Jurnal
Teknologi

Full Paper

HYBRID FUSION OF FACE AND SPEECH INFORMATION FOR BIMODAL EMOTION ESTIMATION

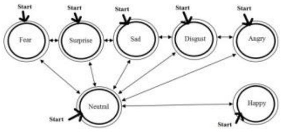
Article history
Received
6 November 2015
Received in revised form
21 March 2016
Accepted
10 March 2016

Krishna Mohan Kudiri[a]*, Abas Md Said[a], M Yunus Nayan[b]

[a]Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia
[b]Applied Sciences Department, Universiti Teknologi PETRONAS, Malaysia

*Corresponding author
kris_g01783@utp.edu.my

Graphical abstract



Abstract

Estimation of human emotions during a conversation is difficult using a computer. In this study, facial expressions and speech are used in order to estimate emotions (angry, sad, happy, boredom, disgust and surprise). A proposed hybrid system through facial expressions and speech is used to estimate emotions of a person when he is engaged in a conversational session. Relative Bin Frequency Coefficients and Relative Sub-Image-Based features are used for acoustic and visual modalities respectively. Support Vector Machine is used for classification. This study shows that the proposed feature extraction through acoustic and visual data is the most prominent aspect affecting the emotion detection system, along with the proposed fusion technique. Although some other aspects are considered to be affecting the system, the effect is relatively minor. It was observed that the performance of the bimodal system was lower than the unimodal system through deliberate facial expressions. In order to deal with the problem, a suitable database is used. The results indicate that the proposed system showed better performance, with respect to basic emotional classes than the rest.

Keywords: Relative Bin Frequency Coefficients (RBFC), Relative Sub-Image Based (RSB), Support Vector Machine (SVM)

Abstrak

Menganggarkan emosi manusia semasa perbualan adalah sukar menggunakan komputer. Dalam kajian ini, ekspresi muka dan ucapan digunakan untuk menganggarkan emosi (marah, sedih, gembira, bosan, meluat dan kejutan). Sistem hibrid melalui ekspresi muka dan ucapan digunakan untuk menganggarkan emosi seseorang apabila dia melibatkan diri dalam sesi perbualan. Pekali Kekerapan Bekas Relatif dan ciri-ciriberdasarkan sub-imej relative masing-masing digunakan untuk modus akustik dan visual. Mesin Vektor Sokongan digunakan untuk pengelasan. Kajian ini menunjukkan bahawa pengekstrakan ciri yang dicadangkan melalui data akustik dan visual adalah aspek yang paling menonjol yang mempengaruhi sistem pengesanan emosi, bersama-sama dengan teknik gabungan yang dicadangkan. Walaupun beberapa aspek lain dilihat mempengaruhisystem tersebut, kesannya adalah agak kecil. Prestasi sistem bimodal didapati lebih rendah daripada sistem unimodal melalui ekspresi muka sengaja. Dalam menangani masalah ini, pangkalan data yang sesuai digunakan. Keputusan menunjukkan bahawa sistem yang dicadangkan menunjukkan prestasi yang lebih baik terhadap kelas emosi asas berbanding daripada yang lain.

Kata kunci: Pekali kekerapan bekas relatif, ciri berdasarkan sub-imej, mesin vektor sokongan

1.0 INTRODUCTION

Human-human communication has been started since the starting of human life. This communication is being through social cues like verbal signs of speech signal, non-verbal signs such as gesture, speech tone and facial expressions. This communication consists of person's emotional and non-emotional data towards his goal. Human encodes their emotions and decodes other's emotions while communicating with each other. Darwin, (1998) started study of emotions, and its significance on human life [1]. From the understanding of emotions, Picard et al., (1997) released a book on the emotional side of computers [2]. This book triggered a new research area called affective computing. According to that, identification of human emotion helps to design an intelligent computing system. And also it is essential for computers to understand human emotions to act like human. Many scientists in the field of computer science have inspired by the above theoretical studies and focused on effective and user-friendly Human Computer Interaction (HCI).

Effective computing is an interdisciplinary field which talks about the relation between computer and emotions. According to that, estimation of human emotions is needed, in order to estimate his needs towards his goals. Emotion plays pivot role in human computer interaction while dealing with user. And also suggested that estimation of human emotions with proper feedback to the user helps to boost the communication process between computer and user. This improves users' trust on HCI. Furthermore, it is also helpful for the computer to make optimal decisions.

In the market, identification of emotions is possible in many ways as follows: physiological methods such as Electromyography (EMG) consists of electrodes to estimate possible muscle contractions using threshold values; Electrocardiograph (ECG) consists of electrodes to study the heart beat rate per second and it intervals; Electroencephalography (EEG) consists of 16 to 32 electrodes to study the interaction between the left and right brains for each emotional class. However, the above techniques are not suitable for social human-human or human-computer interaction due to their voluntariness.

