Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing

Humaid Alshamsi [1] , Veton Kepuska [1] and Hazza Alshamsi [1]
Department of Electrical and Computer Engineering
Florida Institute of Technology
Melbourne, USA

Hongying Meng [1,2]
Department of Electrical and Computer Engineering
Brunel University London
UB8 3PH, UK

***Abstract*— Because emotions are such a significant and integral part of being human, understanding them and knowing how to react to the emotions of others is a fundamental requirement for successful social interaction. We recognize emotions primarily through speech and facial expression. The topic is gaining importance in academic research triggered by research in new techniques such as identifying emotions based on speech context. This investigates the relationship between emotions and the content of our speech. This paper proposes how emotion in speech and facial expression can be recognized in real time using a framework consisting of mobile phone technology backed by cloud computing. This functionality was developed and built into a mobile phone application. Currently, the application works on any Android smartphone to detect and identify emotions in real time. The results are expressed as a percentage of all the possible emotions, such as sad, happy, fear, surprise, anger and so on. The results of the experiment confirm that face and speech emotion recognition was conducted successfully using a smartphone. It was correct in 97.26% of instances when used with standard corpora:**

a. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

b. the Surrey Audio-Visual Expressed Emotion (SAVEE)

Keywords: Speech Recognition; Facial Expression Recognition; Mobile Computing; Support Vector Machine; Computer Vision;

I. INTRODUCTION

Automated emotion recognition is currently a subject of great interest because it has many real-life applications in many areas ranging from advertising [1] to helping to progress our understanding of mental health [2] and from analyzing and summarizing video [2] to removing facial expressions for normalization. While capability to achieve emotion, recognition is at quite an advanced stage under laboratory conditions, that is far from the case in real world scenarios with many issues yet to be resolved. The challenge of designing and developing a system for recognizing human emotions in speech and facial expressions using smartphone technology is primarily the need for a great number of suitable audio-visual clips for testing. In addition, there is the need for storage of that audio-visual corpora, and difficulties for the application in correctly assessing light conditions, head position and poses, and the wearing of masks and sunglasses. Researchers have put forward a range of methods in the field of computer recognition of emotions in speech and facial recognition systems, and these techniques focused mainly on feature descriptors and methods of classification for SER / FER.

By contrast, we propose the Emotii Android application in this paper. It is a framework for a speech and a facial expression recognition system that can identify emotion. Tests demonstrate that Emotii Speech and Facial Expression recognition delivers the capability that enables the smartphone to more accurately assess the emotional state of humans. It follows that proposed Emotii Speech and Facial Expression Recognition on smartphones helps to enhance the human-computer interface and interaction.

II. RELATED WORKS

There is a significant body of scientific literature that deals with emotion recognition. A large proportion of addresses potential options to define and represent emotions. If implementing discrete classes [4] such as happiness, sadness, anger, etc. is the most straightforward approach, then greater interest can be added by representing emotions as the degree of valence and arousal as suggested in [5]. For facial expression recognition, action units may be utilized that focus on movement of various areas of the face [6]. These two representations can be interlinked by mapping the discrete classes onto the arousal valence elements [7] and the action units can be used to deduce them [8].

Another essential section of literature addresses approaches to representing audio-visual content by features, which can be used subsequently by classifiers. A number of earlier studies utilize (i) manually created features such as Gabor features, Local Binary Patterns (LBP), modulation spectrum (ModSpec) or Enhanced AutoCorrelation (EAC) for audio, discrete cosine transforms for representing images, relative spectral transform - linear perceptual prediction (RASTA-PLP), Linear Predictive Coding coefficients (LPC), (ii) standard classifiers (e.g. KNN, SVN) for classification as detailed in [9]. [10] The winner of Emotion Recognition in The Wild Challenge 2015 showed just how relevant are Action Units in emotion recognition. Among the earliest proponents of learning the features was [11] in place of manually created descriptors, utilizing Deep Convolutional Networks. Since then, EmotiW'16 winner [12] put forward the

978-1-5386-7266-2/18/$31.00 ©2018 IEEE

C3D feature that represents faces in a spatiotemporal manner with great efficiency. The literature also looks at fusing the various modalities for recognizing emotions in audio-visual content. A modality can be considered as one of the signals that helps in the perception of emotion. Recent techniques for fusing multiple modalities are two-stream ConvNets [13], Multiple Kernel Fusion [14], or ModDrop [15]. The most commonly used modalities for this purpose are speech and facial images, although context appears to be extremely significant [16]. For example, a basic understanding of the scenario may assist in discriminating between two candidate classes even if it is merely based on relatively simple features that describe the entire image.

Because the majority of recent techniques for emotion recognition are supervised and therefore require training data, it is becoming more important to have such resources available. A number of challenges have resulted in useful data being collected.

III. Facial expression and speech emotion recognition

The pre-processing phase of our research detected and tracked movements in the face using the Hausdorff Distance detection method [17]. We then used Facial Landmarks when assessing facial expression, and the Mel-Frequency Cepstral coefficients method for speech, to extract features. The core approach consisted of extraction and classification of facial features. Using an SVM-based algorithm, we analyzed the histogram values as opposed to the stored values in the training set. A very popular database of video clips of facial expressions is RAVDESS (the Extended Ryerson Audio-Visual Database of Emotional Speech and Song) [18]. We used both that and SAVEE (Surrey Audio-Visual Expressed Emotion Database) [19]. Action Units (AUs) were triggered to best characterize the most common facial expressions in the clips.

Feature Descriptors: The HOG descriptor and BRIEF descriptor [20] have been examined for use in FER as well [21], [22], [23]. The Gabor descriptor [24] is also another common feature descriptor used for FER [25], [26], [27], [28], [29]. The Haar descriptor [30] is one of the most popular descriptors for use in FER [31], [32].

