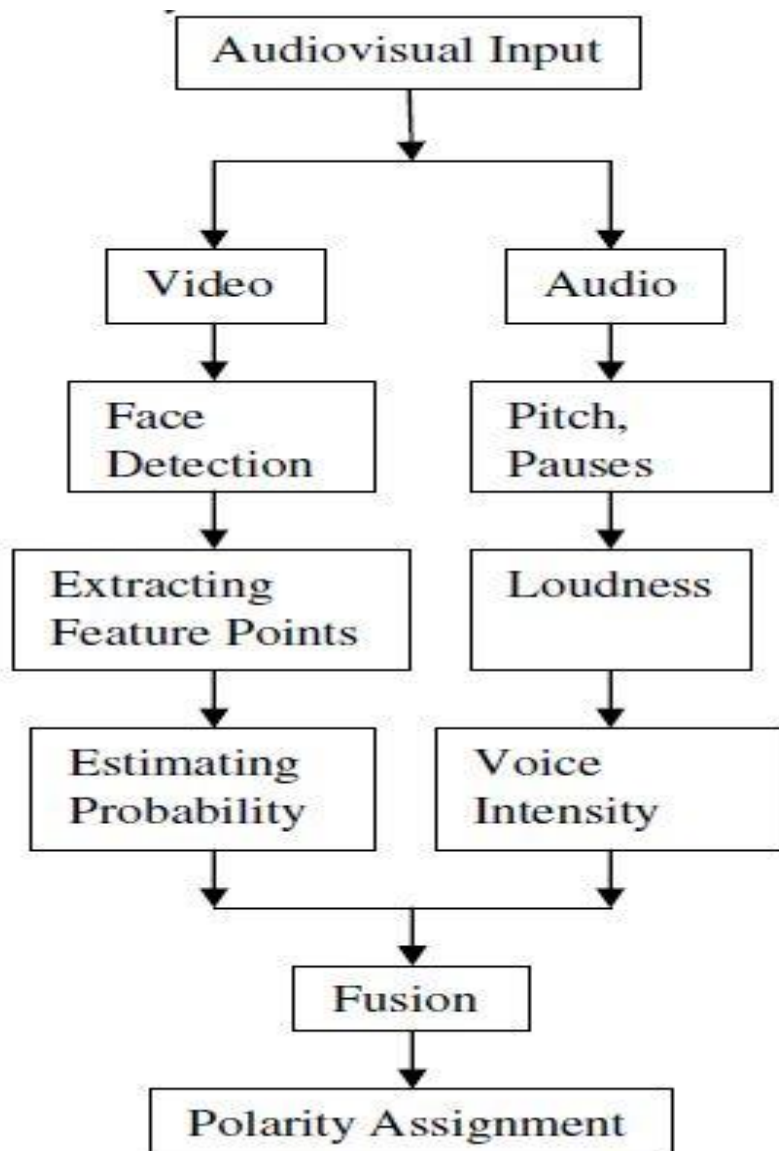


Fig 2: Processing Flowchart of Multimodal Sentimental Analysis

6.SOURCE CODE:

6.1 Expressionrecog.m

```
function[pos, neg] = Expressionrecog( )
    reqToolboxes = {'Computer Vision System Toolbox', 'Image Processing Toolbox'};
    if( ~checkToolboxes(reqToolboxes) )
        error('detectFaceParts requires: Computer Vision System Toolbox and Image
            Processing Toolbox. Please install these toolboxes.');
```

end

```
    s1=evalin('base','z1');%Read Image from workspace
    s2=evalin('base','z2');
    I=im2single(imread(baseimage));
    detector = buildDetector(); %call face detector function
    [bboxbbimg faces bbfaces] = detectFaceParts(detector,I,2); %detect face parts
    lefteye1= bbox(:, 5: 8);
    a=imcrop(I,lefteye1); %cropping the face eyes nose lip to extract feature points
    a=rgb2gray(a);
    righteye1=bbox(:, 9:12);
    nose1=bbox(:,17:20);
    b=imcrop(I,nose1);
    b=rgb2gray(b);
    mouth1=bbox(:,13:16);
    c=imcrop(I,mouth1);
    c=rgb2gray(c);
    for q=1:3
        F=im2single(imread(testimage)); %reading test image
        detector = buildDetector();
        [bboxbbimg faces bbfaces] = detectFaceParts(detector,F,2);
        lefteye2= bbox(:, 5: 8);
        a1=imcrop(F,lefteye2);
        a1=rgb2gray(a1);
        righteye2=bbox(:, 9:12);
        nose2=bbox(:,17:20);
        b1=imcrop(F,nose2);
        b1=rgb2gray(b1);
        mouth2=bbox(:,13:16);
        c1=imcrop(F,mouth2);
        c1=rgb2gray(c1);
        corners1 = detectMinEigenFeatures(a);%Detect Feature points
        eyepts1=corners1.selectStrongest(6);
        corners2 =detectMinEigenFeatures(a1);
        eyepts2=corners2.selectStrongest(6);
        corners3 = detectMinEigenFeatures(b);
        nose1=corners3.selectStrongest(4);
        corners4= detectMinEigenFeatures(b1);
        corners5 =detectMinEigenFeatures(c);
        mouth1=corners5.selectStrongest(8);
```

```

corners6=detectMinEigenFeatures(c1);
mouth2=corners6.selectStrongest(8);
x1=eyepts1.Location(1,:);y1=eyepts1.Location(2,:);%feature points location
x2=eyepts2.Location(1,:);y2=eyepts2.Location(2,:);
x3=nose1.Location(1,:);y3=nose1.Location(2,:);
x4=nose2.Location(1,:);y4=nose2.Location(2,:);
x5=mouth1.Location(1,:);y5=mouth1.Location(2,:);
x6=mouth2.Location(1,:);y6=mouth2.Location(2,:);
eye_dist=int32(sqrt(sum(((x1-y1)-(x2-y2)).^2))); %caluclating distances
nose_dist=int32(sqrt(sum(((x3-y3)-(x4-y4)).^2)));
mouth_dist=int32(sqrt(sum(((x5-y5)-(x6-y6)).^2)));
sup=0; sad=0; smi=0;neu=0;pos=0; neg=0;ang=0;
if(eye_dist<10 && nose_dist<10)
    disp('Test Image Expression is Disgust');
    ang=ang+1;
else if(nose_dist>10 && mouth_dist<10)
    disp('Test Image Expression is sad');
    sad=sad+1;
else if(mouth_dist>40 || eye_dist>40)
    disp('Test Image Expression is suprise');
    sup=sup+1;
else if(mouth_dist>10 && mouth_dist<40)
    disp('Test Image Expression is smile');
    smi=smi+1;
else
    disp('Test image Expression is neutral');
    neu=neu+1;
endendendend
end
if(smi>sup && smi>sad && smi>neu&&smi>ang)
    disp('Final video expressio is smile i.epositive');
    pos=1;
else if(sup>smi&& sup>sad && sup>neu&& sup>ang)
    disp('Final video expressio is suprisei.epositive');
    pos=pos+1;
else if(sad>smi&& sad>sup && sad>neu&& sad>=ang)
    disp('Final video expressio is sad i.e negative');
    neg=neg+1;
else if(ang>smi&& ang>sup && ang>neu&& ang>=sad)
    disp('Final video expressio is sad i.e negative');
    neg=neg+1;
else
    disp('Final video expression is neutral')
endendend
end