Estimation of human emotions using computer triggred in many areas as follows: Alzheimer's disease is a disease that destroys human memory and metal functions. That leads to dangerous short-term memory loss. Due to that, slowly patient loses bodily functions and it ultimately leads to death. Currently this disease affected an estimated 1 in 10 people over age 65. Alzheimer's disease is a specific cause of dementia. In order to detect it, doctors use a variety of screenings including blood tests, mental status evaluations and brain scans. In this case, this proposed research work helps to the doctor in order to estimate the mental state of the patient without using any electrodes on patient's body. This research work helps to the doctor to distinguish Alzheimer's patients from those with other form of dementia.

Battocchi et al., (2005) stated that, human facial expressions are two types namely, standard and non-standard facial expressions [3]. A standard facial expression means deliberate facial expressions. A non-standard facial expression means facial expressions during speech. Identification of human emotions using a computer through deliberate facial expressions is easy compared to facial expressions during speech. It is due to adjacent emotions detection problem. Due to that, many researchers started focusing on other dimensions (EGC, EEG) instead the above. However, identification of human emotions in social human-human or human-computer is only possible through speech and facial expressions. Thus, this research work focused on the above modalities which will be discussed in detail in Section 2. The rest of the paper is organized as follows: Section 2 discusses about related work. Section 3 explains about proposed work. Section 4 presents results and discussion and finally Section 5 gives the conclusion and future work.

2.0 RELATED WORK

2.1 Facial Expressions

Feature extraction from visual is possible in two ways namely geometric based and appearance based. Geometric based feature extraction focuses on complete face to extract emotional data. Whereas to extract emotional information, the appearance based technique focuses on facial skin changes likely, winkles and bulges.

Geometric based method: Lien (1998) used wavelet based method for complete face regarding emotion detection [4]. In addition, PCA also used to optimize the data to process in vertical and horizontal directions. In this case, there is no guarantee that PCA always gives better accuracy. Otsuka and Ohya, (1998) used local regions of eyes and mouth regarding emotion detection [5]. But, the above techniques contain noise and loss of data problem.

Wang *et al*., (2006) identified geometric displacements using manual feature points in the form of lines and dots near to eyes, eyebrows and mouth to detect emotions [6]. However, this method failed to predict feature points automatically. Kaliuby and Robinson (2005) applied color plastic dots to the users face to recognize the facial muscle moments clearly [7]. This method gives better accuracy than conventional methods. However, labeling the points manually is not useful for natural human computer interaction.

Appearance based method: Vukadinovic and Pantic (2005) used Viola-Jones algorithm to detect faces and then the face is subdivided into 20 regions [8]. Next, each region is tested for region of interest through Gabor wavelets. Padgett and Cottrell (1997) used Kernel based approach to detect emotion through facial expression [9]. In this case, Principal Component Analysis (PCA) is used for full face to detect emotions. The advantage of PCA is fast and simple. Moreover, PCA provides better accuracy for person dependent emotion detection system in natural human - computer interaction.

Hung-Hsu *et al*., (2010) developed a novel human facial emotion detection system using angular radial transform, discrete cosine transform and Gabor filter [10]. Suja *et al.*, (2014) used Dual-Tree Complex Wavelet Transform (DT-CWT) and Gabor Wavelet Transform to detect emotions [25]. This system contains loss of data due to filters. Thus, the system cannot provide better accuracy for real time data. From the above two, hybrid technique came into the picture. However, hybrid feature extraction techniques are not suitable for fully automatic emotion detection systems.

From the above studies, identification of emotions through facial expression needs to be suitable for real time data. While the above mentioned methods, every method focused on facial expressions and it's experienced emotion in outside environment. In this case, no one is focused on actual meaning of emotion. According to Darwin theory (1998) people use facial expressions to express their feeling to others in order to reach their goal [1]. Identification of every human emotion through facial expression is not possible except basic emotions. Basic emotions relate to basic needs. Thus, identification of facial expressions for basic emotions is unique and easy to predict.

From the above discussion on 2D approach, it is concluded that no conventional method from the above approaches is supporting fully automatic system effectively. According to that, appearance based approach is used by the proposed research work in order to implement fully automatic emotion detection system for real-life conditions. To do so, the proposed research work is used RSB features which will be discussed in Section 3.