In our research, we assessed the facial landmarks descriptor [17], which is fairly new features descriptor in facial expression recognition. As far as we know, we were the first to utilize this facial landmark descriptor in developing a real time mobile phone application for FER, and also Mel-Frequency Cepstral coefficients technique in SER.

Classification Methods: The most common classifier used in FER is the SVM classifier [27], [28] [33], [34], [35]. A number of studies reported that SVM in conjunction with a linear kernel delivered similar test results compared with a radial basis function (RBF) kernel [27], [36], [28]. The K-Nearest Neighbors (K-NN) classifier is another that has also been used recently for FER [37]. To date, their application has delivered very accurate results with a few errors. The application is developed on an Android platform and can run on any Android device. The databases are used in system testing.

A. System Overview

Figure 1 illustrates a block diagram of automated speech emotion recognition and facial expression running on a smartphone utilizing cloud computing. The principal phases are (i) feature abstraction and (ii) feature classification. The feature abstraction phase stage requires pre-processing phases, such as obtaining a sequence of images (at 8 frames per second) utilizing a video camera, identifying the facial regions of the images, and standardizing the image's lighting properties. This application utilizes Mel-Frequency Cepstral Coefficient (MFCC) to identify speech emotion and Facial Landmarks for facial expressions to process feature abstraction.

A support vector machines (SVM) algorithm is involved in the feature classification process and is generally accepted as being an effective nonparametric pattern classification technique in image processing. After the emotion has been aggregated, it is transmitted to the cloud via Globalscape (CuteFTP) where the result is established. The application to perform this work was developed using an Android studio tool, which is user friendly and runs on any Android phone. The code tracks facial features and recognizes emotion in speech. It can also detect when the subject is wearing glasses.

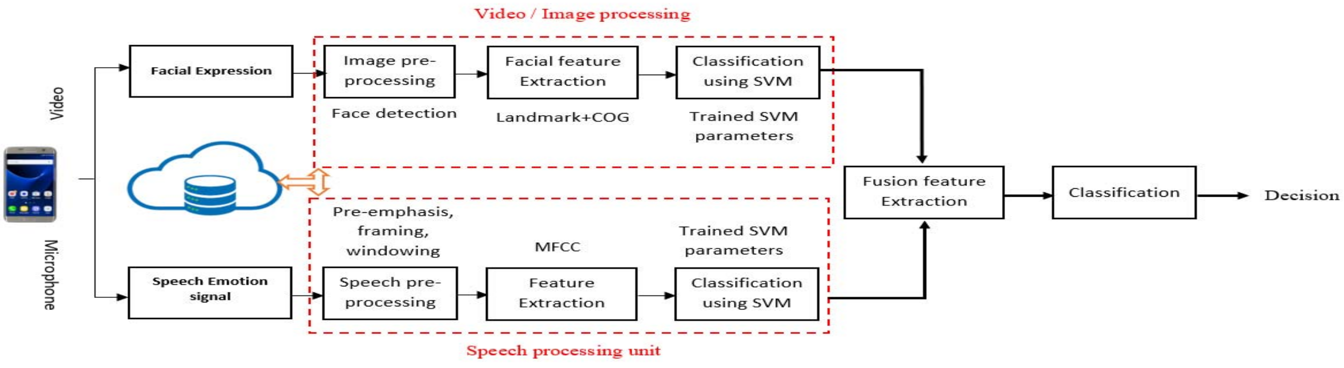


Fig. 1. Over View Structure of The Real Time Facial Expression and Speech Emotion Recognition System.

B. Extracting Features from Facial Expression

The locations of the mouth, eyebrows, eyes, nose etc. are indicated by points on certain areas of the image called facial landmarks. These points are tracked in time to follow the movements of facial muscles. If we consider all facial landmarks as a connected graph, the assumption is that the graph's density is different for each facial expression; we use the already trained facial landmark finder to estimate locations of 68 (x, y) coordinates used to map facial structures to the face. Having detected the facial landmarks, we calculate the "Center of Gravity" (COG) of all the landmarks (Formula [1,2]) as illustrated in Fig.2 (a & b). Being useful mathematical tools, graphs provide a great deal of information about the interrelationships of spatial points which are facial landmarks in this case. Using spectral graph analysis, we can obtain features from these landmarks and extract a characteristic vector that depicts a graph's areas of density. The tip of the nose is taken to calculate the face offset correction by finding the angle of the nose as illustrated in Fig.2 (c) (formals from [3-8]).

$$X_{COG} = \frac{1}{68}\sum_{i=1}^{68} x_i \tag{1}$$

$$Y_{COG} = \frac{1}{68}\sum_{i=1}^{68} y_i \tag{2}$$

Where X_{COG} is the x-coordinate of COG and Y_{COG} is the y-coordinate of COG.