```

6.2 buildDetector.m

```
function detector = buildDetector( thresholdFace, thresholdParts, stdsize )

    if(nargin< 1 )
        thresholdFace = 1;end
    if(nargin< 2 )
        thresholdParts = 1;end
    if(nargin< 3 )
        stdsize = 176;end
    nameDetector = { 'LeftEye'; 'RightEye'; 'Mouth'; 'Nose'; };
    mins = [[12 18]; [12 18]; [15 25]; [15 18]; ];
    detector.stdsize = stdsize;detector.detector = cell(5,1);

    for k=1:4
        minSize = int32([stdsize/5 stdsize/5]);
        minSize = [max(minSize(1),mins(k,1)), max(minSize(2),mins(k,2))];
        detector.detector{k} = vision.CascadeObjectDetector(char(nameDetector(k)),
'MergeThreshold', thresholdParts, 'MinSize', minSize); %violo-johnes cascade detector
    end
    detector.detector{5} = vision.CascadeObjectDetector('FrontalFaceCART',
'MergeThreshold', thresholdFace);
end
```

6.3 detectFaceParts.m

```
function [bbox,bbX,faces,bbfaces] = detectFaceParts(detector,X,thick)

    if(nargin< 3 )
        thick = 1;end
    bbox = step(detector.detector{5}, X); %detect face parts
    bbsize = size(bbox);
    partsNum = zeros(size(bbox,1),1);
    nameDetector = { 'LeftEye'; 'RightEye'; 'Mouth'; 'Nose'; };
    mins = [[12 18]; [12 18]; [15 25]; [15 18]; ];
    stdsize = detector.stdsize;
    for k=1:4
        if( k == 1 )
            region = [1,int32(stdsize*2/3); 1, int32(stdsize*2/3)];
        elseif( k == 2 )
            region = [int32(stdsize/3),stdsize; 1, int32(stdsize*2/3)];
        elseif( k == 3 )
            region = [1,stdsize; int32(stdsize/3), stdsize];
        elseif( k == 4 )
            region= [int32(stdsize/5),int32(stdsize*4/5); int32(stdsize/3),stdsize];
        else
            region = [1,stdsize;1,stdsize];
        end
    end
```

```

bb = zeros(bbsize);
for i=1:size(bbox,1)
XX = X(bbox(i,2):bbox(i,2)+bbox(i,4)-1,bbox(i,1):bbox(i,1)+bbox(i,3)-1,:);
XX = imresize(XX,[stdsize, stdsize]);
XX = XX(region(2,1):region(2,2),region(1,1):region(1,2),:);
b = step(detector.detector{k},XX);
if( size(b,1) > 0 )
    partsNum(i) = partsNum(i) + 1;
    if( k == 1 )
        b = sortrows(b,1);
    elseif( k == 2 )
        b = flipud(sortrows(b,1));
    elseif( k == 3 )
        b = flipud(sortrows(b,2));
    elseif( k == 4 )
        b = flipud(sortrows(b,3));
    end
    ratio = double(bbox(i,3)) / double(stdsize);
    b(1,1) = int32( ( b(1,1)-1 + region(1,1)-1 ) * ratio + 0.5 ) + bbox(i,1);
    b(1,2) = int32( ( b(1,2)-1 + region(2,1)-1 ) * ratio + 0.5 ) + bbox(i,2);
    b(1,3) = int32( b(1,3) * ratio + 0.5 );
    b(1,4) = int32( b(1,4) * ratio + 0.5 );
    bb(i,:) = b(1,:);end end
bbox = [bbox,bb];
p = ( sum(bb') == 0 );
bb(p,:) = [];
end
b(1,3) = int32( b(1,3) * ratio + 0.5 );
b(1,4) = int32( b(1,4) * ratio + 0.5 );
bb(i,:) = b(1,:);
end
end
bbox = [bbox,bb];
p = ( sum(bb') == 0 );
bb(p,:) = [];
end
end

```

6.4 Audioreview.m

```

function [ posa,nega ] = Audioreview()
[File_NamePath_Name]=uigetfile({'*.xlsx'},'Select Audio Review Data');
audio=strcat(Path_Name,File_Name);
data=xlsread(audio); %reading audio data
intensity=data(:,1);
pitch=data(:,2);
a=max(intensity);b=min(intensity);x=mean(intensity);c=max(pitch);
d=min(pitch);y=mean(pitch);
disp('max Intensity:',a);

```

```

disp('min Intensity:',b);
disp('mean of Intensity:',x);
disp('max Pitch:',c);
disp('min Pitch:',d);
disp('mean of Pitch:',y);
count=0;xy=0;
for i=1:100
    if(pitch(i)==0)
        count=count+1;
    endend
disp('Pausescount',count);
posa=0;nega=0;
if(c>400 && count>35)
    disp('Audio review is postive');
    posa=posa+1;
else if(x>50 && count<35)
    disp('Audio review is negative');
    nega=nega+1;
else
    disp('Audio is neutral');
end
end
end
end

```

6.5 Fusion.m:

```

function [ finalotp ] = fusion()
    finalotp="";
    [posa,nega]=audioreview();
    [pos,neg]=Expressionrecog();
    finalp=posa+pos;
    finaln=nega+neg;
    if(finalp>finaln)
        finalotp='Product Review has Postive Sentiment';
    else if(finalp==finaln)
        finalotp='Product Review has Neutral Sentiment';
    else
        finalotp='Product Review has Negative Sentiment';
    end end
end
end

```

6.6 recoggui.m:

```
function pushbutton1_Callback(hObject, eventdata, handles)
    [File_NamePath_Name]=uigetfile({'*.jpg'},'File selector');
    axes(handles.axes1);
    trainimage=imread(fullfile(Path_Name,File_Name));
    assignin('base','z1',trainimage);
    imshow(trainimage);
function pushbutton2_Callback(hObject, eventdata, handles)
    [File_NamePath_Name]=uigetfile({'*.jpg'},'File selector');
    axes(handles.axes2);
    testimage=imread(fullfile(Path_Name,File_Name));
    assignin('base','z2',testimage);
    imshow(testimage);
function pushbutton3_Callback(hObject, eventdata, handles)
    output=expressionrecog();
    set(handles.edit1,'String',output);
function pushbutton4_Callback(hObject, eventdata, handles)
    areview="";
    [posanega]=Audioreview();
    if(posa>nega)
        areview='Audio Expression is postive';
    else if(nega>posa)
        areview='Audio Expression is Negative';
    else
        areview='Audio Expression is Negative';
    end end
    set(handles.edit1,'String',areview);

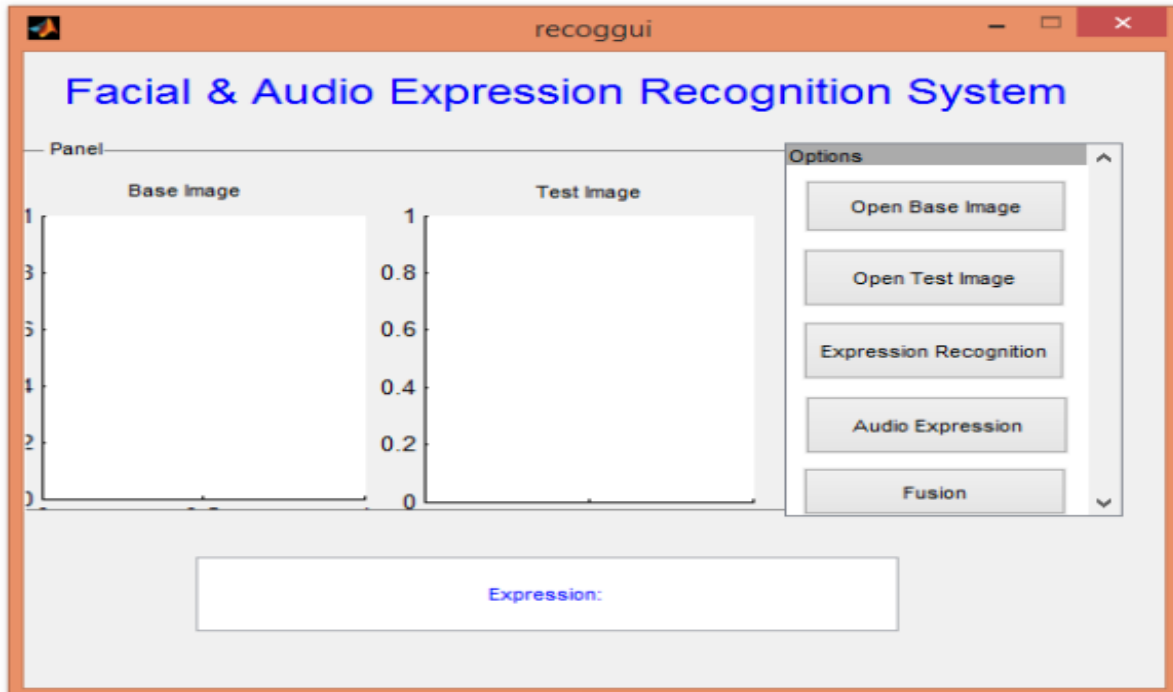
% Executes on button press in pushbutton5.

function pushbutton5_Callback(hObject, eventdata, handles)
    finalotp=Fusion();
    set(handles.edit1,'String',finalotp);
```