2.2 Speech

Pantic *et al*., (2007) developed a system and used pitch, speech rate and intensity of the signal to identify emotions through speech [11]. Pitch is a famous and basic element of speech signal in the frequency domain. Additionally, speech rate is calculated through number of samples. Finally, the intensity of the signal computed through the strength of the input speech signal. Sometimes, the speech intensity also depends on some other criteria's like, mike position and environmental noise. However, this system achieved around 70% of the system accuracy to classify human emotions from the given input signal.

Liu *et al*., (2007) extracted formant features from the vocal chord vibrations to detect emotions [12]. Formant contains range of a frequency of the input signal which is used mostly on gender detection. Human audible range always vary between 1000 Hz to 20,000 HZ. According to that, human audible range exists in between formant 1 to 3. In this case, formant one and two useful for male voice and formant two and three useful for female voice. Kun Han *et al.*, (2014) used MFCC and prosodic features to detect emotions [24].

Busso *et al*., (2004) showed the emotion detection system through speech using prosodic features [13]. The performance of the system depends on the proper frequency spectrum due to sufficient length of the input signal. Thus, this system is difficult to apply for real-life conditions. From the above studies, suitable feature extraction is needed to escape from the above problems. Furthermore, arousal of emotion is always not possible through single modality. It is used to change between different modalities. Human emotion arousal does not affect speech and face immediately. Thus, emotion detection system in HCI is strictly multimodal.

According to the study by Iohnstone and Scherer (2000), the accuracy of the emotion detection system depends on sufficient length of the input speech signal [14]. This creates a proper frequency spectrum, which helps to predict the basic emotions. However, for real-life conditions, this is not possible. Huang *et al*., (2004) showed that prosodic or acoustic feature shows better accuracy only for person dependent and language dependent approaches [15]. However, this is not applicable for real-life conditions. In order to counter the above problems, the proposed research work is used Relative Bin Frequency Coefficients (RBFC) for speech which will discussed in Section 3.

2.3 Fusion through Speech and Facial Expressions

Multimodal system means, one system takes multiple inputs and processes together to produce a single output. In order to achieve this, the system needs to process inputs simultaneously and jointly together to produce desired output. According to the proposed researched work, joint analysis through speech and facial expressions needs to implement because:

- Massaro and Egan (1996) stated the importance of joint analysis of speech and facial expressions to predict emotions [16]. Russell *et al*., (2003) also showed support to joint analysis of speech and facial expressions [17]. According to their studies, Integration of the above modalities helps to classify emotions.
- Identification of facial expressions through camera contains many limitations like, lighting condition problem, head pose, occlusion. On the other hand, Identification of speech through microphone contains limitations like, noise and distance between microphone and speaker. Thus, there is no guarantee that, both channels are available on every time. In order to solve this problem, joint analysis is needed.
- Joint analysis of both speech and facial expression helps to understand emotions even in speech related facial configurations.

Fasel and Luettin (2003) suggested that, there should be a weight for each modality, while integrating them [18]. For instance, speech modality mostly produces better results to predict aggressive emotion. On the other hand, facial expressions help to produce better results to predict happy emotion. Thereby, depends on their importance, weights are needed to predict emotions for multimodalities.

Paleari and Lisetti (2006) discussed about signal level, feature level and decision level fusions [19]. According to their framework, they focused on multimodal emotion detection with different fusion techniques. But they did not report any practical implementation and results.

Signal level fusion is applicable only for tightly synchronous signals, which are similar in nature. Signal level fusion is a mixture of the both signals together to produce one output. Here, these two signals as follows: voice signal from stereo mikes, or images from stereo cameras. However, this signal level fusion is not suitable for the multimodal emotion detection system. Because, here one is speech modality and another one is image modality. Both are different in nature. Thus, signal level fusion is not suitable for the proposed research work.

Feature level is possible suddenly after feature extraction of different modalities. It is a direct combination between features. Feature level fusion is not suitable for uncorrelated data. The decision level fusion is possible as follows: emotion detection is done separately, and then the final decision is done through classified data. Using this model, information between correlated data from different modalities is not going to be there. This is the drawback of decision level fusion technique.

The goal of this audio-visual fusion is to smooth the speech related facial configurations effect on facial expressions to predict emotions. To do so, this study enables a computer to, understand user emotions effectively. This study focused on an integration of speech and facial expressions at the same time to detect user emotions.

According to human-human communication, human facial expressions are influenced by both speech and internal emotions. In order to smooth the above to understand emotions, study of joint of complementary and conflicting information of both modalities is needed. In this research work, a proposed method is used to joint both modalities which will discuss in Section 3**.**

3.0 PROPOSED WORK

3.1 Facial Expressions

Emotion detection system through facial expressions consists of three steps namely, face detection, feature extraction and classification. Face detection means face prediction in given input image, if there is any. This is a challenging task in real time conditions due to the following issues: pose and angle of the face can vary due to the camera position, target face may occlude or close up to the other objects and lightning condition problem. Furthermore, face is unique to every person and same person may look different with eye glasses, beard and moustache. Next, to evaluate the performance of the face detector in real time, HCI depends on the following factors: detection speed, detection accuracy, required training time and number of training samples. Consider the above mentioned metrics in this research work, Ada-boost based face detector by Viola and Jones (2004) is used [20].

