International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Multimodal Emotion Recognition using Facial Expressions, Body Gestures, Speech, and Text Modalities

Mahesh G. Huddar, Sanjeev S. Sannakki, Vijay S. Rajpurohit

Abstract: ***Automatic emotion recognition from multimodal content has become an important and growing research field in human-computer interaction. Recent literature has used either audio or facial expression for emotion detection. However, emotion and body gestures are closely related to one another. This paper explores the effectiveness of using text, audio, facial expression and body gesture modalities of multimodal content and machine learning and deep learning based models for building more accurate and robust automatic multimodal emotion recognition systems. First, we get the best accuracy from the individual modalities. Then we use feature level fusion and ensemble based decision level fusion to combine multiple modalities to get better results. Proposed models were tested on IEMOCAP dataset and results show that proposed models with multiple modalities are more accurate compared to unimodal models in classifying emotions.***

Keywords**:** ***Affective Computing, Multimodal Emotion Classification, Deep Learning, Ensemble, IEMOCAP, LSTM***

I. INTRODUCTION

Due to the advancement in World-Wide-Web (WWW) and the availability of the internet at an affordable rate, people post their opinion about a product or an entity or an event in audio-visual format. People share their opinion about newly released movies, or new products, or hotel, or tourist destination or any other topic in the form of audio-visual content [1]. This gives an opportunity for industries to improve their revenue and trust in their products. The knowledge extracted from audio-visual content can improve the quality of life. Emotions play a crucial role in social interaction, decision making and outcome [2]. Physical expressions and emotions are closely related to one another. While emotion can be expressed in many forms in human-computer interaction such as text, speech, facial expressions, and body gestures, but emotion recognition from audio modality and facial expressions were extensively studied in recent literature [3][4]. A smile can convey happiness or pleasure, a rise of an eyebrow can convey doubt and anger may be conveyed by clenched teeth. A higher pitch is ambiguous as it may convey either a person is happy or angry. Along with the high pitch, if the person is smiling, we can conclude that the person is happy or frowning then we can classify the person's emotion as anger. Body gestures such as leaning forward may convey a person's interest and collapsed posture may indicate the depression about an object, entity or a product. To the best of our knowledge, about 95% of the literature on emotion recognition has used either facial expression or speech or both, at the expense of other modalities such as text and body gestures. In some cases, fake smile representing disagreement or sarcasm could be difficult to recognize by using individual modalities. For example, the text "The movie is sick" could be difficult to classify, but if the person is smiling then the emotion of the person could be classified as happy or pleasure. "The movie is sick" and frown could be classified as anger or displeasure. In this paper, we extensively study the impact of body gestures and the combination of modalities (such as face + speech, face + text, face + body gestures, text + speech, text + body gestures, speech + body gestures, and finally all modalities) in automatic emotion recognition. In the following section, we review the recent literature in emotion recognition and describe the data set used for the experiments. The proposed approach to emotion recognition using each modality and the combination of modalities are presented. Also, different methods of performing data fusion for bimodal and multimodal emotion recognition are discussed. Finally, the results of the proposed method are compared and conclude the paper with future work.

II. RELATED WORK

IEMOCAP dataset [5] is the largest open-sourced dataset for multimodal emotion detection. IEMOCAP dataset contains two-way acted dyadic conversations among multiple speakers. This dataset consists of 93 videos of approximately 12 hours of audio-visual content including facial expressions, body gestures, text transcriptions, and speech. Most of the recent literature on IEMOCAP dataset has concentrated on emotion detection usingeither facial expressions or speech or both. One of the early works on IEMOCAP dataset is [6] which outperform the state-of-the-art methods such as "HMMs (Hidden Markov Models), SVMs (Support Vector Machines)" [7] by 20% in classification accuracy. They extracted segment level features and constructed the feature vector of dimension 750; the feature vector was feed to multilayer perceptron-based architecture followed by 3 hidden layers each with 256 computational units (neurons).

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Mahesh G. Huddar*, Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, India.
Sanjeev S. Sannakki, Computer Science and Engineering, Gogte Institute of Technology, Belgaum, India.
Vijay S. Rajpurohit, Computer Science and Engineering, Gogte Institute of Technology, Belgaum, India.