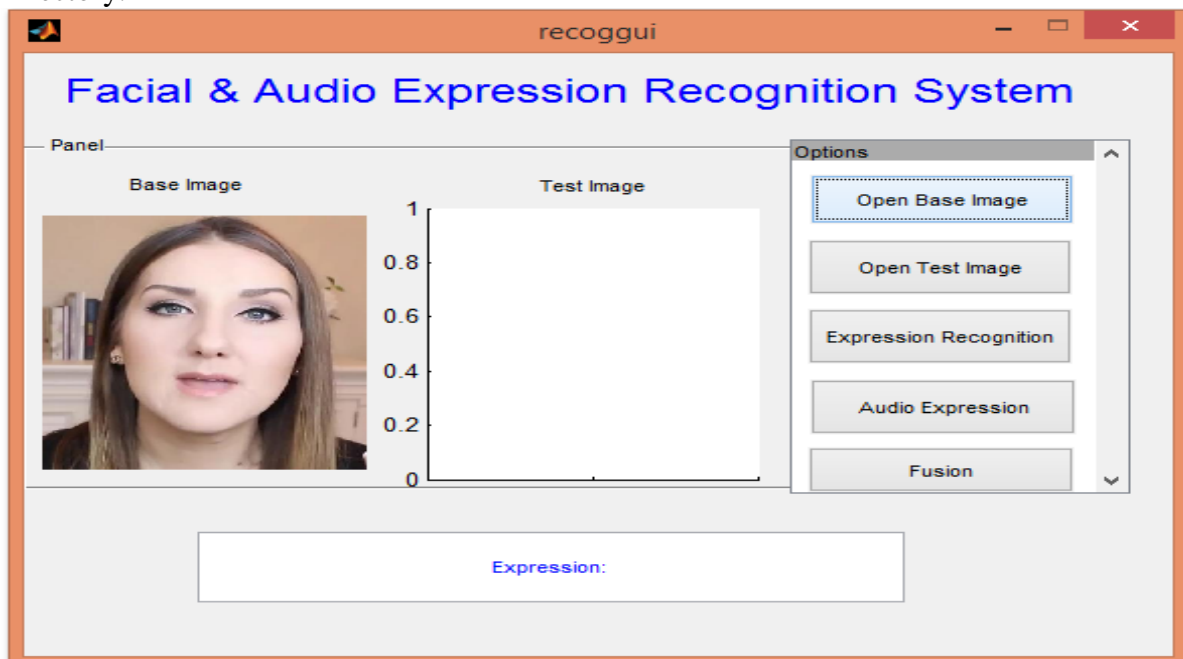
7.TESTING:

7.1 Steps To Perform Test:

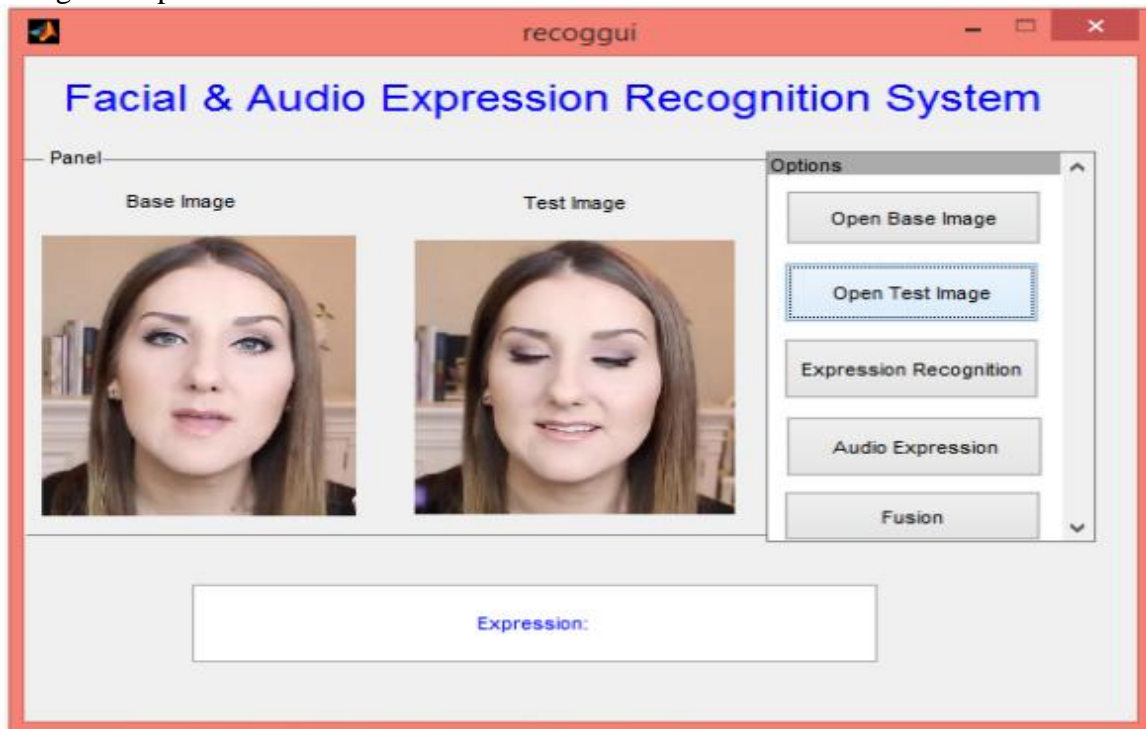
Step 1:Open Matlab and change the directory path to the location where project files are located and run recognitongui.m file then it opens GUI like this