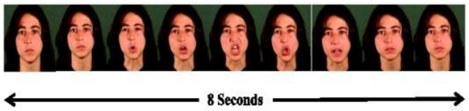


Figure 1 Video file of 8 seconds duration [3]

Figure 1 shows a video file. This video contains both speech and facial expressions. Let's consider facial expressions, it is changing every second. The emotion of that person changes from one class to another. This change is not random in nature. This change is completely depends on state diagram of Plutchik emotional circle (1980) [21] which will be discussed on Section 3.4. However, in order to detect emotions, from the video file, proposed feature extraction is needed.

Algorithm for proposed Relative Sub-Image Based Features (RSB) as follows:

Step 1: Read image and pre-processing.
Step 2: Divide input image into sub-images.
Step 3: Calculate average of each sub-image.
Step 4: Calculate relative difference between each sub-image to its edge cent sub-images.

The proposed RSB (2012) uses pixel intensities to process the data which is similar to human observation in real-life conditions [23]. It is also can able to discriminate adjacent emotions. For instance, human-human or human-computer communication is a mixture of deliberate and non-deliberate emotions. Angry and disgust are adjacent emotions. In order to predict these two emotional classes through facial expressions during speech is difficult for unimodal emotion detection systems. In this case, the proposed RSB features are suitable to deal with the problem.

3.2 Speech

Algorithm for proposed Relative Bin Frequency Coefficients (RBFC) (2012) through speech [22] as follows:

Step 1: Read speech signal.
Step 2: Pre-emphasize the speech signal.
Step 3: Divide the signal into frame size of 20 ms and a shift of 10 ms, and multiple each frame with hamming window.
Step 4: Compute the frequency spectrum from the above.
Step 5: Compute RBFC from the frequency domain data.

This is a frequency independent feature extraction. So, it is able to create a better model for short input signal, and it is easy to implement. This feature helps to create a compact model with less cost and there is no loss of data. Furthermore, it also helps to deal with adjacent emotion detection problem through speech.

3.3 Classification

SVM (2010) is a classifier which uses maximal margin for classification between two classes, namely positive and negative [10]. So far, this is the classifier which has been used for complex patters recognition application tasks. Support Vector Machine (SVM) is a linear classifier which does not only work well on the training samples, but also works equally well on previously unseen samples. The linear SVM is one of the classifier to separate two classes and this is the only classifier which can classify the data in high dimensional space.

3.4 Hybrid Method

Hybrid method is a combination of feature level and decision level fusion systems. Human machine interaction is always possible through active conversation between human and machine in a social environment. In this case, identification of human emotions through facial expressions is difficult in nature. In order to deal with the above, identification of emotions should be possible through facial expressions and speech. Human emotions are asynchronous in nature which is already discussed in the above.

According to that, identification of human emotions is not always possible through both modalities at the same time. The correlation between the both modality changes between time to time. In order to counter the above problems, the hybrid system is needed. Figure 2 shows the block diagram of a proposed hybrid system. According to the diagram, the proposed system contains six stages. This system helps to detect emotions to input video file.

In this system, stage 1 computes emotion for each frame of video file. In this process, stage 3, 4, 5 and 6 helps to do unimodal emotion detection through speech, feature level multimodal fusion, unimodal emotion detection through facial expressions and decision level multimodal fusion respectively. Finally stage 2 computes the final emotion through analysis unit (AU).

In this case, s_1, s_2 and f_1, f_2 shows speech and facial expressions input data. x, y and z shows output of unimodal emotion detection through speech, feature level multimodal system and unimodal emotion detection through facial expressions respectively. Next, i shows feature level fused vector which goes as a input for feature level multimodal analysis unit. k shows decision level fused vector in stage 6. z shows the output of input video file from decision level multimodal analysis unit. Z^{-1} shows previous output in order to estimate current output. Finally this predicted vector goes to final analysis unit (stage 2) to estimate emotion.

