# Development of Real Time Facial Emotional Video Analysis with Speech Data Fusion using Random Forest Algorithm

Anusha A U[1], Anusha C[2], Bhumika S[3], Harshini S[4],
Students Mahanthesha U[5], Faculty, Dept. of ECE, GSSSIETW, Mysuru
GSSS Institute of Engineering and Technology for Women, Mysuru

***Abstract*— The main intention of this project is to represent the real time facial emotional analysis with the help of speech dataset and video dataset. The utilizers of the detection system is the multi-modal fusion and also uses various timescale feature of speech. The analysis of audio signals can be done by using many ways such that finding the amplitude & also detects the maximum peaks of the signals in our paper work we use audio, video dataset and find the human emotions like happy, sad, clam, angry, fear, disgust and surprise, with the help of machine learning algorithm and HAR emotion detection can be achieved. Pre-recorded and real time emotions can be analyzed and detected. And also produces the conclusion whether the frame has speech or non- speech.**

**The proposed work deals with the novel technique of data fusion for real time heterogeneous background facial emotional signal analysis with speech signal using the most popular machine learning algorithm called random forest and multi scale wavelet transform techniques. This concept implemented using MATLAB 2020 simulation Software for better outcomes.**

***Keywords: Face detection, Random forest Algorithm, Video surveillance system***

## I. INTRODUCTION

Nowadays, Humans communicate with the help of their emotions, Human emotions can be considered universal language. We know that an emotion plays a major role in a Human life. Facial expressions play a vital role while interacting socially and which is responsible for conveying human feelings and their emotions. Which was difficult to obtain in the early days but now many ways are there to find Human emotions using Machine Learning, Deep learning etc. Human emotions gives information about Humans mind set whether he/she is sad angry fear etc. The main goal about our project is to find the study about the Human emotions which should be recognized by the computer so the main intention is to build better interaction between Human and computer and to a build the machine that serves better for users' which is actually needed more in natural and in an effective way. Thus, human emotions recognition works are gaining a lot attention and interest from all groups especially from engineers. Emotion recognition technologies will benefit users in human-to-computer and human-to-human dialogue systems by improving services and adapting to human emotions. There is an increasing need for real-time emotion detection systems. Many studies are being conducted on emotion recognition via speech, which has resulted in numerous advancements in terms of various aspects as well as machine learning algorithms. Many studies focus on offline systems that employ auditory information on the phone or voice level.

In order to understand the offline emotion recognition research, we used human audio and video datasets in the real-time system. Human activities are divided into four categories: action, gesture, interactions, and group activities. In our project, we provide a real-time emotion categorization system that uses audio and video data sources without any additional information and provides speed and accuracy while meeting the objectives. Furthermore in future work the architecture contains all information in it which needs to be considered. In our work we extract the data's like audio and video then check the emotions and also our main intention to check it with real time action with given dataset. The major goal is to use video analysis and voice data at various levels of speech, such as supra-frame, intra-frame, and lexical level, to detect various human emotions, and the timescale feature is complementary. There has been a lot of proposed work on detecting human emotions and classifying them at various levels of extracted feature extraction. Considering the intra-frame features, Kwon et al found the human emotion recognition with the help of MFCC (Mel Frequency Cepstral Coefficients) and also with the help of various learning algorithms. The goal of our project is to detect few human emotions they are happy, sad, angry, fear, clam, surprise, disgust. Which will help in many ways in many fields.

**Aim:** Analyzing the performances of fusion of two activities which includes audio and video signals of person are mapping and testing for HAR system.

**Objectives:**

- Analysis of Human Action Recognition paves a way to develop a video analytics system which helps to recognize the Human facial expression.
- The objective of the project is to recognize the Real Time Human Facial Expressions such as: happy, anger, fear, disgust, etc., by extracting required features.
- Facial expression recognition has connections in many applications like for example sociology, security systems, surveillance environments, entertainment environments, and healthcare systems, human computer interaction.

## II. PROPOSED WORK

The system includes four major parts, which include speech acquisition, feature extraction, machine learning, and information fusion. Figure.1 describes the basic concepts of the system.