$$x_{relative_i} = x_i - X_{COG} \tag{3}$$

$$y_{relative_i} = y_i - Y_{COG} \tag{4}$$

$$EUC_i = \sqrt{(x_i - X_{COG})^2 + (y_i - Y_{COG})^2} \tag{5}$$

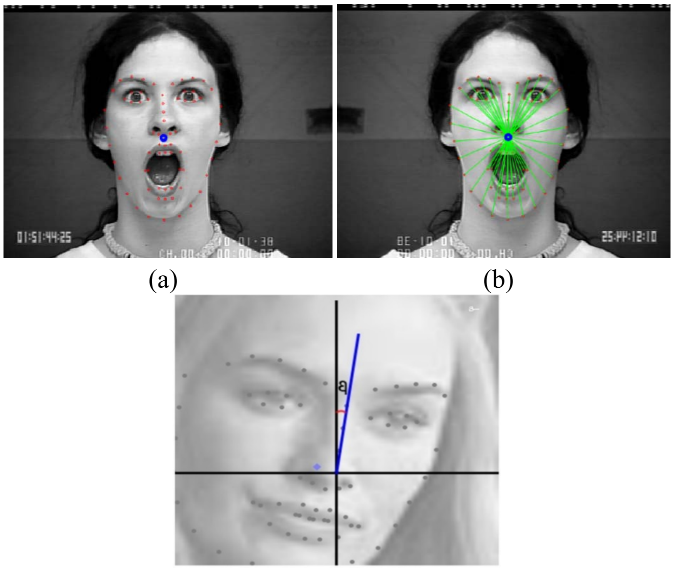
$$\beta_{nose} = \tan^{-1}\left(\frac{y_{28} - y_{31}}{x_{28} - x_{31}}\right) \tag{6}$$

$$\theta_i = \tan^{-1}\left(\frac{y_i - Y_{COG}}{x_i - X_{COG}}\right) - \beta_{nose} \tag{7}$$

where $i = 1,2,...., 68$

$$F = \{x_{relative_i}, y_{relative_i}, EUC_i, \theta_i\} i = 1^{68} \tag{8}$$

The Euclidian type of distance EUC_i is between the central point and point i and is essentially the angle of the nose using point-31 at the tip and point-28 at the top. The overall relative angle is defined by θ_i and is corrected when the face itself is not perfectly horizontal. The actual feature vector length is approximately 272.



(a) (b)

(c)

Fig. 2. (a) Facial Landmarks and COG using CK+ dataset (b) Line mapping between Facial Landmarks and COG. (c) Face offset correction.

C. Extracting Features from Speech Emotion

Selecting which features contain the most relevant emotion information from speech signals is one of the most important processes in the SER system. Multiple previous studies have shown that viable and valid parameters for detecting and differentiating specific emotions include speech energy, fundamental frequency, formant frequency and Mel-frequency cepstrum coefficients. The basis of extracting features rests with separating out speech into discrete smaller intervals or frames. Because of vocal fold tension, the pitch signal conveys much data regarding emotion. Vocal fold vibrations constitute the fundamental frequency in speech and the second most relevant characteristic is speech energy because differing emotions cause the speech signal energy to change. MFCC is the most commonly used of all the available spectral features in SERs [38], [39]. Its benefits include superior tolerance of noise, better capability for distinction, and simple calculation. In this study, Praat software [40] provides MFCC features to extract using a 20ms window length and a 10ms time interval. The Hamming window is generally favored for its excellent sidelobe suppression and high-frequency resolution. Because they contribute no useful information, silent areas in the audio files of the database were identified and removed by thresholding energy and measuring their zero-crossing rate. Hearing in humans is not on a linear scale and therefore MFCC uses the Mel scale [41]. This frequency scaling uses logarithmic intervals over 1000Hz, but linear spacing below. For any given frequency f Hz the formula below calculates the Mel frequency [42]:

$$Mel(f) = 2595 \times \log\left(1 + \frac{f}{700}\right) \tag{9}$$

Uniform filters in a triangular series that overlap constitute the Mel scale filter bank. Filters have a constant bandwidth of 100 with center frequencies set at 50. This mimic our understanding of the auditory system [43] and corresponds with the Mel frequency scale's spacing.

D. *Extracting Features from Multimodal Fusion*

Feature level of Multimodal fusion represents a method for combining all feature vectors. The overall modalities should match the modality of the various features that extracted the algorithms. Fig. 3 presents a block diagram representing feature level fusion of emotions contained in audio-visual data to achieve recognition.

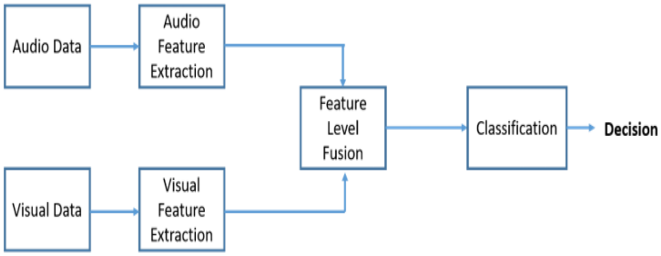


Fig. 3. Block diagram illustrating Audio-Visual Emotion Recognition through Feature-Level Multimodal Fusion.

Fusion at feature level concatenates the vectors from features of both audio and visual data to arrive at the feature vector. Fig. 4 illustrates the concatenation of the two feature vectors that result from feature level fusion.

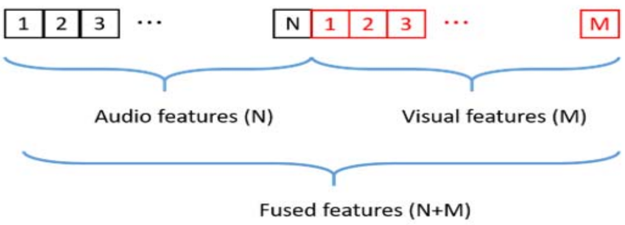


Fig. 4. Block diagram illustrating Feature-Level Multimodal Fusion.

$$f_a(i), \quad i = 1,2,\dots,N \tag{10}$$

$$f_a = [f_a^1, \; f_a^2, \; \cdots \; , f_a^N] \tag{11}$$

$$f_v(k), \quad k = 1,2,\dots,M \tag{12}$$

$$f_v = [f_v^1, \; f_v^2, \; \cdots \; , f_v^M] \tag{13}$$

$$f_f(j), \quad j = 1,2,\dots,N+M \tag{14}$$

$$f_f = [f_a^1, \; f_a^2, \; \cdots \; , f_a^N, \; f_v^1, \; f_v^2, \; \cdots \; , f_v^M] \tag{15}$$

f_a here represents the feature vector relating to audio data, which includes N features, f_v denoted by the visual data vector consisting of M features, and f_f being the fused feature vector. The length of the fused feature length equals the sum of the two original features (length (f_f) = N+M).