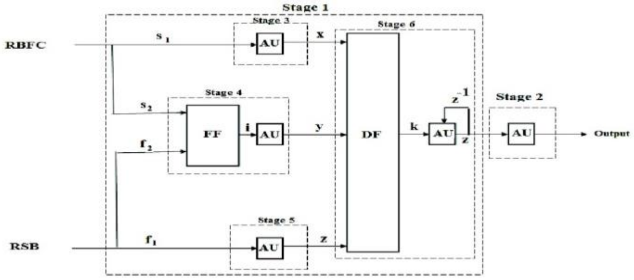


Figure 2 Hybrid system

Stage 6 at analysis unit is using current and previous outputs through state diagram of emotions from Plutchik Wheel and gives output of the system. Figure 3 gives the simplified form of Plutchik Wheel. There each circle represents an emotional class. The intersection of two classes gives another emotion. The intersection also shows a link or a way to change one emotional class to another emotional class. For example, a person cannot be happy after angry emotion within a second vice versa. On the other hand, a person can be angry immediately after disgust emotion, because they are correlated to each other. From the above discussion, Figure 3 illustrates the state diagram for human emotions which helps the proposed system in order to detect emotions at stage 6. According to Figure 3, the fear, emotional class can change to surprise or neutral, but it is not possible for this class to change directly to sad or disgust or angry or happy. The surprise emotion class can change to fear or sad or neutral, but it is not possible to change directly to disgust or angry or happy.

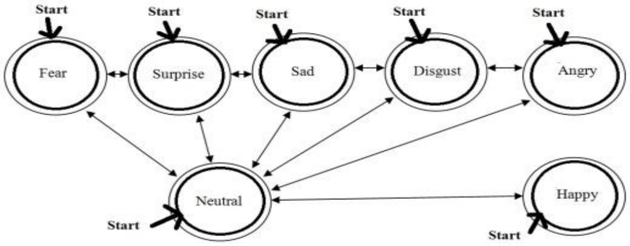


Figure 3 State diagram of emotions through Plutchik Wheel

The sad emotion class can change to surprise or disgust or neutral, but, not to fear or angry or happy. Disgust emotion can change to sad or angry or neutral, but, not to fear or surprise or happy. Angry emotion can change to disgust or neutral, but, not to fear or surprise or sad or happy. Happy emotion can change to neutral only. Any emotional state can start and end or it can remain in the same state. For example, let's consider current emotion shows angry and previous emotion shows disgust, then output is angry emotion. If current emotion is angry and previous emotion is happy, then output is neutral emotion. Because according to Plutchik Wheel, this case is practically not possible. So, it's unknown emotion which also called a neutral emotion.

3.5 Database

The Database of Facial Expressions (DaFEx) (2005) is a database of acted one, which is created by professional actors from Italy (4 males and 4 females) [3]. The database is suitable for unimodal and bimodal system through speech and facial expressions. It contains 1050 video samples, and each sample of 4 or 27 seconds duration. Next, the samples are divided into two parts namely, with speech and without speech. Further, the samples are divided into 6 classes of emotions, namely angry, disgust, fear, happy, sad and surprise. Every emotion is expressed in three intensities namely, high medium and lower. Here, each sample in the database labeled with the help of 80 observers. Thus, this is a suitable database for real time applications.

4.0 RESULTS AND DISCUSSION

4.1 Emotion Detection through Deliberate Facial Expressions

Table 1 shows the size of the database. The database contains a total of 1050 samples, 150 samples in each class. From that 90% of samples is used for training and regaining 10% samples are used for testing. Figure 4 shows the comparison between the performance of the proposed RSB features with other conventional (PCA and Gaussian Wavelet) methods through deliberate facial expressions. Identification of angry emotion is easy through facial expressions. Due to that, angry emotional class achieved better classification accuracy than other emotional classes.

Table 1 DaFEx Database without utterance

Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
150	150	150	150	150	150	150

On the other hand, this angry emotional class achieved 92% classification accuracy using proposed RSB features through SVM. Identification of disgust and surprise emotions is difficult through facial expressions. Facial expressions of sad and disgust look similar to each other. On the other hand, a facial expression of surprise and fear looks similar to each other. Due to that, these both disgust and surprise emotional classes achieved less accuracy than other emotional classes. But, proposed RSB achieved 80% and 83% accuracy for disgust and surprise emotions which is higher than conventional methods.

Furthermore, happy and sad emotional classes are easy to predict through facial expressions. Because, these two emotions effects more number of facial muscles than other emotional expressions. Here also, proposed RSB achieved 91% and 90% accuracy for happy and sad emotions which is higher than the other conventional methods. On the other hand, the proposed RSB features achieved total average accuracy 87.71% which is also higher than conventional features.