Videos contain both audio signals and images. The segregation of the frames and audios can be useful in many processes. Speech detection can be done on videos by analyzing the audio signals from them. The analysis of the audio signals can be done in many ways, such as finding the amplitude and detecting the maximum peaks in the signals. Signal analysis may be performed using the feature-based approach. The feature-based approaches can be more reliable for signal analysis since they are unaffected by other disturbances. The method to identify the spoken word in the video is done using the method of feature extraction. The feature extraction method computes sound signal values to detect the voice activity in the frames. The frames in which the speech is detected are identified. The variation between the values of the characteristics retrieved from the frames. The energy and the zero-crossing rate are obtained as features. The variation between the energy of the two frames will be nil if no spoken word is detected.

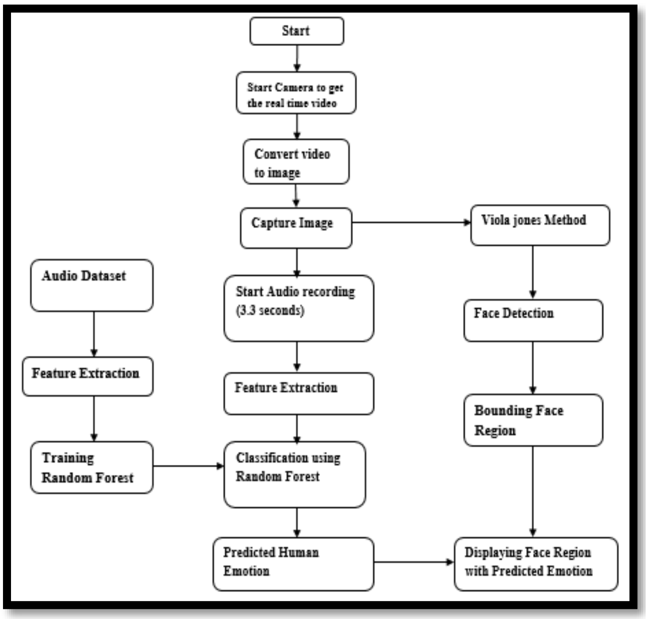


Fig 1: System Architecture

The video frames at which speech has occurred can be recognised and based on that classification can be finished. The input video is loaded. The video is converted into frames and voice signals. The separated frames were then pre-processed. In the pre-processing step, basic resizing operations and filtering processes are applied. The filtering procedure identifies the noise and substitutes the pixels with alternative pixels. The non-linear digital filtering technique recognises noisy pixels and replaces the noisy pixels identified using the midpoint of nearby pixels. The procedure is executed on the entire image in a specified window size. The face and mouth were sensed from the frames depend on the Viola-Jones detection method. Characteristics have been derived from the voice signals. Energy and ZCR was derived as features. Signal energy was commonly identified by identifying the signal length and signal amplitudes. The signal amplitude relates to the signal's value. The ZCR is the zone of the signal where the amplitude decreases. The energy and the ZCR were not affected by any other outer parameters and hence are more reliable. The variation between the derived features from each frame was calculated. The variation within the frames that are having voice activity will be one and the variation within the frames that are not having voice activity will be zero. A binary decision tree produces the decision of whether the frame has speech or non-speech. The extracted decision is useful for the recognition of the persons, Categorizing of the frames, and so on. The suggested method has better accuracy when equated to the other methods since the energy and zero-crossing rate is not affected by other external parameters.

## RANDOM FOREST ALGORITHM

Random forest is a prominent supervised learning technique in machine learning. The term "forest" refers to the process of assembling decision trees, which are typically trained using the "bagging" method. The “bagging method” is a collection of learning models that work together to improve the overall outcome. The random forest decision tree's classifiers are depicted in Figure 2.

One of the most significant benefits of random forest is that it can be used to solve classification and regression issues, and it is one of the most popular machine learning algorithms today. Let's take a look at random forest categorization, which is also referred to as the building block of machine learning. A decision tree or a bagging classifier are both examples of random forests. Fortunately, we don't need to mix a decision tree and a bagging classifier because we can just utilize the random forest classifier-class. It is simple to deal with regression with the help of random forest by employing regressor algorithms, as indicated in the flow chart in Figure 3.