Another work [3] follows [6], and they use the recurrent neural network (RNN) based architecture to train the model. They have extracted frame level 32 features such as 12 Mel-frequency cepstral coefficients (MFCC) and their derivatives, pitch (F0), zero-crossing rate, voice probability. The proposed network model contains 2 hidden layers each with 128 Bidirectional long short-term memory (BLSTM) cells. In [4] to improve the RNN based emotion detection accuracy, they use CTC as the loss function unlike cross-entropy as a loss. They extracted and used 34 features including chroma-based features and spectrum-based features such as flux and roll-off and 12-dimensional MFCC. They extracted speech features at an interval of a 0.2-second window. In some cases, the whole utterance may not have the emotion cues, but few words in utterance may contain the emotionality cues of the utterance which can be handled well by CTC based loss function. Unlike in [3][4] and [6], they use only the perfect acted data, in [8] they used all the data from the dataset for emotion detection. In recent literature, bimodal techniques were developed for emotion classification. Researchers used either feature level (feature concatenation or early fusion) [9], or decision level fusion (late fusion) [10], or hybrid fusion [11] to design audio-visual emotion detection systems. In [12] they fused audio-visual features (feature level fusion) to create a bimodal emotion detection system and showed that bimodal systems achieve higher accuracy than unimodal systems. In [13] authors used decision fusion with audio and textual features for emotion detection. In [14] authors extracted acoustic, lexicon and visual features and used an ensemble approach to ensemble classification of SVM classifier. Their proposed ensemble approach achieves better results than conventional methods. Authors in [15] have extracted acoustic cues and linguistic cues, they fused them at feature level using 3-D activation valance for emotion recognition. In [16] authors extracted textual, speech and visual features using convolutional neural networks. They analyzed sentiment and emotion using multiple kernel learning. In [17] authors performed multimodal sentiment analysis using tensor fusion network. Most of the research in multimodal emotion recognition has focused on using text, audio, and facial expression based mode. In this paper, we explore the usage of text, speech, facial expression along with body gesture and deep learning based models to improve the overall classification accuracy.

III. DATASET AND FEATURE EXTRACTION

IEMOCAP dataset has approximately twelve hours of audio-visual recordings from 10 actors. Each recording contains scripted or improvised dialogues between a male actor and a female actor. Once the dataset is collected, the dataset is divided into small segments or utterances of duration ranging from 3 to 15 seconds. Each utterance/segment is evaluated manually by multiple assessors. Each assessor is given 10 options (anger, excited, neutral, sadness, happiness, disgust frustration, surprise, fear, other) to assign for each utterance. Recent literature has considered only 4 of them angry, excitement (happiness), neutral and sadness. We have also restricted our experiments to 4 options so as to remain consistent with the literature. Each utterance in the dataset is assigned a label by multiple assessors. Hence we consider emotion label where at least 2 assessors were consistent with their decision, again to remain consistent with the literature. Along with the .wav file (speech), the dataset contains the transcript of each of the utterance. One actor (either male actor or female actor) always wears the Motion Capture camera which records the body gestures (hand movement and head rotation) and facial expression of the actor. The train/test distribution of IEMOCAP dataset for the experiment is shown in Table 1.

Table 1 Train / Test distribution of IEMOCAP dataset

	Happy	Angry	Sad	Neutral
Train	1194	933	839	1324
Test	433	157	238	380

The Mocap data (body gesture and facial expression) contains a column of tuples. For facial expression, each tuple contains 165 columns, 18 and 6 columns for hand movement and head rotation respectively. For the Mocap data, all the feature values of specific modalities are sampled between the start time values and finish time values and partitioned them into an array of size 200. We then average along the columns of each of the 200 arrays. Finally, for each utterance, we obtain a (200, 165) dimension feature vector for facial expression and (200, 24) dimension feature vector for body gesture modality. We extract 34 features for speech data they include 13 energy-based Mel-frequency cepstral coefficients (MFCC) features, 8 Time-Spectral features like short-term entropy, spectral spread, zero crossing rate, spectral roll-off, spectral centroid, short-term energy, spectral entropy, and flux and 13 chroma-based features. These features are extracted in 0.1-second frame length at a sample rate of 16 KHz. We consider a maximum of 100 frames (10 seconds) per utterance. Finally, for each utterance, we obtain a (100, 34) dimension feature vector. For each transcribed text utterance, we use the pre-trained Glove embedding's [18] to obtain a feature vector of (500, 300) dimension for deep learning based models and we have used TF-IDF features for machine learning based models.