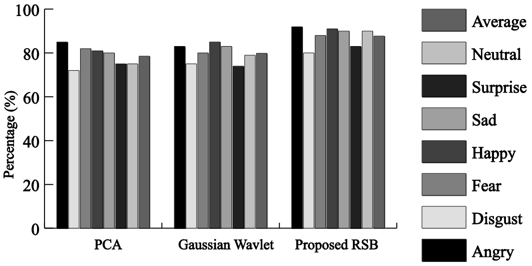
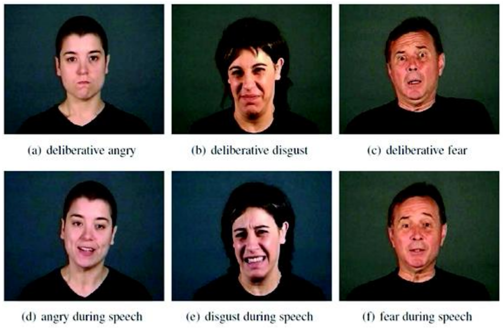


Figure 4 Comparison between conventional features with proposed RSB

Identification of human emotions through facial expressions in a social environment is always playing a higher role. On the other hand, human communication in social environment always contains a mixture of facial expressions with utterance and without utterance. Sadly, human facial expressions are different from without utterance and with utterance as shown in Figure 5. In order to deal with the above, identification of human emotions through facial expressions during speech is needed.


(a) deliberative angry (b) deliberative disgust (c) deliberative fear

(d) angry during speech (e) disgust during speech (f) fear during speech

Figure 5 Human facial expressions for with utterance and without utterance [3]

4.2 Emotion Detection through Facial Expressions during Speech

Table 2 shows a DaFEx database during speech. Similarly like Table 1, here also each class contains 150 samples. All these samples are confirmed to the particular emotional classes with the help of a group of observers. From that, 90% and 10% samples used for training and testing, respectively.

Table 2 DaFEx Database with uttrance

Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
150	150	150	150	150	150	150

In Figure 6, neutral class achieved higher accuracy of 86% using proposed RSB features. This is done due to the problem (adjacent emotions problem) of facial expressions with utterance. On the other hand, proposed RSB achieved better accuracy for angry, fear, happy and sad than conventional features at 85%, 83%, 85% and 85% respectively.

Finally, the proposed RSB features also achieved higher average accuracy of 82.57% than conventional features which is also illustrated in Figure 6. This is due to the adjacent nature between the emotions. Furthermore, neutral emotion confused with every other class of emotions. It's due to nature of neutral class.

Furthermore, Table 3 compares both unimodal emotion detection systems through facial expression with utterance and without utterance, respectively. According to Table 3, the performance of the facial expression without utterance is higher while comparing with facial expression with utterance. Thus, the deliberate facial expression plays a major role. In order to overcome the above problem, observation of facial cures, namely; eyes, eyelids and lips helps to get improved accuracy. However, this is going to make a system semi-automatic. On the other hand, this increases system complexity. In order to deal with the above, another option is to provide a large amount of training data which can help to achieve better accuracy but, practically this is difficult. Thus, speech modality is considered.

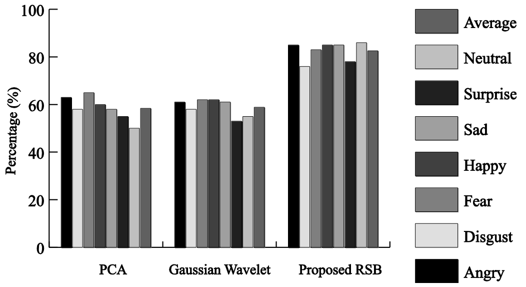


Figure 6 Comparison between conventional features with proposed RSB

Table 3 Comparison between emotion detection through facial expressions with speech and without speech

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Without utterance (%)	83	75	82	85	83	78	86
With utterance (%)	80	72	79	81	78	75	83

4.3 Hybrid Emotion Detection through Speech and Facial Expressions

In this experiment, DaFEx database is used which contains a total of 6 classes of emotions with neutral class and produced results using leave one out cross validation which is shown in Table 4. Each class contains 150 samples and each sample of length 10 seconds duration.

Table 4 DaFEx database of video samples

Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
150	150	150	150	150	150	150

4.3.1 Person Dependent Test

To overcome the above problems, the proposed hybrid system is tested in Table 5. Here, the performance of Hybrid system is better than both feature and decision level systems. This was due to the mixed nature of both fusion techniques. In the Hybrid system, the performance of the angry emotion achieved 95% higher accuracy than the rest. Other than that, the prediction rates of remaining emotional classes achieved almost similar prediction rate. Here, when the data is synchronous, the feature level fusion responds better than decision level fusion. On the other hand, when the data are asynchronous in nature, the decision level fusion is going to respond better. Due to that, the hybrid system performs better in both conditions. Table 6 illustrates the comparison between the proposed system with conventional systems. In this case, feature level and decision level techniques achieved almost similar accuracies. But, proposed system achieved promising results than the rest.