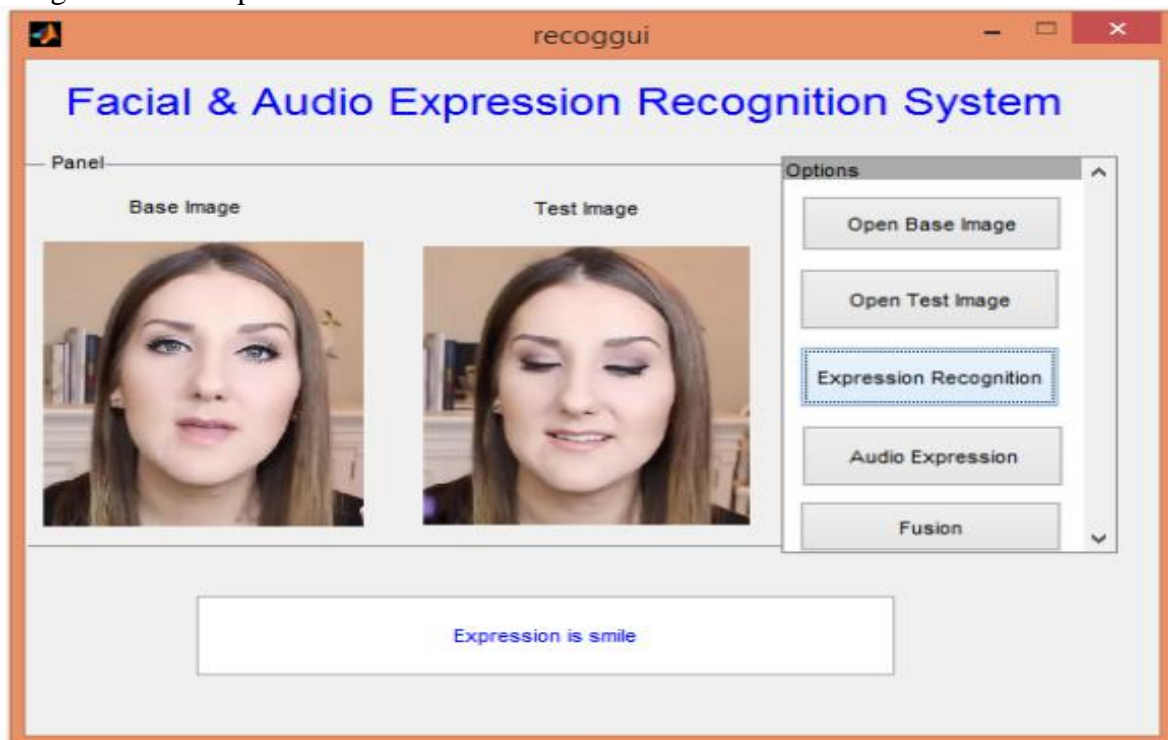
Step2: Click on the Open Base Image to open the base image and select the image from File Directory.



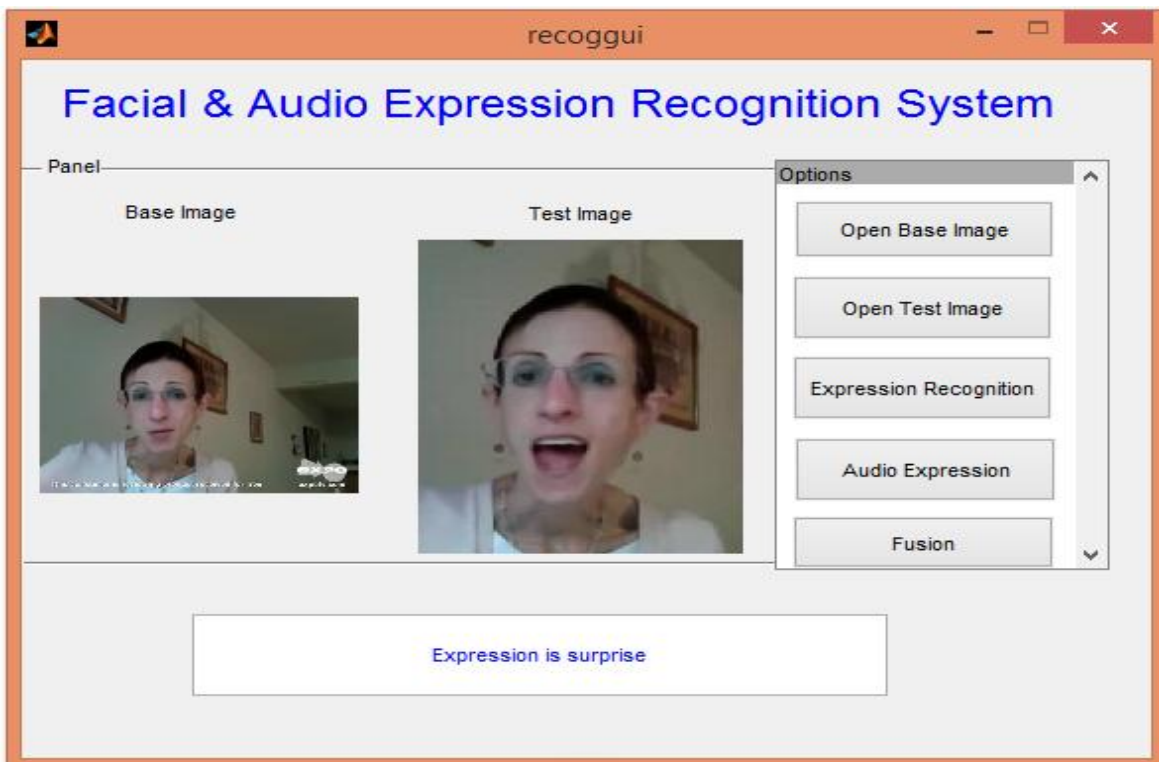
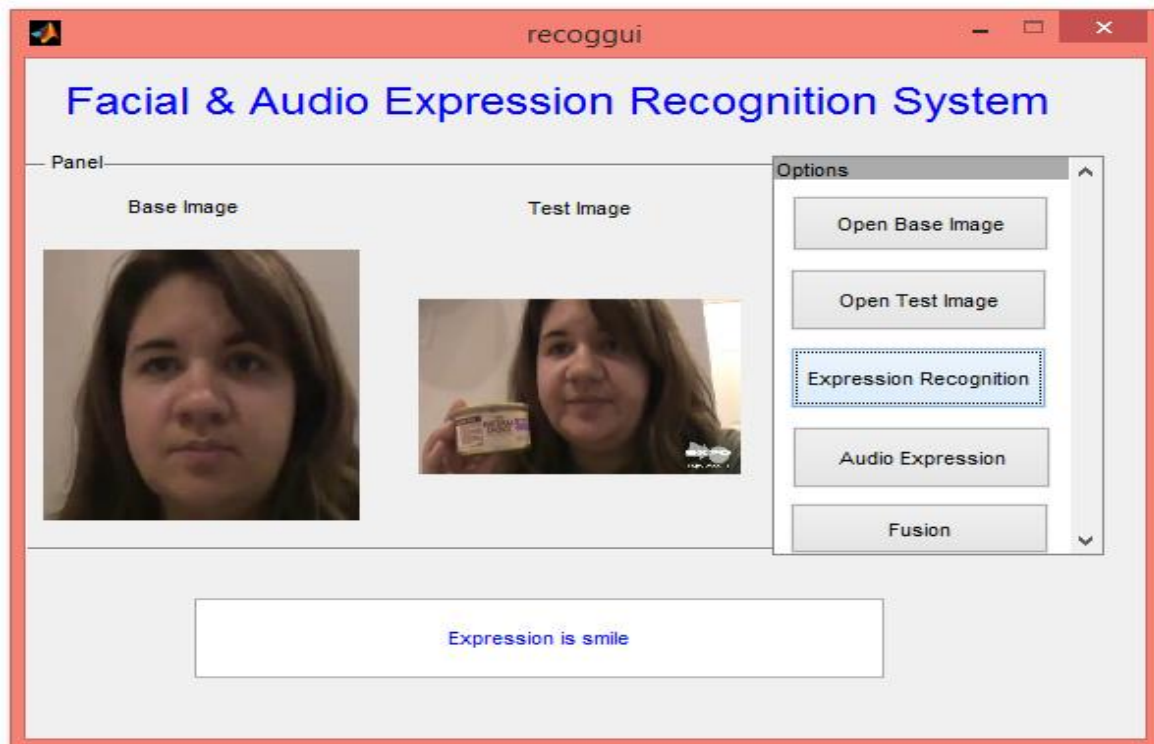
Step 3: Then click on Open Test Image button and select the test image to recognize the change in expression.