IV. EMOTION RECOGNITION MODELS

Extensive experiments on the IEMOCAP dataset show that the following models achieve maximum classification accuracy.

A. *Machine Learning Based Models*

To start with our experiments, machine learning based models are designed. Base classifiers such as Support Vector Machines, Decision Tree, Linear Regression, K-Nearest Neighbor and Random forest are used to design unimodal models.

1) Feature Level Fusion

The feature level or early fusion is performed by concatenating the features from two modalities, three modalities, and all modalities at a time.

The base classifiers are trained and tested using these fused feature vectors.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

2) Ensemble Approach

The ensemble is the process of combining the results of two or more classifiers. Initially, we have used traditional ensemble techniques such as voting, averaging and weighted averaging for building decision level models. The traditional weighted averaging ensemble method uses random weights to calculate the average prediction score. To overcome the randomness in selecting the weight, we have proposed an optimal weighted ensemble technique. The proposed method calculates the weight for each classifier based on the accuracy of the classifier against the overall accuracy. Then the average prediction score is calculated using the prediction score of classifier and weight associated with the classifier. For each utterance, the maximum prediction score is calculated, based on the score the associated label or class is assigned to the utterance. The assigned class is compared to the original class to calculate the correct predictions and wrong predictions. Finally, the accuracy is calculated using correct prediction over correct and wrong predictions. Table 2 shows the proposed algorithm for optimal weighted averaging based ensemble approach.

Table 2 Algorithm – Proposed Optimal Weighted Averaging based Ensemble approach

1. $foreach\ classifier\ Ci\ in\ classifier\ ensemble$
$$Weight_{C_i} = \frac{Accuraccy_{C_i}}{\sum_{j=1}^{n} Accuraccy_{C_j}}$$
Where $Accuraccy_{C_i}$ and $Accuraccy_{C_j}$ represents the accuracy of C_i^{th} and C_j^{th} classifier respectively, n is the number of classifiers.
2. $$Avg\_Pred = \frac{1}{n}\sum_{i=1}^{n} Weight_{C_i} * Pred_{C_i}$$
Where AVG_Pred the average prediction for the individual class label is, $Pred_{C_i}$ is the prediction of C_i^{th} classifier.
3. For each utterance in the test set, find the largest of Avg_Pred and assign Class Label 0, 1, 2 or 3 to $Pred\_Class$.
4. For each utterance in the test set
   $if\ Pred\_Class == Target\_Class$
   $Correct\_Pred = Correct\_Pred + 1$
   $else$
   $Wrong\_Pred = Wrong\_Pred + 1$
5. Calculate Accuracy
$$Accuracy = \frac{Correct\_Pred}{Correct\_Pred + Wrong\_Pred}$$

B. Deep Learning Based Models

1) Text-based Model

Emotion recognition from text is performed on transcripts of text obtained from each utterance. The method is similar to natural language processing. Transcript of utterance is converted into word-vectors using Glove Embedding's [18] of 300 dimensions with 840B tokens, 2.2M vocab. The model takes word-vectors as input and classifies the utterance into one of the 4 emotion classes. The proposed model (Model-1) uses 1-D Convolutional Layers with 256, 128, 64 and 32 filters and kernel size 3 followed by a dense layer with 256 computational units feedingto softmax layer with 4outputneurons. Model -2 uses two stacked LSTM (Long Shot-Term Memory) layers with 512 and 256 computational units followed by a fully connected dense layer with 512 computational units, feedingto softmax layer with 4outputneurons. Dropout probability is set to 0.2, and Relu is used as an activation function in each layer. Adam optimizer is used in the fully connected model and Adadelta is used in LSTM based models.