E. *Method of Classification*

Although somewhat simplistic, this algorithm formula is highly effective in classification and pattern recognition for machine learning. The SVM benefits from a superior level of excellent training data. It main purpose [44], [45] is utilizing a kernel function for transforming they inputs into a high dimensional feature space. This makes the input samples separable in a linear manner [46]. Fig. 2 shows its use with an optimal separation hyperplane.

The biggest advantage of an SVM is that it demonstrates excellent classification performance because it has limited training data. Classification of linearly separable data points is executed using this formula [47]:

$$\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1 \tag{16}$$

$$\langle w \cdot x \rangle + b_0 \geq -1, \forall y = -1 \tag{17}$$

Where (x, y) is the pair from the training set. Here, $x \in R^n$ and $y \in \{+1, -1\}$.

$\langle w \cdot x \rangle$ represents the inner product of w and x while b_0 signifies the bias condition.

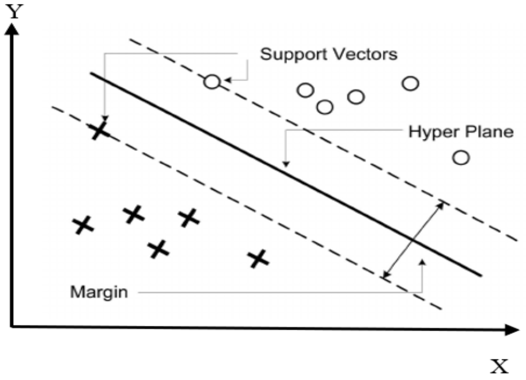


Fig. 5. Structure of Support Vector Machine.

This study employs a SVM that uses the Radial Basis Kernel (RBF) [24] function as well as a linear kernel function, which is defined by this formula:

$$\text{Kernel } (x, y) = (x \cdot y) \tag{18}$$

This formula gives the radial basis kernel function:

$$\text{Kernel } (x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} \tag{19}$$

IV. System Evaluation

The proposed system was initially evaluated against a public dataset to demonstrate its efficiency for recognition of speech emotion and facial expression.

A. Speech Emotion and Facial Expression Datasets

Two facial expression datasets are utilized by this study: The Extended Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18] and the Surrey Audio-Visual Expressed Emotion Database (SAVEE) [19]. Both contain facial and speech data. The Computer Vision system supplied the facial elements using a facial detector, while speech elements were generated from an Android library by speech recognition.

1. RAVDESS Dataset

The RAVDESS dataset was used by research in [18] for comparing differences and similarities in audio and visual signals in speaking and singing with emotion.

The dataset contains audio-visual clips from 24 performers (12 male, 12 female) who speak and sing the same two sentences using a variety of emotions using normal and strong emotional intensity, repeating it twice. When speaking, performers use these eight emotions: neutral, surprise, happiness, sadness, disgust, fear, anger. When singing, they expressed happiness, sadness, calm, fear, anger as well as neutral

2. SAVEE Dataset

The SAVEE [19] database contains audio-visual, audio and video samples and is an extract of 480 British speech samples from the TIMIT DB by 4 male performers between the ages of 27 and 31 using 7 emotions. Video samples were captured at 60 fps and audio at 44.1 kHz. The count of samples of each emotion in the audio files is as follows: Happiness (60), Sadness (60), Fear (60), Anger (60), Disgust (60), Surprise (60), Neutral (120).

B. Performance

The study analyzed the performance of the fusion features and assessed the SVM algorithm. Testing was executed against two datasets and results are presented in these tables.

TABLE I. CONFUSION MATRIX WHEN USING THE PROPOSED METHOD ON RAVDESS DATABASE. NUMBERS REPRESENT % ACCURACY.

	N	*F*	*Su*	*H*	*S*	*A*	*D*
Natural (N)	**96.49**	1.75	0	0	0	0	0
Fear (F)	0.87	**97.39**	0	0	0	0.87	0.87
Surprise(S)	0.87	0	**97.39**	0	0.87	0	0
Happy (H)	0	0	0	**97.39**	0.87	0	0.87
Sad (S)	0	0	1.74	0	**97.39**	0.87	0
Angry (A)	0	0.87	0	0	0.87	**97.39**	0.87
Disgust (D)	0	0	0.87	0	0	0	**97.39**

Table 1 presents the results when deploying the RAVDESS database, which contains audio, visual and audio-visual clips from 12 male and 12 female performers speaking and singing the same two sentences but in various different emotions and with normal emotional intensity and strong intensity, and repeat them twice.

Table 2 presents the results from using the SAVEE database.

TABLE II. CONFUSION MATRIX USING THE PROPOSED METHOD ON SAVEE DATABASE. THE NUMBERS REPRESENT % ACCURACY.