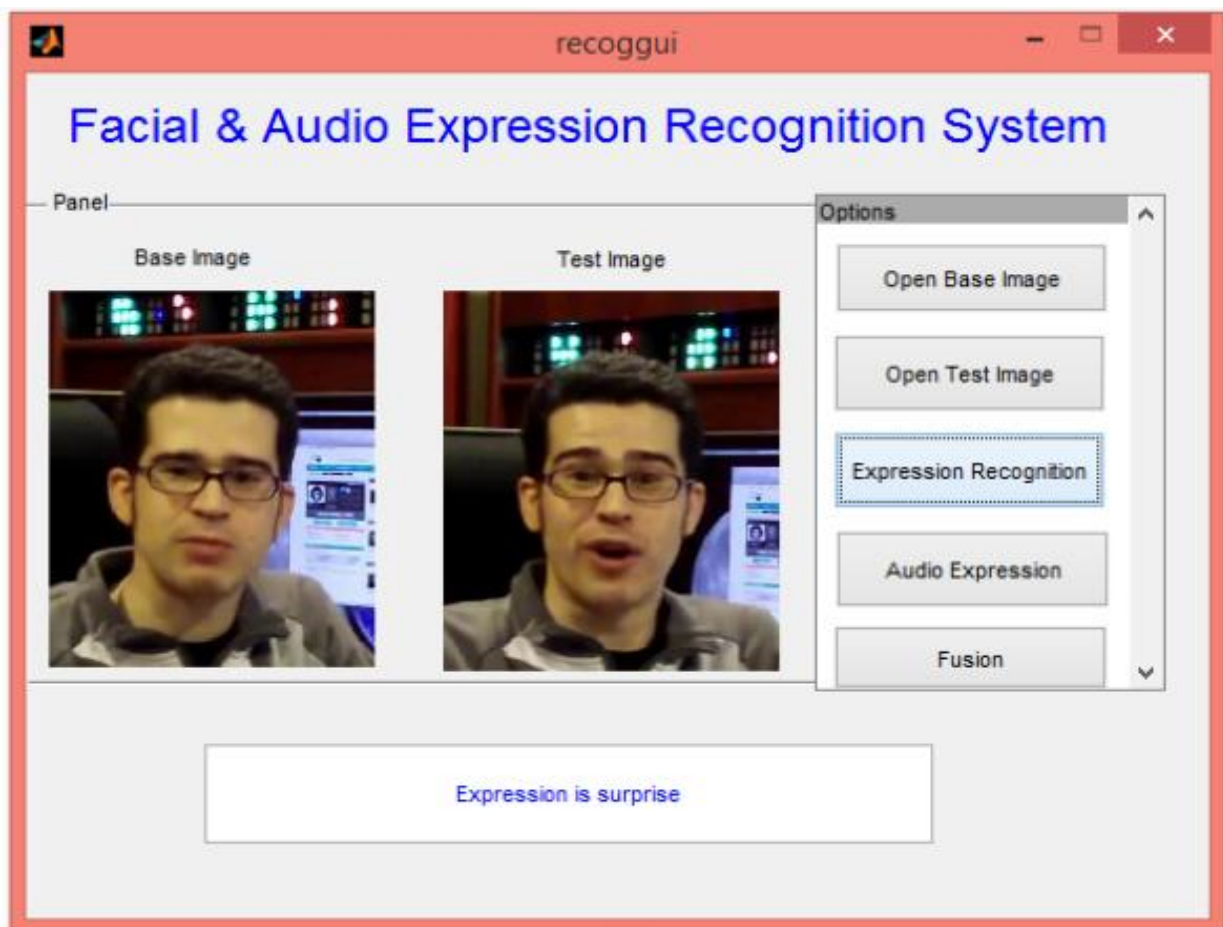
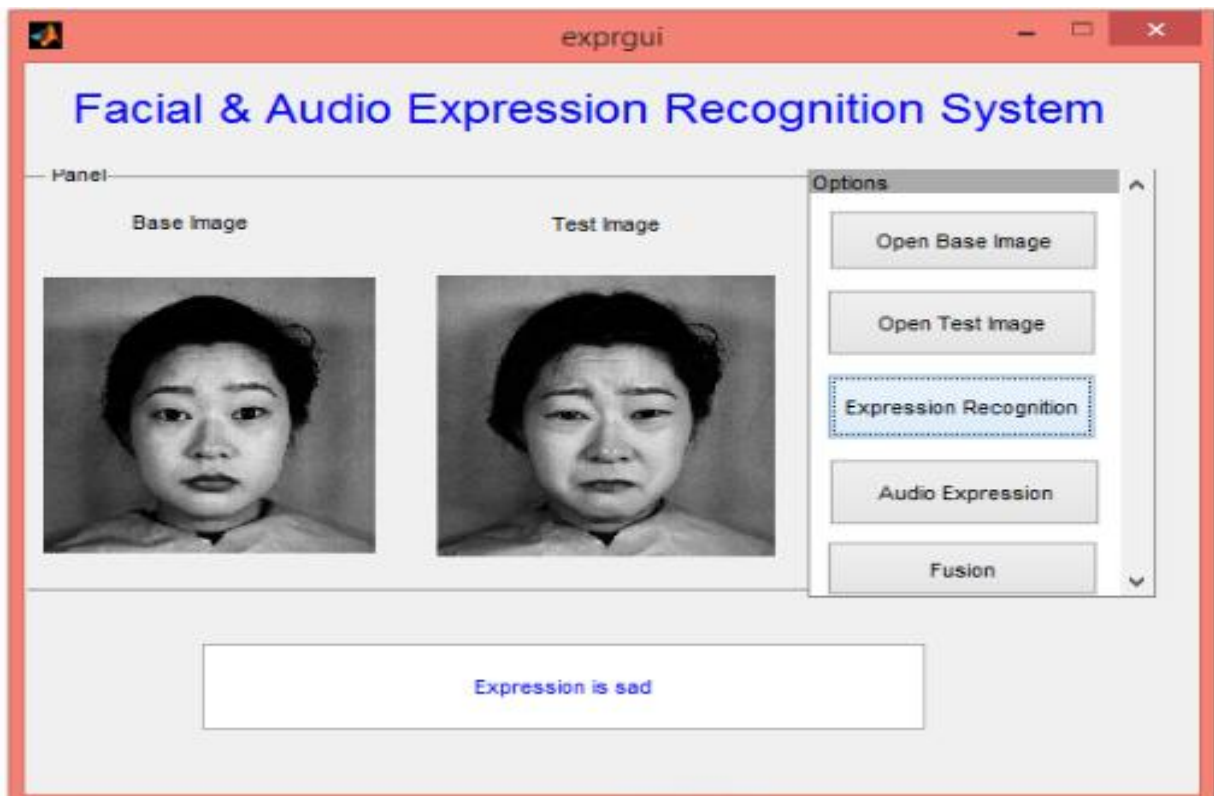


Step4: Click on Expression Recognition Button to recognize the change in expression from base image to test image. Can select many test images by keeping base image as constant to recognize video expression.