Table 5 Comparison between different fusion techniques

	Feature Level Fusion (%)	Decision Level Fusion (%)	Proposed Hybrid System (%)
Angry	88	62	**95**
Disgust	62	80	**90**
Fear	85	64	**92**
Happy	65	90	**95**
Sad	60	88	**95**
Surprise	82	60	**93**
Neutral	85	88	**95**
Average	75.28	76	**93.57**

4.3.2 Person Independent Test

That is the reason why hybrid system is getting better classification rate than unimodal systems. In a practical scenario, HCI is not unimodal in nature. Mainly dealing with human emotions is purely multimodal. Thus, the hybrid system gives better performance. Here also, proposed system achieves better accuracy than other conventional systems in person independent test which is illustrated in Table 7.

Table 6 Person dependent comparisons between proposed system with conventional systems

Features	Fusion	Classifier	Database type	Average accuracy
Prosodic features for speech and Local Gabor filter for facial expressions	Feature level fusion	SVM	DaFEx	77%
MFCC, PCA for speech and Local Gabor Filters for facial expressions	Decision level fusion	SVM and Bayesian	DaFEx	75%
Color, texture features from facial expressions. Prosodic features from speech.	Hybrid	SVM	DaFEx	81.20%
Proposed System	**Hybrid**	**SVM**	**DaFEx**	**93.57%**

4.4 Hybrid Emotion Detection for Real-life Conditions

To detect emotions in real time conditions through speech and facial expressions, the same procedure is followed like DaFEx database type samples namely; deliberate and with speech. Four subjects (two male and two female) participated in this test. The complete test is divided into two parts, namely training and testing. This test took 7 days to complete the process. First training phase was started. In this training phase, each subject displayed their emotions in front of camera into six emotional classes, namely angry, disgust, fear, happy, sad and surprise. Subjects need to express emotions in two possibilities, namely deliberate and with utterance. This experiment is conducted in Sony VAIO Dual core laptop which is running with Windows 7 operating system.

Table 7 Person independent comparison between proposed system with conventional systems

Features	Fusion	Classifier	Database type	Average accuracy
Prosodic features for speech and Local Gabor filter for facial expressions	Feature level fusion	SVM	DaFEx	63%
MFCC, PCA for speech and Local Gabor Filters for facial expressions	Decision level fusion	SVM and Bayesian	DaFEx	66%
Color, texture features from facial expressions. Prosodic features from speech.	Hybrid	SVM	DaFEx	70.5%
Proposed System	**Hybrid**	**SVM**	**DaFEx**	**90.28%**

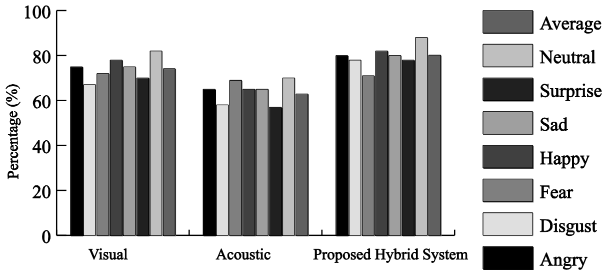


Figure 8 Person independent comparison between the unimodal and hybrid system

Table 8 Comparison between proposed system with conventional systems

Features	Fusion	Classifier	Database type	Average accuracy
Prosodic features for speech and Local Gabor filter for facial expressions	Feature level fusion	SVM	Real time	54.50%
MFCC, PCA for speech and Local Gabor Filters for facial expressions	Decision level fusion	SVM and Bayesian	Real time	57%
Color, texture features from facial expressions. Prosodic features from speech.	Hybrid	SVM	Real Time	60.57%
Proposed System	**Hybrid**	**SVM**	**Real time**	**73.57%**

The proposed system takes input from two input sensors for acoustic and video data. Here Logitech HD Webcam C270 is used for input sensors. It contains screen resolution of 1280 x 720 pixels. It can record voice with 8000 samples per second. It can able to detect target effectively range from one meter distance to user and computer. Through this experiment, the average data captured from each subject are around 250 images. Next, RBFC, RSB features are extracted from these samples and used to train SVM.

Table 8 shows the person independent comparison between both unimodal systems with hybrid system. In this case, except fear emotional class, all other emotional classes achieved higher accuracy in hybrid system which is also illustrated in Figure 8. This is due to the accuracies of both unimodal systems. However it is noted that the performance of the hybrid system is directly proportional to the difference between the accuracies of the unimodal systems.