	N	*F*	*Su*	*H*	*S*	*A*	*D*
Natural (N)	**92.59**	0	3.70	0	3.70	0	0
Fear (F)	0	**87.50**	6.25	0	0	0	6.25
Surprise(S)	0	0	**92.31**	0	7.69	0	0
Happy (H)	0	0	13.33	**86.76**	0	0	0
Sad (S)	0	0	15.38	0	**84.62**	0	0
Angry (A)	5.56	0	0	0	0	**94.44**	0
Disgust (D)	0	0	0	0	5.56	0	**94.44**

V. TESTING THE SYSTEM

A. Feature Extraction

The process and techniques used to obtain audio features were as follows: Extracted features will be useful for testing in the future and so we stored them in a database. Training samples make up 70% of the dataset, while the remaining 30% is considered as being test data. When the application is initialized, it requests the training audio first, followed by the test audio clips. Once loaded, the feature abstraction functionality performs the processing on the audio files.

B. Testing of the System

It was necessary to validate that the training database was formulated correctly and that the system would operate correctly. The test involved using the test audios to interrogate the training database. We deployed a classifier designed to assess the test images using the training images using code we developed in MATLAB. That code was deployed against the SAVEE and RAVDESS databases. Both that code and the SVM classifier are detailed below, along with the test results.

The audio signal sampling frequency is 44.1 KHz. Frame size consists of 236 samples with 118 overlapping sample frames. 12 Mel coefficients with low frequency of 300Hz and high frequency 3700 Hz. The 20 Mel scale filter bank was deployed as depicted below.

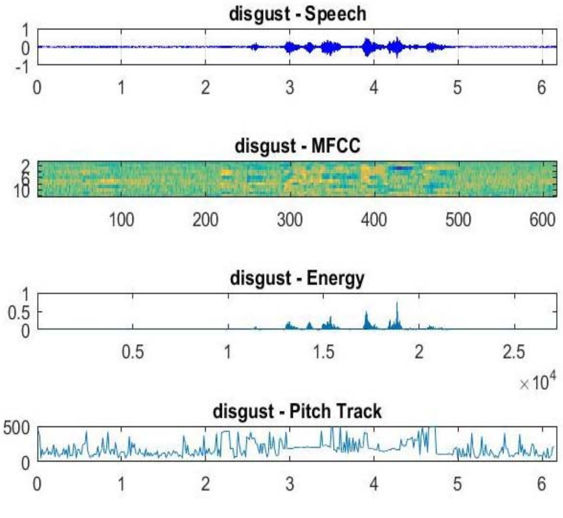


Fig. 6. Extracting the Disgust feature using MATLAB.

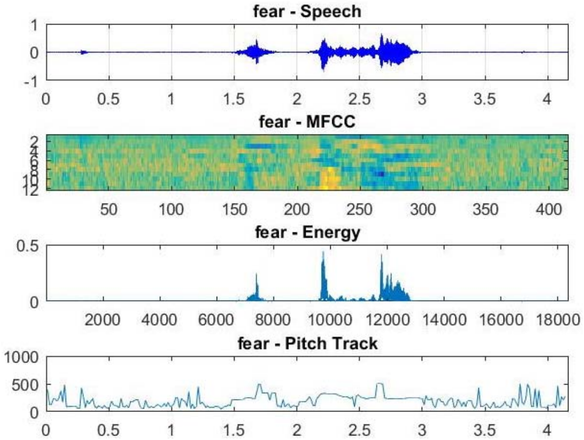


Fig. 7. Extracting the Fear feature using MATLAB.

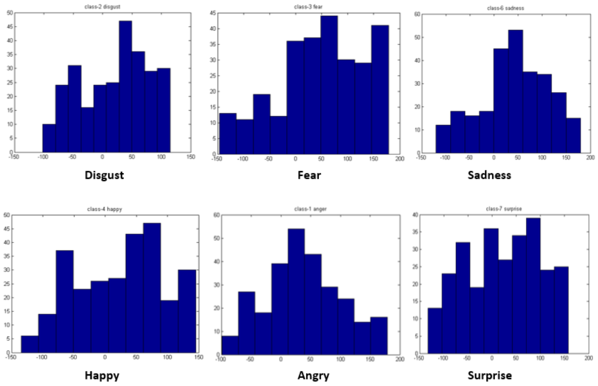


Fig. 8. Histogram of Facial Landmarks descriptors using RAVDESS dataset using MATLAB.

C. Testing the proposed on mobile phone

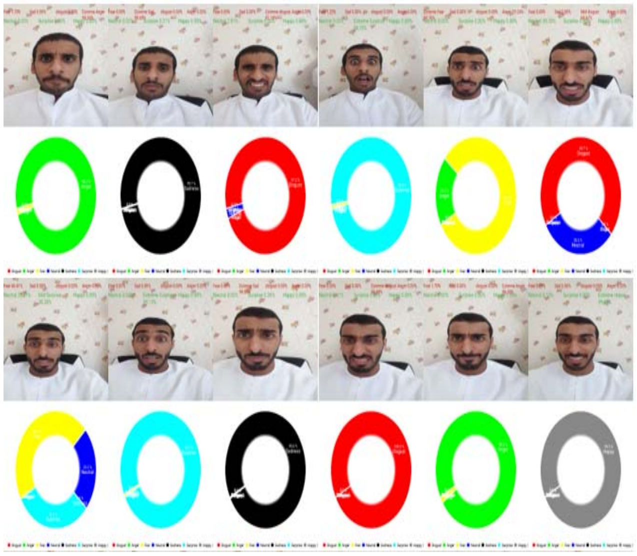


Fig. 9. Output of automated facial expressions recognition module.

Fig. 9 shows simple smartphone testing results with the percentages of facial expression recognition in real time, where the pie chart in Green represents Anger; Gray represents Happiness; Red represents Disgust; Black represents Sadness; Blue represents Neutral.