Step5: In the same way we can perform test on different faces to recognize the change in expression





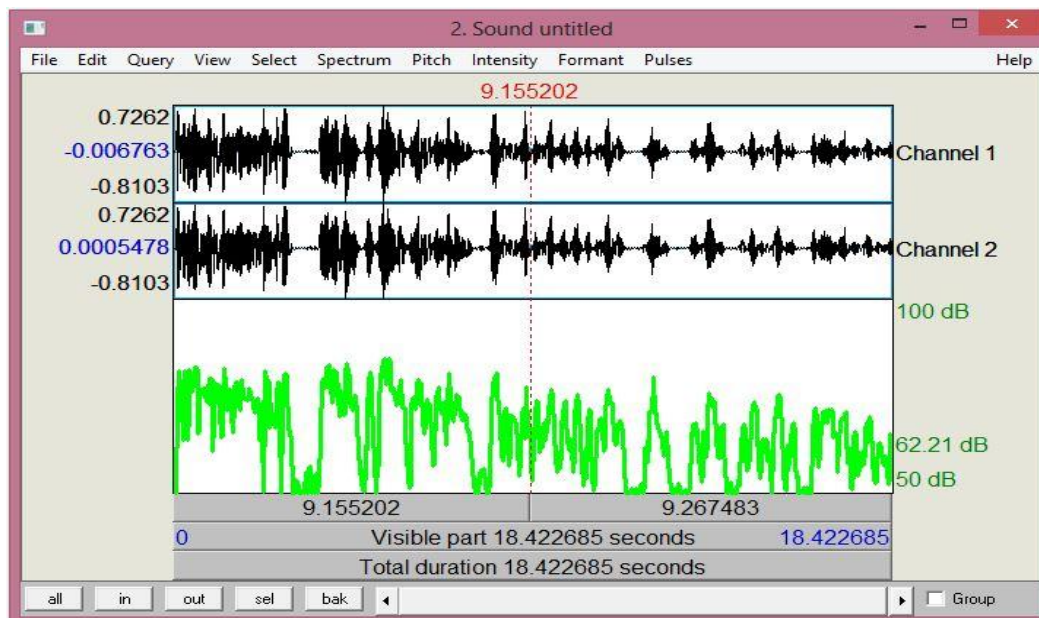
7.2 Fusion of Both Audio and Video Data:

To calculate the overall polarity of the video review we have to fusion the polarity from audio review and video review ,to do so click on Fusion button which ask to select audio input and calculate it polarity and then ask to select set of test images , compare the change in expression with base image and calculate the video polarity as mostly occurring expression and then combine both audio and video polarities and display in GUI the final product review sentiment as positive ,negative or neutral. Process is as follows.

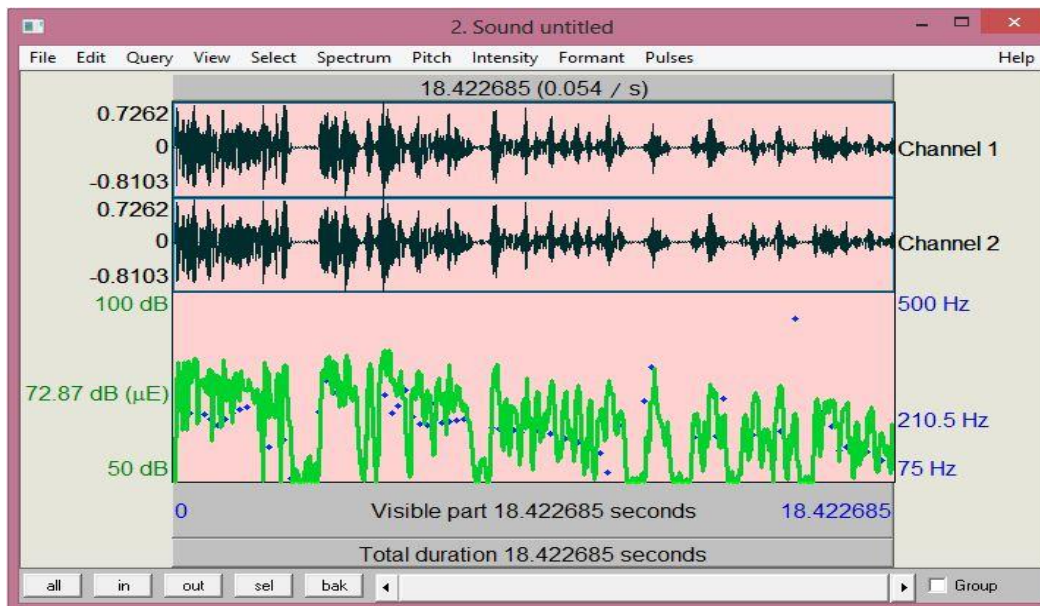
7.2.1 Audio Data

Using PRAAT tool we can analyze audio input pitch and intensity levels there graph are as follows

1.Intensity: For give input audio file by clicking on intensity->show intensity, it displays variation of intensity from start to end of the audio file in dB.To get the Intensity list of the audio file we have to select intensity->intensity listing.



2.Pitch: For give input audio file by clicking on pitch->show pitch, it displays pitch levels variation from start to end of the audio file in Hz. To get the pitch list of the audio file we have to select pitch->pitch listing.



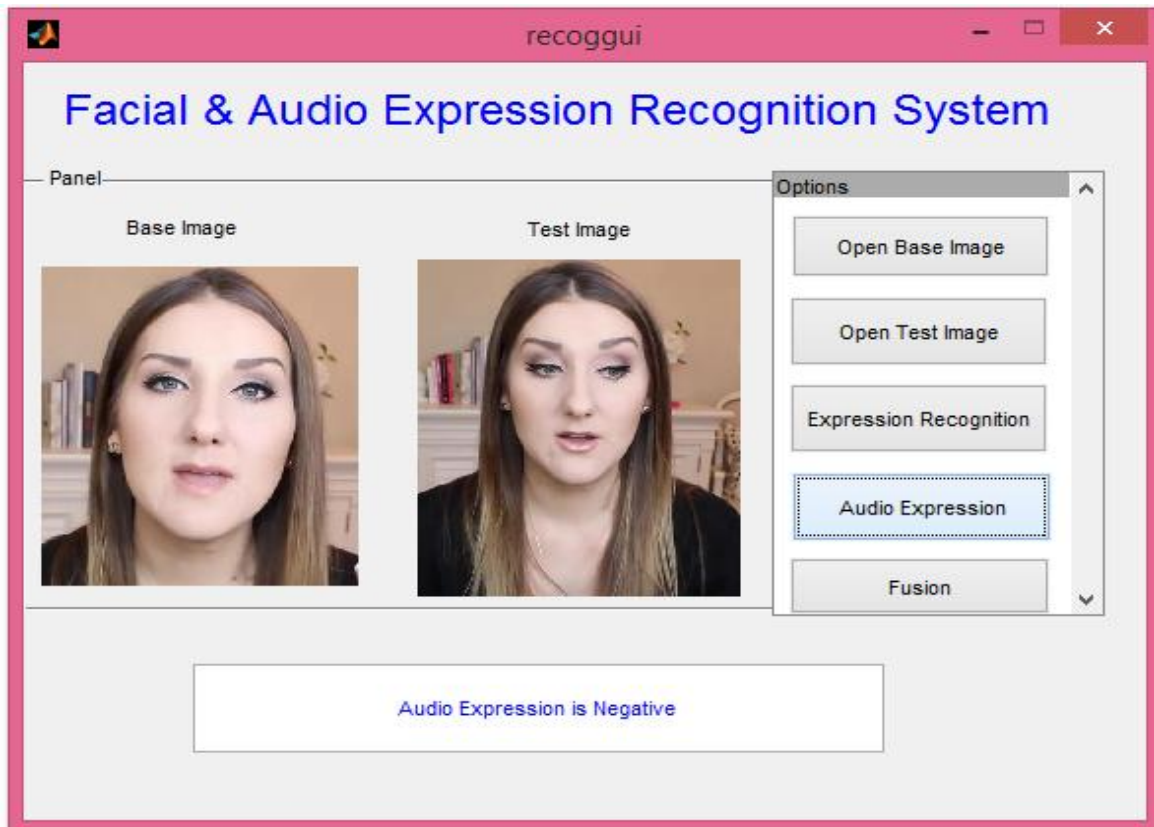
3. Data Listing:

Intensity listing and pitch listing of audio review file is stored in xls sheet as shown below. This data is given as input to Audioreview.m file to define the polarity of audio data.