5.0 CONCLUSION

The comprehensive study on the effects of various aspects such as standard facial expressions and non-standard facial expressions on proposed system has conducted. It was seen that non-standard facial expressions has a significant effect on the proposed emotion detection system at real-life conditions. Overall, every objectives of the work was achieved. Future work needs to focus on many issues about both unimodal and bimodal systems. The proposed RSB feature has lighting problem. The simple solution for this problem might be texture features. Similarly, proposed RBFC has high dimensionality problem. The simple solution for this problem might be PCA to reduce dimensionality. From the experiments, it is indicated that, performance of the system is less while dealing with real-life conditions. Regarding this issue, it might be better to have more number of training samples to provide strong modality. On the other hand, labeling of that training data is another open challenge.

References

[1] Darwin, C., Ekman, P. and Prodger, P. 1998. *The Expression of the Emotions in Man and Animals*.

[2] Picard, R. W. and Picard, R. 1997. *Affective Computing*. 252(1).

[3] Battocchi, A., Pianesi, F. and Goren-Bar, D. 2005. A First Evaluation Study of A Database of Kinetic Facial Expressions (DAFEX), *IEEE 7th International Conference on Multimodal Interfaces*. 558-565.

[4] Lien, J. J. J. 1998. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*.

[5] Otsuka, T. and Ohya, J. 1998. Spotting Segments Displaying Facial Expression from Image Sequences Using HMM. *IEEE International Conference on Automatic Face and Gesture Recognition*. 442-447.

[6] Wang, J., Yin, L., Wei, X. and Sun, Y. 2006. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. *IEEE Computer Society Conference on Computer Vision and Pattern*. 1399-1406.

[7] El Kaliouby, R. and Robinson, P. 2005. Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures, *Real-time Vision for Human-computer Interaction*. 181-200.

[8] Vukadinovic, D. and Pantic, M. 2005. Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. *IEEE International Conference on Systems, Man and Cybernetics*. 1692-1698.

[9] Padgett, C. and Cottrell, G. W. 1997. Representing Face Images for Emotion Classification *Advances in Neural Information Processing Systems*. 894-900.

[10] Tsai, H. H., Lai, Y. S. and Zhang, Y. C. 2010. Using SVM to Design Facial Expression Recognition for Shape and Texture Features. *International Conference on Machine Learning and Cybernetics*. 2697-2704.

[11] Pantic, M. and Baartlett, M. S. 2007. *Machine Analysis of Facial Expressions*.

[12] Liu, J., Chen, C., Bu, J., You, M. and Tao, J. 2007. Speech Emotion Recognition based on a Fusion of All-class and Pair wise-class Feature Selection, *International Conference on Computational Science*. 168-175.

[13] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U. and Narayanan, S. 2004. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. *6th International Conference on Multimodal Interfaces*. 205-211.

[14] Iohnstone, T. and Scherer, K. 2000. Vocal Communication of Emotion. *Handbook of Emotion*. 220-235.

[15] Huang, X., Li, S. Z. and Wang, Y. 2004. Statistical Learning of Evaluation Function for ASM/AAM Image Alignment. *Biometric Authentication*. 45-56.

[16] Massaro, D. W. and Egan, P. B. 1996. Perceiving Affect from the Voice and the Face. *Psychonomic Bulletin and Review*. 215-221.

[17] Russell, J. A., Bachorowski, J. A. and Fernndez-Dols, J. M. 2003. Facial and Vocal Expressions of Emotion. *Annual Review of Psychology*. 329-349.

[18] Fasel, B. and Luettin, J. 2003. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*. 259-275.

[19] Paleari, M. and Lisetti, C. L. 2006. Toward Multimodal Fusion of Affective Cues. *1st ACM International Workshop on Human-centered Multimedia*. 99-108.

[20] Viola, P. and Jones, M. J. 2004. Robust Real-time Face Detection. *International Journal of Computer Vision*. 57(1): 137-154.

[21] Plutchik, R. 1980. A General Psychoevolutionary Theory of Emotion. *Theory of Emotions*.

[22] Mohan Kudiri, K., Md. Said, A. and Nayan, M. Y. 2012. Emotion Detection Using Sub-image Based Features Through Human Facial Expressions. *IEEE International Conference on Computer and Information Science*. 332-335.

[23] Mohan Kudiri, K., Md. Said, A. and Nayan, M. Y. 2012. Emotion Detection Using Relative Amplitude-based Features. *IEEE International Conference on Computer and Information Science*. 522-525.

[24] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. *Annual Conference of International Speech Communication Association*. 223-227.

[25] Suja, P., Shikha Tripathi and Deepthy, J. 2014. Emotion Recognition from Facial Expressions Using Frequency Domain Techniques. *Advances in Intelligent Systems and Computing*. 299-310.