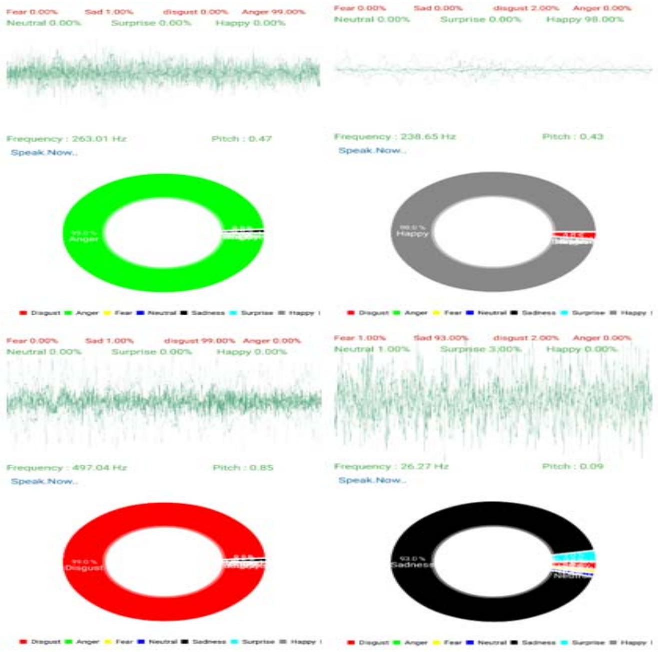

Fear 0.00% Sad 1.00% disgust 0.00% Anger 99.00%
Neutral 0.00% Surprise 0.00% Happy 0.00%
Frequency : 263.01 Hz Pitch : 0.47
Speak.Now..
99.0 % Anger
Fear 0.00% Sad 0.00% disgust 2.00% Anger 0.00%
Neutral 0.00% Surprise 0.00% Happy 98.00%
Frequency : 238.65 Hz Pitch : 0.43
Speak.Now..
98.0 % Happy
Disgust Anger Fear Neutral Sadness Surprise Happy
Fear 0.00% Sad 1.00% disgust 99.00% Anger 0.00%
Neutral 0.00% Surprise 0.00% Happy 0.00%
Frequency : 497.04 Hz Pitch : 0.85
Speak.Now..
99.0 % Disgust
Fear 1.00% Sad 93.00% disgust 2.00% Anger 0.00%
Neutral 1.00% Surprise 3.00% Happy 0.00%
Frequency : 26.27 Hz Pitch : 0.09
Speak.Now..
93.0 % Sadness
Disgust Anger Fear Neutral Sadness Surprise Happy

Fig. 10. Output of automated speech emotion recognition module.

Fig. 10 shows simple smartphone testing results with the percentages of speech emotional recognition in real time, where the pie chart in Green represents Anger; Gray represents Happiness; Red represents Disgust; Black represents Sadness; Blue represents Neutral.

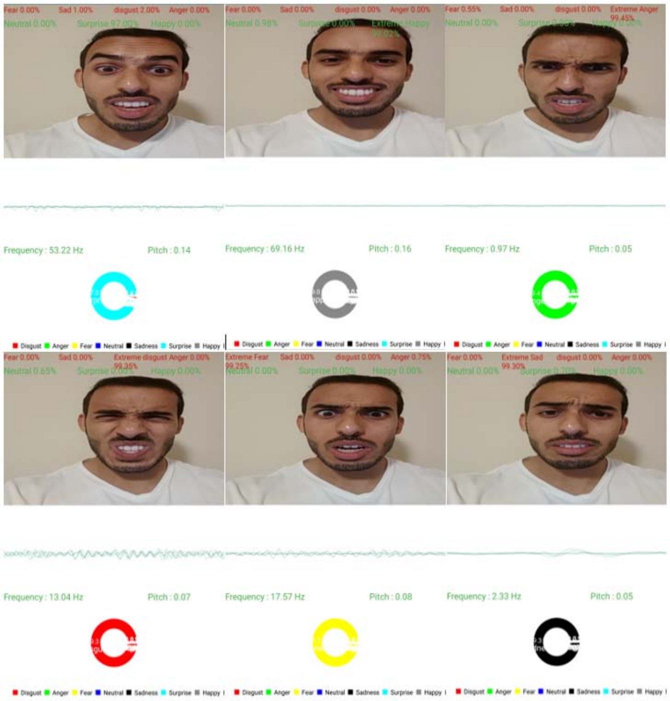

Fear 0.00% Sad 1.00% disgust 2.00% Anger 0.00%
Neutral 0.00% Surprise 97.00% Happy 0.00%
Fear 0.00% Sad 0.00% disgust 0.00% Anger 0.00%
Neutral 0.98% Surprise 0.00% Extreme Happy 99.02%
Fear 0.55% Sad 0.00% disgust 0.00% Extreme Anger 99.45%
Neutral 0.00% Surprise 0.00% Happy 0.00%
Frequency : 53.22 Hz Pitch : 0.14
Frequency : 69.16 Hz Pitch : 0.16
Frequency : 0.97 Hz Pitch : 0.05
Disgust Anger Fear Neutral Sadness Surprise Happy
Fear 0.00% Sad 0.00% Extreme disgust 99.35% Anger 0.00%
Neutral 0.65% Surprise 0.00% Happy 0.00%
Extreme Fear 99.25% Sad 0.00% disgust 0.00% Anger 0.75%
Neutral 0.00% Surprise 0.00% Happy 0.00%
Fear 0.00% Extreme Sad 99.30% disgust 0.00% Anger 0.00%
Neutral 0.00% Surprise 0.70% Happy 0.00%
Frequency : 13.04 Hz Pitch : 0.07
Frequency : 17.57 Hz Pitch : 0.08
Frequency : 2.33 Hz Pitch : 0.05
Disgust Anger Fear Neutral Sadness Surprise Happy

Fig. 11. Output of automated multimodale emotion recognition on mobile phone.