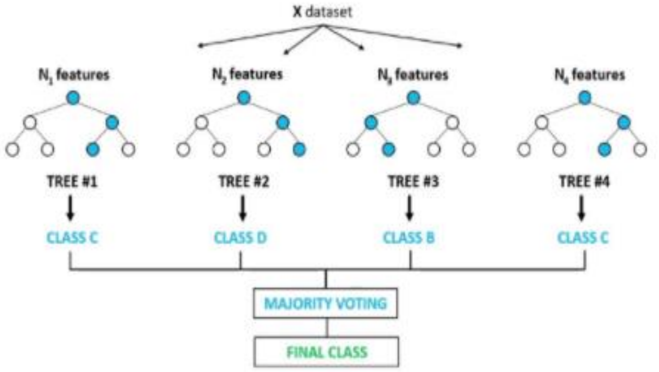


Fig2: Random Forest Classifiers Decision Tree

The following steps and diagram depict the working process:

**Step 1**: Pick K data points at random from the training set.

**Step 2:** Create decision trees based on the data points you've chosen.

**Step 3:** Choose an N for the number of decision trees we wish to build.

**Step 4:** Repeat the first and second steps.

**Step 5:** Find the forecasts of each decision tree for new data points, and allocate the new data points to the category that receives the majority vote.

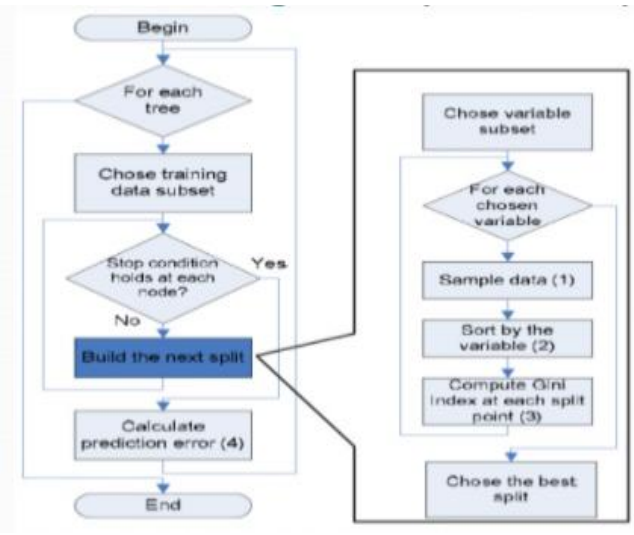


Fig3. Flow chart illustrating Random forest algorithm

The entropy of Random forest algorithm is calculated by the following expression

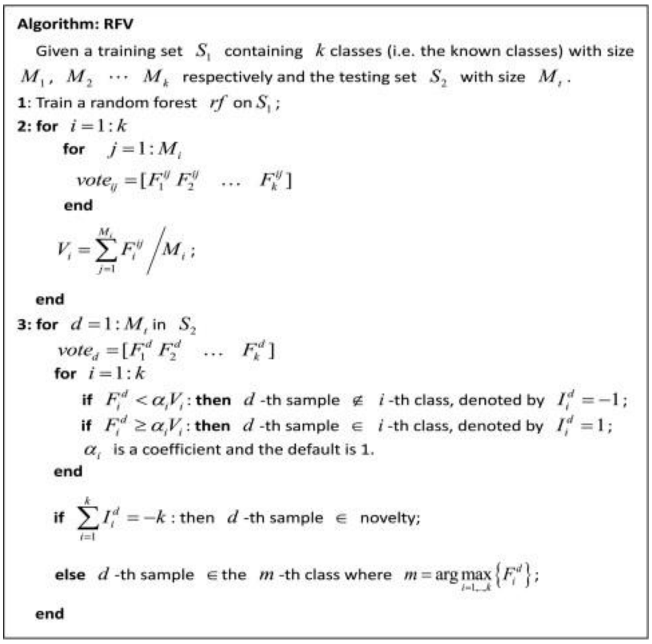
$$Entropy = \sum_{i=1}^{C} - p_i * \log_2 (p_i) \qquad \text{----------- 1}$$

The Mean Square Error for Random forest Algorithm is given by

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2 \qquad \text{------------2}$$

Where N= Number of Data Points, fi= Returned value, yi= Actual value.