2) Audio-based Model

The input to the Audio-based emotion recognition is (100, 34) dimension feature vector. The first model (Model-1) is built using three fully connected MLP layer with Relu as activation function and softmax layer with 4 output neurons. Model-2 uses 2 LSTM (Long Short-Term Memory) layers with 512 and 256 computational units respectively followed by 512 units dense layer, followed by a Softmax layer with 4 output neurons. Each unit uses Relu as the activation function.

3) Facial Expression-based Model

The input to the facial expression-based emotion recognition model is (200, 165) dimension feature vector. The first model (Model-1) is built using 2-D convolutions with 32, 64 and 128 filters, followed by a dense layer with 256 computational units feedingto softmax layer with 4outputneurons. As the input feature vector has a larger dimension we use 2 LSTM layers each with 512 and 256 computational units respectively in Model-2, followed by a dense layer with 512 computational units, finally softmax layer with 4 output neurons. Relu activation function is used in each computational unit.

4) Body Gesture-based Model

The input to the body gesture-based emotion recognition model is (200, 24) dimension feature vector.

As the input feature vector has a smaller dimension we use

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

one LSTM layer with 256 computational units in Model-1 and 512 computational units in Model-2, followed by a dense layer with 512 computational units, finally softmax layer with 4 output neurons. Relu activation function is used in each computational unit.

Table 3 Model Configurations for Emotion Recognition

Parameter / Model	Text-based Model		Audio-based Model		Face-based Model		Body Gesture-based Model	
	Model-1	Model-2	Model-1	Model-2	Model-1	Model-2	Model-1	Model-2
Input	(500,300)	(500,300)	(100,34)	(100,34)	(200,165)	(200,165)	(200,24)	(200, 24)
Hidden Layer	1-D Convolutions 256, 128, 64 & 32 filters with kernel size 3	512 & 256 Units	3 Fully Connected MLP Layers	512 & 256 Units	2-D Convolutions 32, 64, 128 filters with kernel size 3	512 & 256 Units	256 Units	512 Units
Dense Layer	256 Units	512 Units	-	512 Units	-	512 Units	512 Units	512 Units
Optimizer	Adam	Adadelta	Adam	Adadelta	Adam	Adadelta	Adam	Adadelta
Activation	Relu							
Output	Softmax Layer with 4 Units							
Loss	Categorical Cross-Entropy							
Dropout	0.2							
Batch	100							

5) Multimodal Emotion Recognition Model

In multimodal emotion recognition, two or more modalities such as text, audio, and video are used. Bimodal (a combination of two modalities at a time), trimodal (a combination of three modalities at a time) and all modalities at a time models are built using the unimodal models. For combining models, first, we have selected the best performing models from each modality. Then, we concatenate the output of final hidden layers (without output neurons) of individual modality into a hidden layer with 512 dimensions, followed by the softmax layer with 4 output neurons. Relu is used as an activation function and Adadelta as the optimizer. Figure 1 shows the concept where all the modalities are combined. Finally, rather than considering Face and body gesture as two modalities, they were combined to form one modality (say video) with (200, 189) sized feature vector.