Fig. 11 shows simple smartphone testing results with the percentages of recognized emotions in real time, where the pie chart in Green represents Anger; Gray represents Happiness; Red represents Disgust; Black represents Sadness; Blue represents Neutral.

D. Testing Summary

Assessing the performance of the proposed system to evaluate its effectiveness was the final stage in our study. The study's purpose at the outset was to establish the feasibility or otherwise of deploying a mobile platform together with cloud capability. We conducted testing through all phases of the project from the very beginning. The first step was to isolate each function's information separately. A critical requirement was to validate that each line of code performed exactly what it was designed to do and that the algorithm executed properly. To fulfil the requirements of code testing, we varied between building a number of suitable datasets, and using tiny visual and audio samples. Just to take one instance: when calculating the outcomes of functions, we physically counted and assessed both the output of code results and of objective results and stored them in a numeric matrix database. We ensured that we tested the performance of each stage when it was completed and then, after those results were assimilated into the overall results, we evaluated the system performance. Sometimes, recommended algorithms unexpectedly contributed nothing to the project, meaning we had to discard them and begin again. Disappointingly, the SVM failed to live up to expectations despite receiving high praise in at least three research papers. A feature of the work was the extremely high number of multiclass features, which were difficult to train using the SVM algorithm. The solution was found to be a combination of certain Android tools in conjunction with MATLAB. The application code was developed with the help of the Android tool while the application's functionality was simulated by the MATLAB.

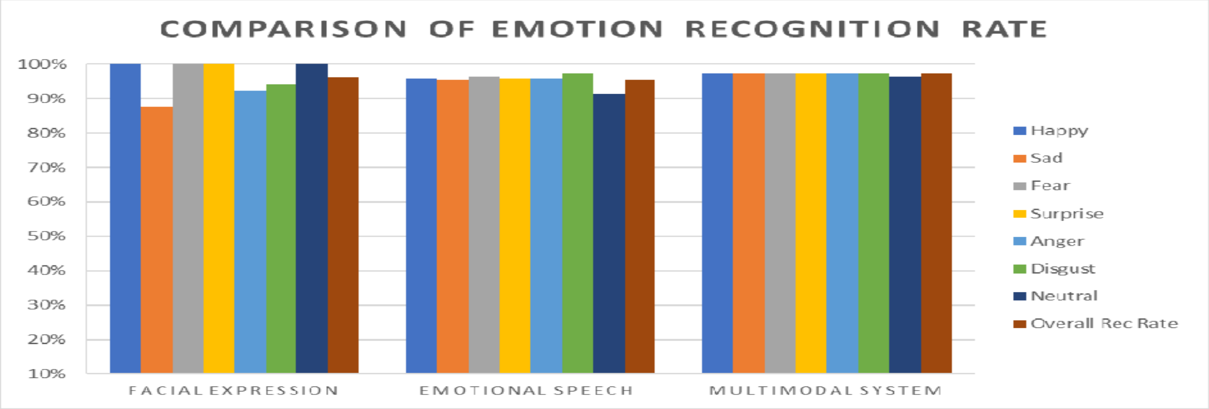


Fig. 12. *Performance comparison on Facial expression , Speech information and Multimodal system(Facial and speech recognition).*

VI. Conclusion and discussion

This project presents a merging framework to perform recognition and analysis of emotions in speech and facial expressions. The proposed approach set out to identify 7 emotions by using different classifiers: neutral, happy, sad, surprise, fear, anger and disgusted. The Support Vector Machine (SVM) achieved the best accuracy. The primary objective was to identify a number of suitable features suitable for emotion classification. A secondary purpose was to analyze unimodal emotion recognition systems for strengths and weaknesses when dealing with speech and facial expressions. The multimodal emotion recognition (facial expression and speech emotional recognition) framework was tested against instances of the RAVDESS database. Average recognition accuracy using a smartphone reached 96.3% for facial expressions and speech recognition achieved 95.43%

Information fusion uses the method of fusing the speech recognition results with those from facial recognition. This combined approach improves the recognition accuracy by 1.83 compared to using speech alone and by 1% compared to using visual recognition alone.

The method proposed in this paper has successfully achieved a high degree of accurate facial expression and speech emotion recognition, which is an enabler for even greater progress in the fields of Human-Computer Interaction (HCI) and Artificial Intelligence (AI). Our next objective is to build on this success, further improve the techniques, and eventually progress to testing it on a multilingual dataset as a future point.

References

[1] A. Kolakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wrobel, "Emotion recognition and its application in software engineering," 2013 6th International Conference on Human System Interactions (HSI), 2013.

[2] P. Washington, C. Voss, N. Haber, S. Tanaka, J. Daniels, C. Feinstein, T. Winograd, and D. Wall, "A Wearable Social Interaction Aid for Children with Autism," Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA 16, 2016.

[3] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution

and Summarization," IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 255–270, Jan. 2018.

[4] Robert Plutchik and Henry Kellerman. 2013. Theories of emotion. Vol. 1. Academic Press.

[5] L. F. Barrett and J. A. Russell, "The Structure of Current Affect," Current Directions in Psychological Science, vol. 8, no. 1, pp. 10–14, 1999.

[6] "Facial Action Coding System," The SAGE Encyclopedia of Communication Research Methods.

[7] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," Image and Vision Computing, vol. 65, pp. 23–36, 2017.

[8] P. Khorrami, T. L. Paine, and T. S. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?," 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.