The mathematical description of random forest is explained below



```
Algorithm: RFV
  Given a training set $S_1$ containing $k$ classes (i.e. the known classes) with size
$M_1$, $M_2$ $\cdots$ $M_k$ respectively and the testing set $S_2$ with size $M_t$ .
1: Train a random forest $rf$ on $S_1$ ;
2: for $i = 1 : k$
      for $j = 1 : M_i$
        $vote_{ij} = [F_1^{ij}\ F_2^{ij}\ \ \dots\ \ F_k^{ij}]$
      end
      $V_i = \sum_{j=1}^{M_i} F_i^{ij} \Big/ M_i$ ;
   end
3: for $d = 1 : M_t$ in $S_2$
      $vote_d = [F_1^d\ F_2^d\ \ \dots\ \ F_k^d]$
      for $i = 1 : k$
        if $F_i^d < \alpha_i V_i$ : then $d$ -th sample $\notin$ $i$ -th class, denoted by $I_i^d = -1$ ;
        if $F_i^d \geq \alpha_i V_i$ : then $d$ -th sample $\in$ $i$ -th class, denoted by $I_i^d = 1$ ;
        $\alpha_i$ is a coefficient and the default is 1.
      end
      if $\sum_{i=1}^{k} I_i^d = -k$ : then $d$ -th sample $\in$ novelty;
      else $d$ -th sample $\in$ the $m$ -th class where $m = \arg\max_{i=1,\dots,k} \{F_i^d\}$ ;
   end
```

III. FEATURE EXTRACTION:

### 1) MEL-FREQUENCY CEPSTRUM:

In speech signal processing, the mel-frequency cepstrum (MFC), it is the representation of a short-term power spectrum of a sound based on a linear cosine transform of a logorithmic power spectrum on a nonlinear mel-scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that are accumulatively form the MFC. They are obtained from some sort of cepstral representation of an audio clip (a nonlinear "spectrum-of-a-spectrum"). The main difference between the cepstrum and the mel-frequency cepstrum i.e, in the MFC, the frequency bands which are equally spaced in the mel-scale,which estimates the human auditory system's response more closely than that of the linearly-spaced frequency bands which is used in the normal cepstrum. This frequency warping can be allowed for better representation of sound, like for example: an audio compression.

MFCCs are usually explained as shown below:

1. Firstly to take Fourier transform of input signal.
2. Create the triangular / overlapping window by mapping the spectral power on to a mel scale
3. To take the logarithmic of power at each stages of mel frequencies.
4. Represent the discrete cosine transformation of the input signal using the mel log power.
5. The resulting spectrum is obtained by magnitude of MFCCs.

### 2) ZERO CROSSING TECHNIQUE:

The zero crossing technique is a slite where waveform will cross over a zero level axis. When we will implement any editing actions, like cutting, pasting, or dragging, ensure that material is included at zero crossing which is displayed in Figure4.

If we do not perform the operations at zero crossing technique properly the result can have discontinuities in wave, which will be recognized as clicks / pops in the sound.

To turn-on zero crossing on Edit Tab of an Audio Editor just to ensure that selections that we made will be adjust every time so that they will always begin and closed at nearest zero crossing.

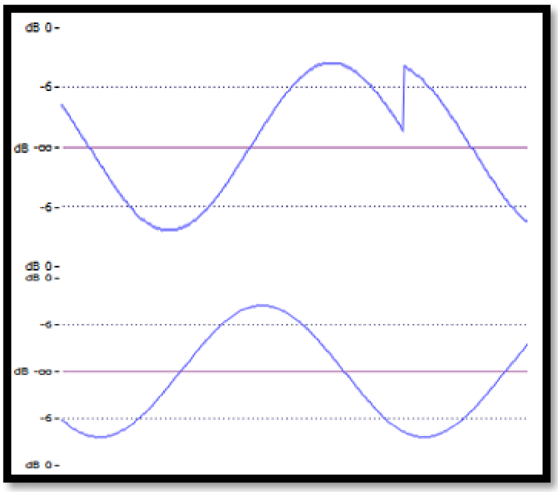


**Fig4: Zero Crossing Representation**

### 3) MULTISCALE WAVELET TRANSFORM FEATURES

Energy, Variance, Standard Deviation and Waveform Length are the feature extraction algorithm which are sane for any type of signals. Although this was mostly developed for the myoelectric, EMG and the signal feature extraction for prostheses control. The algorithms employs the coefficients of wavelet transformation to extract few features which including:

- One energy (E) feature
- Two standard and variance features
- One waveform length feature
- One entropy feature

These are needed to specify the window size, window increment, sampling frequency and so on.

## ADVANTAGES / APPLICATIONS

The Real-Time human facial expression detection with speech data fusion concept will be used in the different fields of study mainly in Sociology, Security systems, Surveillance environments, Human computer interaction, Entertainment environments, Healthcare systems etc.,.