1	Intensity_dB	pitch_Hz
2	-58.019665	205.6829
3	8.058885	176.5756
4	25.464599	278.0282
5	28.045637	200.7687
6	43.337303	192.3444
7	57.909858	179.1834
8	68.608204	414.4837
9	75.34309	134.3512
10	79.284777	140.1309
11	81.287055	138.9117
12	81.70938	155.4558
13	80.840343	0
14	79.248957	150.0831
15	77.862281	0
16	76.738276	272.325
17	74.409742	121.8841
18	70.571548	0
19	67.629874	155.8273
20	64.984136	0
21	65.107922	0
22	72.227989	310.3285
23	77.519375	327.1309
24	80.389643	0
25	81.557791	0

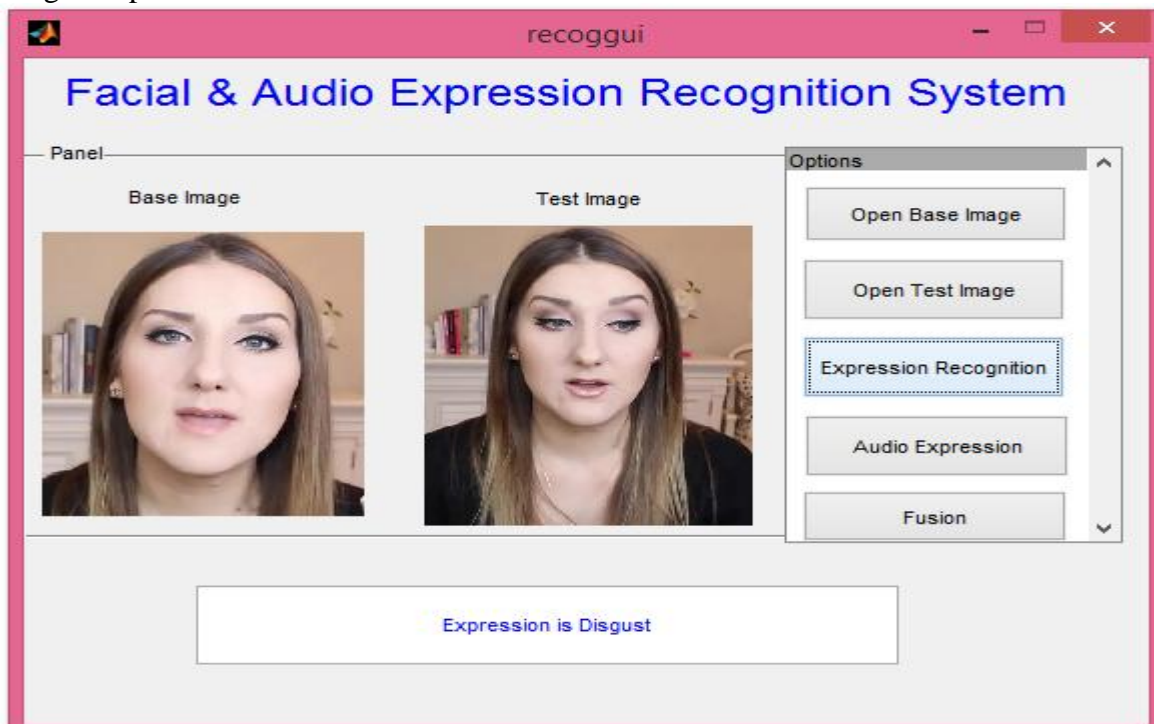
4.Audio Output:

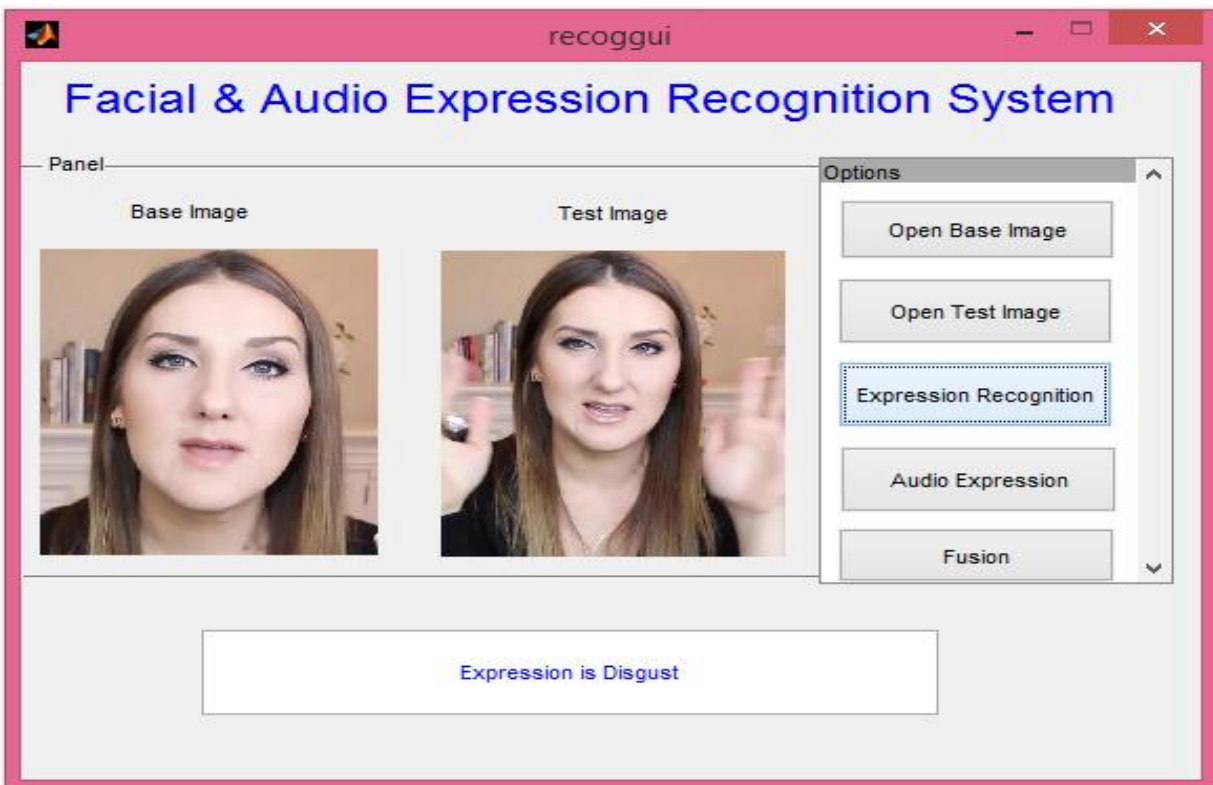
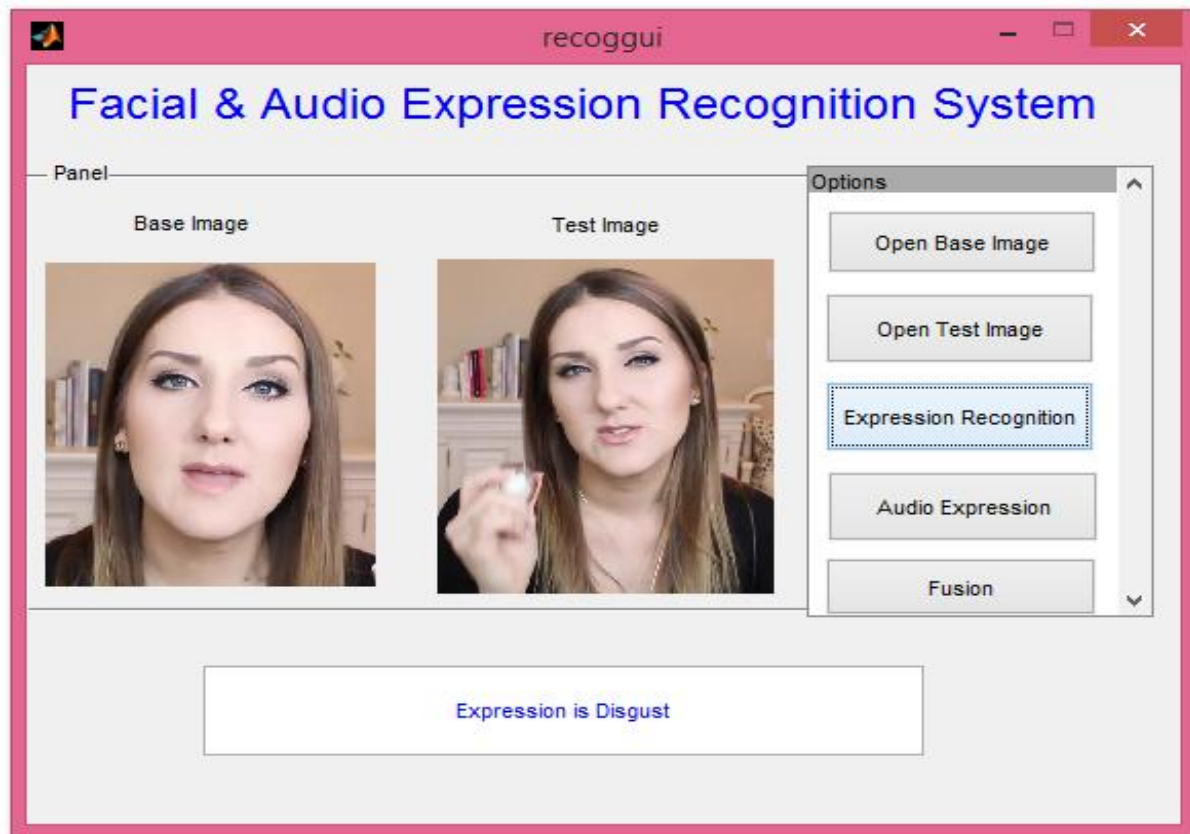
To know the polarity of corresponding video files audio output click on Audio Expression button and select the review .xls data of particular video then the data get loaded in matlab and audioreview.m file calculate the pitch, intensity, mean intensity, number of pauses in data and output the sentiment of audio review in GUI.