[9] M. Kächele, M. Schels, S. Meudt, G. Palm, and F. Schwenker, "Revisiting the EmotiW challenge: how wild is it really?," Journal on Multimodal User Interfaces, vol. 10, no. 2, pp. 151–162, Dec. 2016.

[10] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild," Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI 15, 2015.

[11] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification," Procedings of the British Machine Vision Conference 2012, 2012.

[12] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016, 2016.

[13] A. Tran and L.-F. Cheong, "Two-Stream Flow-Guided Convolutional Attention Networks for Action Recognition," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

[14] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning," Proceedings of the 16th International Conference on Multimodal Interaction - ICMI 14, 2014.

[15] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive Multi-Modal Gesture Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1692–1706, Jan. 2016.

[16] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," Current Directions in Psychological Science, vol. 20, no. 5, pp. 286–290, 2011.

[17] G. Ghinea, R. Kannan, and S. Kannaiyan, "Gradient-Orientation-Based PCA Subspace for Novel Face Recognition," IEEE Access, vol. 2, pp. 914–920, 2014.

[18] S. R. Livingstone, W. F. Thompson, M. M. Wanderley, and C. Palmer, "Common cues to emotion in the dynamic facial expressions of speech and song," Quarterly Journal of Experimental Psychology, vol. 68, no. 5, pp. 952–970, 2015.

[19] S. Haq and P. J. B. Jackson, "Machine Audition: Principles, Algorithms, and Systems," Hershey PA, 2010, pp. 398-423.

[20] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast Keypoint Recognition Using Random Ferns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 3, pp. 448–461, 2010.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05).

[22] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," Face and Gesture 2011, 2011.

[23] C. Orrite, A. Gañán, and G. Rogez, "HOG-Based Decision Tree for Facial Expression Classification," Pattern Recognition and Image Analysis Lecture Notes in Computer Science, pp. 176–183, 2009.

[24] D. Gabor, "Theory of communication," Journal of the Institution of Electrical Engineers - Part I: General, vol. 94, no. 73, pp. 58–58, 1947.

[25] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition.

[26] I. Gonzalez, H. Sahli, V. Enescu, and W. Verhelst, "Context-Independent Facial Action Unit Recognition Using Shape and Gabor Phase Information," Affective Computing and Intelligent Interaction Lecture Notes in Computer Science, pp. 548–557, 2011.

[27] T. Wu, M. S. Bartlett, and J. R. Movellan, "Facial expression recognition using Gabor motion energy filters," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010.

[28] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," 7th International Conference on Automatic Face and Gesture Recognition (FGR06).

[29] C.-C. Lee and C.-Y. Shih, "Gabor Feature Selection and Improved Radial Basis Function Networks for Facial Expression Recognition," 2010 International Conference on Information Science and Applications, 2010.

[30] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271).

[31] C. Xu, C. Dong, Z. Feng, and T. Cao, "Facial Expression Pervasive Analysis Based on Haar-Like Features and SVM," Communications in Computer and Information Science Contemporary Research on E-business Technology and Strategy, pp. 521–529, 2012.

[32] S. U. Jung, D. H. Kim, K. H. An, and M. J. Chung, "Efficient rectangle feature extraction for real-time facial expression recognition based on AdaBoost," 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.

[33] C. Shan, S. Gong, and P. W. Mcowan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[34] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of Facial Expression Extracted Automatically from Video," 2004 Conference on Computer Vision and Pattern Recognition Workshop.

[35] G. M. Nagi, R. O. Rahmat, F. Khalid, and M. Taufik, "Region-Based Facial Expression Recognition in Still Images," Journal of Information Processing Systems, vol. 9, no. 1, pp. 173–188, 2013.

[36] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583).

[37] S. Eleftheriadis, O. Rudovic, and M. Pantic, “Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition,” IEEE Transactions on Image Processing, vol. 24, no. 1, pp. 189–204, 2015.

[38] P. Shen, Z. Changjun, and X. Chen, “Automatic Speech Emotion Recognition using Support Vector Machine,” Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, 2011.

[39] E. Mower, M. J. Mataric, and S. Narayanan, “A Framework for Automatic Human Emotion Classification Using Emotion Profiles,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057–1070, 2011.

[40] Praat: doing Phonetics by Computer. [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed: 16-Oct-2018].

[41] “Neural Networks: A Comprehensive Foundation,” Haykin, Neural Networks: A Comprehensive Foundation | Pearson. [Online]. Available: https://www.pearson.com/us/higher-education/product/Haykin-Neural-Networks-A-Comprehensive-Foundation-2nd-Edition/9780132733502.html. [Accessed: 16-Oct-2018].

[42] S. Singh and E. Rajan, “MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors,” International Journal of Computer Applications, vol. 21, no. 6, pp. 1–6, 2011.

[43] R. Nakatsu, “A speech recognition machine for connected words,” ICASSP 80. IEEE International Conference on Acoustics, Speech, and Signal Processing.

[44] R. Akbani and T. Korkmaz, “Applications of Support Vector Machines in Bioinformatics and Network Security,” Application of Machine Learning, Jan. 2010.

[45] D. Xi and S.-W. Lee, “Face detection and facial feature extraction using support vector machines,” Object recognition supported by user interaction for service robots.

[46] P. Ekman, “An argument for basic emotions,” Cognition and Emotion, vol. 6, no. 3-4, pp. 169–200, 1992.

[47] C. Xu, C. Dong, Z. Feng, and T. Cao, “Facial Expression Pervasive Analysis Based on Haar-Like Features and SVM,” Communications in Computer and Information Science Contemporary Research on E-business Technology and Strategy, pp. 521–529, 2012.