## IV. RESULTS

The following are the results obtained from the novel technique of data fusion for real time heterogeneous background facial emotional signal analysis with speech signal using the most popular machine learning algorithm called random forest and multi scale wavelet transform techniques. A square box indicated using yellow color which illustrate the detected face and on top of that letters indicate the facial emotions of the person shown from Figure.5 to Figure.11.

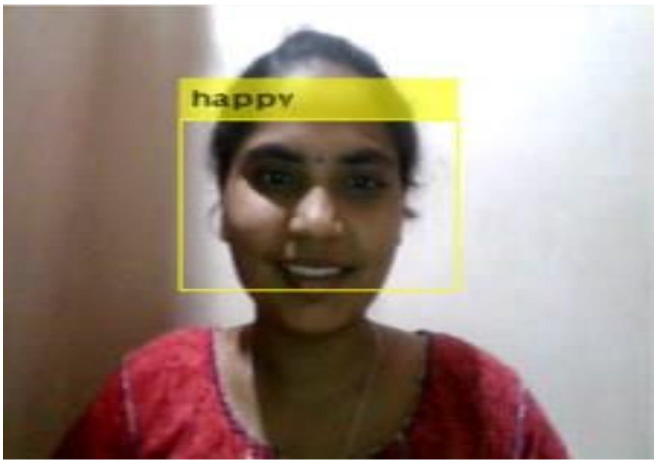


Fig 5. Illustration of Real Time Facial Emotion “Happy”

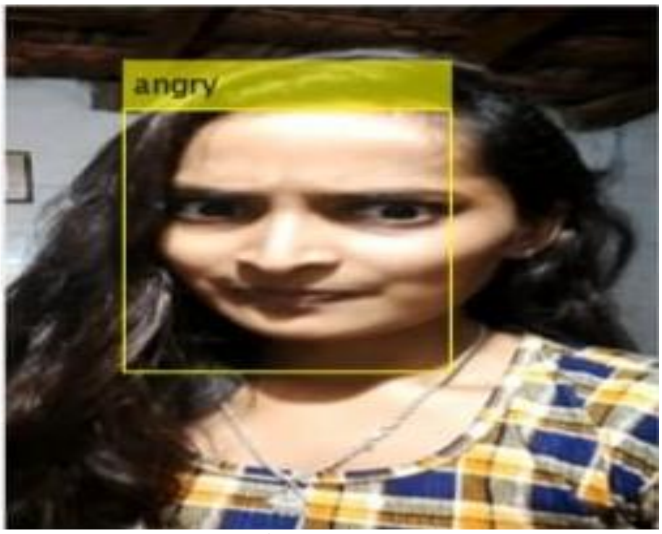


Fig 6. Illustration of Real Time Facial Emotion “Angry”

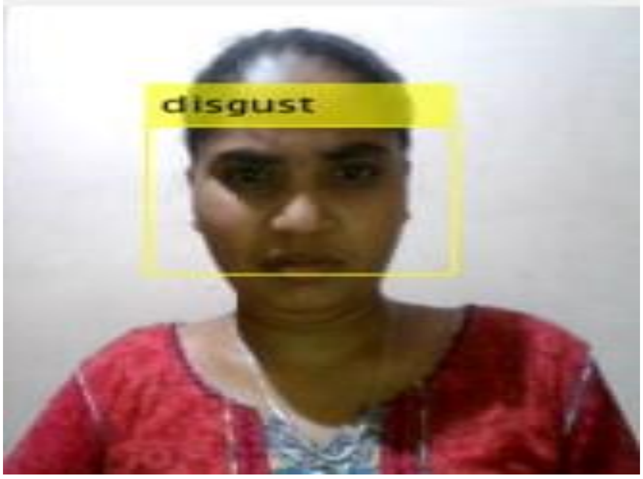


Fig 7. Illustration of Real Time Facial Emotion “Disgust”