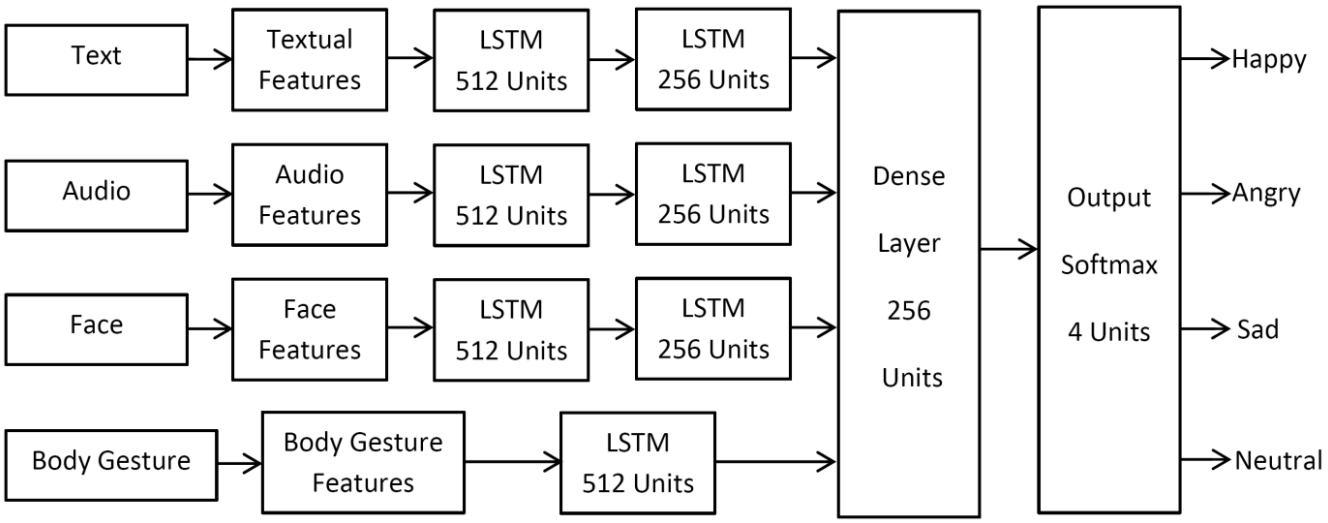

Figure 1 Multimodal Emotion Recognition using Text, Audio, Face and Body Gesture

V. EXPERIMENTAL RESULTS

The proposed models are evaluated on IEMOCAP open-source dataset. Models are implemented using Python-based Scikit-learn and Keras libraries. Initially, unimodal models are tested. Then two models, three models and all models are combined to form a multimodal system. Experimental results of base classifiers, feature level fusion models and proposed ensemble methods are shown inTable 4. Experimental results of deep learning based model are shown in Table 5.FromTable 4, we can observe that facial expression modality has performed better compared to other modalities and audio being the worst.

In feature level fusion, 73.86% is the highest accuracy achieved by combining all the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

modalities. The proposed ensemble model achieves the highest actuary of 74.36%. Table 5 shows the results obtained using deep learning models. From results, we observe that 74.93% is the maximum accuracy achieved by combining all modalities with the text is the best and body gesture being the worst modalities. Also, we can observe that the performance of bimodal models, trimodal and all modality models is significantly better than unimodal, bimodal and trimodal models respectively. In both the cases, Model with all the modalities obtains the best classification accuracy, also the performance of text, audio, and video model is almost same as the model with text, audio, face, and body gesture (all modalities) model.Figure 2 shows the Experimental Results of Base Classifiers, Feature Level Fusion Models and Proposed Ensemble.Figure 3 and Figure 4 show the results of deep learning based models.

Table 4 Experimental results of base classifiers, feature level fusion models and proposed ensemble. Legend: SVM- Support Vector Machine, DT - Decision Tree, KNN – K-Nearest Neighbor, LR – Linear Regression, OWA - Optimal Weighted Averaging

Modality/Classifier	Feature Level Fusion					Ensemble Approach		
	SVM	DT	KNN	RF	LR	Voting	Averaging	ProposedOWA
Text	59.72	38.96	52.49	59.92	59.51	60.73	60.63	**61.24**
Audio	45.88	38.96	42.62	39.06	45.07	**45.17**	43.74	43.85
Face	63.89	47.91	62.36	63.07	64.60	**65.82**	64.60	64.19
Body Gesture (BG)	41.40	42.32	60.53	59.00	41.30	53.92	60.43	**61.85**
Text-Audio	**63.99**	39.17	42.01	48.73	**63.99**	58.90	62.26	63.17
Text-Face	**73.45**	48.02	63.17	62.26	71.82	71.52	70.70	72.33
Text-BG	62.26	40.49	60.83	57.48	62.05	**65.31**	64.50	64.50
Audio-Face	**67.96**	47.51	58.29	62.97	65.92	66.63	64.50	65.51
Audio-BG	50.97	39.98	54.63	49.95	50.05	**54.53**	51.07	51.17
Face-BG	66.23	47.61	63.07	66.53	66.63	**68.16**	65.62	66.02
Text-Face-Audio	71.31	47.61	58.49	60.83	72.23	71.52	71.41	**72.74**
Text-Face-BG	72.84	47.61	62.87	66.84	72.02	71.72	72.53	**73.04**
Text-Audio-BG	66.12	39.98	54.22	51.78	65.51	63.58	66.53	**67.24**
Audio-Face-BG	66.43	48.22	58.49	65.21	66.33	68.26	66.84	**67.04**
Text-Audio-Face-BG	73.86	48.32	58.49	64.70	73.25	73.14	73.55	**74.36**
Text-Audio-Video	**73.86**	48.32	58.49	63.58	73.25	72.53	72.43	73.45