7.2.2 Video Data:

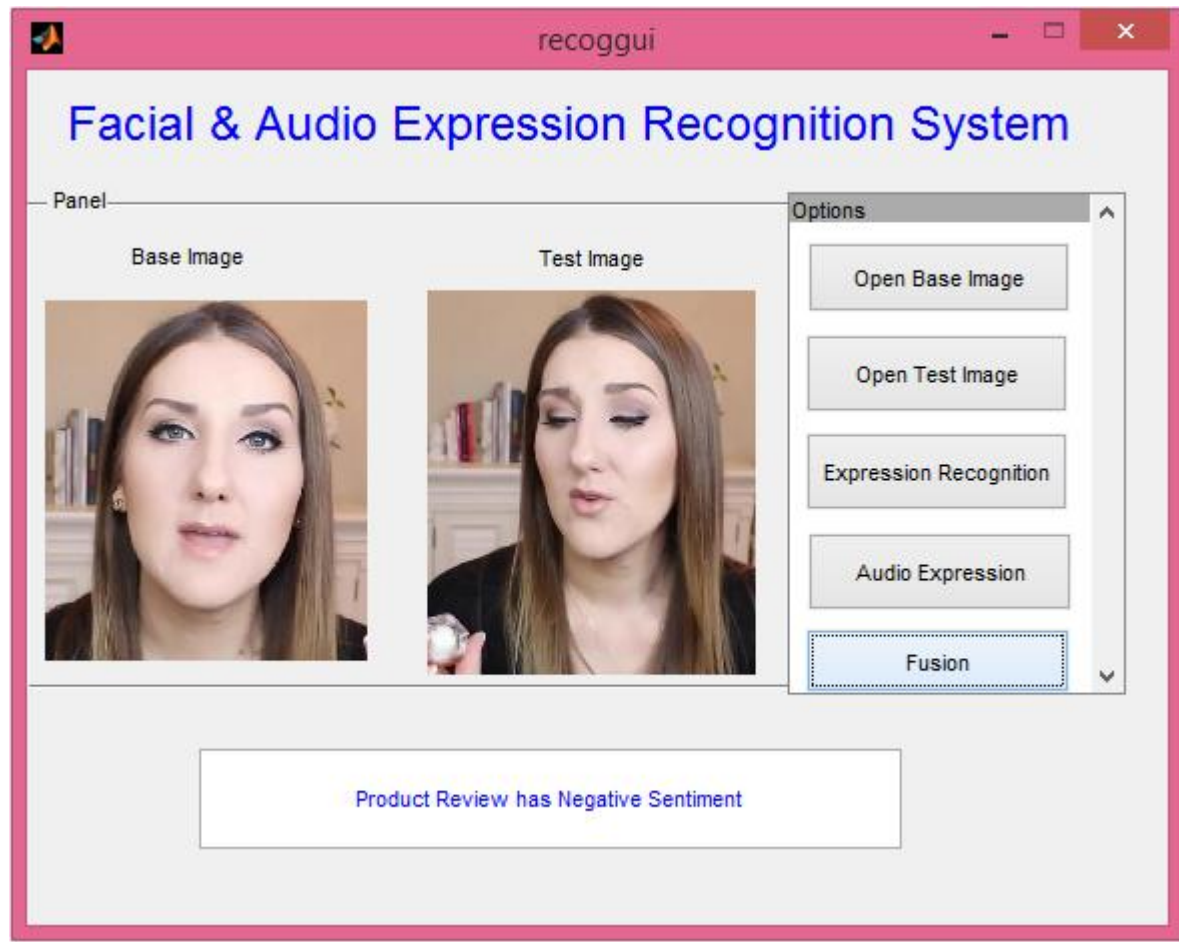
By selecting all the images in corresponding review folder we can get expression of each image and polarity can be assigned based on mostly occurring expression. Here the all images of particular review are selected as follows





7.2.3 Product Review Polarity:

Audio data and video data is fused to determine the product review polarity as positive, negative or neutral. After giving audio input and video input by clicking fusion button then it display the product review polarity in GUI



8. CONCLUSION AND FUTURE WORK:

In this paper, we addressed the task of multimodal sentiment analysis, and explored the joint use of both audio and visual modalities for the purpose of classifying of opinions in online videos. Through experiments performed on Youtube dataset, consisting of product review videos, where people express their opinions about different topics, we showed that the integration of audio, and visual features can improve significantly over the individual use of one modality at a time.

By this Multimodal sentiment analysis using audio visual format we may attain accuracy upto 80% when compared to traditional text sentimental analysis.

In our project for analyzing visual data we extract the frames from video and compare the neutral image with other images to recognize the expression. As a future work involves exploration of our proposed approach for larger videos, including more feature points in order to identify other known emotions. Automatic identification of several other emotions is challenging. Fusion process can also be improved by compiling several new facial emotions with voice features.

REFERENCES:

Research Papers

Sumit K Yadav , MayankBhushan and SwathiGupta's "Multimodal Sentiment Analysis using Audio Visual Format"

SoujanyaPoria, Erik Cambria and Newton Howard's "Fusing audio,visual and textual clues for sentimental analysis from multimodal content"

Veronica Perez Rosas,RadhaMihalcea, Louis-Philippe Morency's "Multimodal Sentimental Analysis of Spanish Online Videos"

NavleenKaur and MadhuBahl's "Emotions Extraction In Color Images Using Hybrid Gaussian AndBeizer Curve Approach"

Amrl EI Magrabhy and Othman Enany's "Detect and Analyze Face Parts Information Using Viola-Jones and Geometric Approaches."

LINKS:

www.mathworks.m

www.stackoverflow.com

www.quora.com