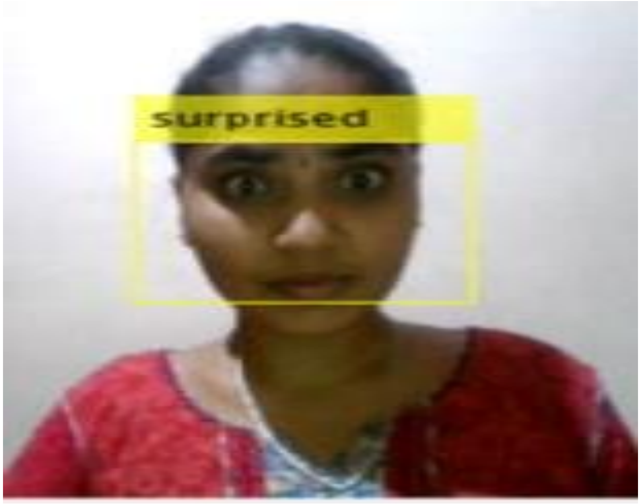


Fig 8. Illustration of Real Time Facial Emotion “Surprise”

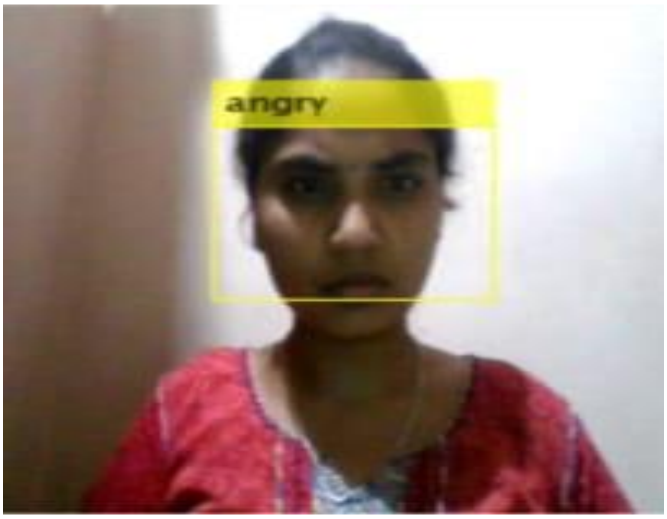


Fig 9. Illustration of Real Time Facial Emotion “Angry”

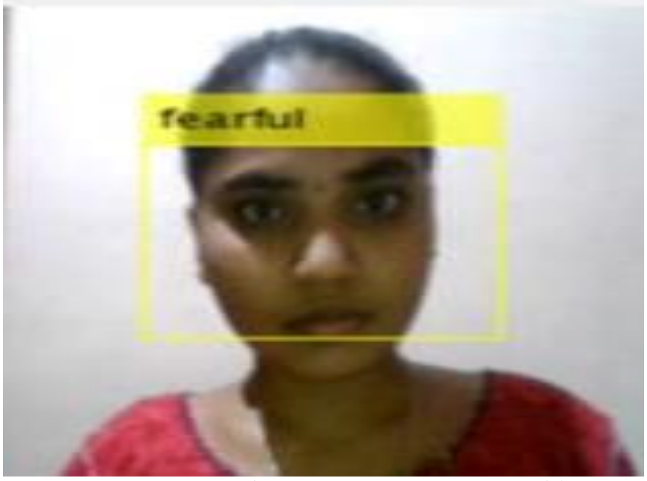


Fig 10. Illustration of Real Time Facial Emotion “Fearful”

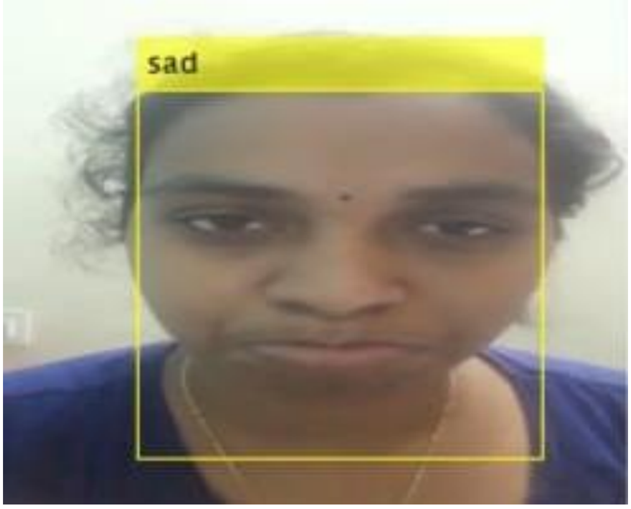


**Fig 11. Illustration of Real Time Facial Emotion “Sad”**

## IV. CONCLUSION

This project described the advance image and speech processing parameters and observed the results obtained from data fusion technique for real time heterogeneous background human facial emotional signal with speech signal using the robust machine learning algorithm called random forest and multi scale wavelet transform techniques using the sophisticated advanced simulation MATLAB tool.

## V. FUTURE WORK

The future applications of real time human action recognition system using artificial intelligence and machine learning techniques can also be made functional using novel concept data fusion techniques. In this study the fusion of multiple actions of a single person or multiple human actions can be detected and recognized to obtain the best output performance to the given input videos.

## ACKNOWLEDGMENT

The author’s likes to convey gratitude to everyone who is directly involved or indirectly contributed, suggested and coordinated in raw information collection and algorithm implementation for the publication of this article.

## REFERENCES

[1] Dr. H S Mohan and Mahanthesha U, “Human action Recognition using STIP Techniques”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-7, May 2020

[2] J. F. Allen, “Maintaining knowledge about temporal intervals,” Commun. ACM, vol. 26, no. 11, pp. 832–843, Nov. 1983.

[3] C. Fernandez, P. Baiget, X. Roca, and J. Gonzalez, “Interpretation of complex situations in a semantic-based surveillance framework,” Image Commun., vol. 23, no. 7, pp. 554–569, Aug. 2008.

[4] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” IEEE Trans. Intell. Transp. Syst., vol. 11, no. 1, pp. 206–224, Mar. 2010.

[5] Y. Changjiang, R. Duraiswami, and L. Davis, “Fast multiple object tracking via a hierarchical particle filter,” in Proc. 10th IEEE ICCV, 2005, vol. 1, pp. 212–219.

[6] A. Loza, W. Fanglin, Y. Jie, and L. Mihaylova, “Video object tracking with differential Structural SIMilarity index,” in Proc. IEEE ICASSP, 2011, pp. 1405–1408.

[7] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 5, pp. 564–577, May 2003.

[8] V. Papadourakis and A. Argyros, “Multiple objects tracking in the presence of long-term occlusions,” Comput. Vis. Image Underst., vol. 114, no. 7, pp. 835–846, Jul. 2010.

[9] Mahanthesh U, Dr. H S Mohana “Identification of Human Facial Expression Signal Classification Using Spatial Temporal Algorithm” International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016.

[10] NikiEfthymiou, Petros Koutras, Panagiotis, Paraskevas, Filntisis, Gerasimos Potamianos, Petros Maragos “Multi-View Fusion for Action Recognition in Child-Robot Interaction”: 978-1-4799-7061-2/18/$31.00 ©2018 IEEE.

[11] Nweke Henry Friday, Ghulam Mujtaba, Mohammed Ali Al-garadi, Uzoma Rita Alo, analysed “Deep Learning Fusion Conceptual Frameworks for Complex Human Activity Recognition Using Mobile and Wearable Sensors”: 978-1-5386-1370-2/18/$31.00 ©2018 IEEE.

[12] Van-Minh Khong, Thanh-Hai Tran, ”Improving human action recognition with two-stream 3D convolutional neural network”, 978-1-5386-4180-4/18/$31.00 ©2018 IEEE.

[13] Nour El Din Elmadany , Student Member, IEEE, Yifeng He, Member, IEEE, and Ling Guan, Fellow, IEEE ,”Information Fusion for Human Action Recognition via Biset /Multiset Globality Locality Preserving Canonical Correlation Analysis”

IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 27, NO. 11, NOVEMBER 2018.

[14] Pavithra S, Mahanthesh U, Stafford Michahial, Dr. M Shivakumar, “Human Motion Detection and Tracking for Real-Time Security System”, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016.

[15] Lalitha. K, Deepika T V, Sowjanya M N, Stafford Michahial, “Human Identification Based On Iris Recognition Using Support Vector Machines”, International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE) Vol 2, Issue 5, May 2016.

[16] RoozbehJafari, Nasser Kehtarnavaz “A survey of depth and inertial sensor fusion for human action recognition”, https://link.springer.com/article/10.1007/s11042-015-3177-1, 07/12/2018.

[17] Rawya Al-Akam and Dietrich Paulus, ”Local Feature Extraction from RGB and Depth Videos for Human Action Recognition”, International Journal of Machine Learning and Computing, Vol. 8, No. 3, June 2018.