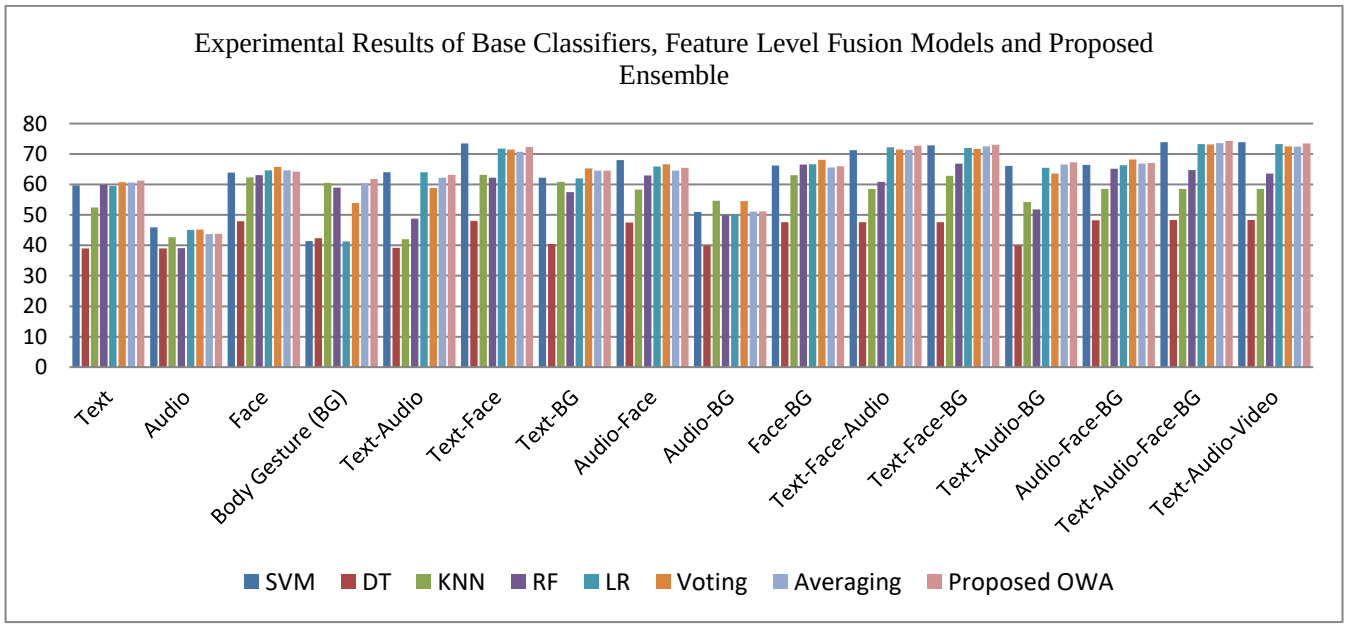

Figure 2 Experimental Results of Base Classifiers, Feature Level Fusion Models and Proposed Ensemble

Table 5 Experimental Results of Deep Learning Based Model

Modality / Model	Model-1	Model-2
Text	66.02	66.12
Audio	50.66	50.97
Face	48.99	49.39
Body Gesture (BG)	37.75	40.28
Text-Audio	66.28	
Text-Face	69.16	
Text-BG	66.90	
Audio-Face	67.58	
Audio-BG	58.35	
Face-BG	68.87	
Text-Face-Audio	72.52	
Text-Face-BG	73.86	
Text-Audio-BG	67.13	
Audio-Face-BG	66.96	
Text-Audio-Face-BG	**74.62**	
Text-Audio-Video	**74.60**	

.

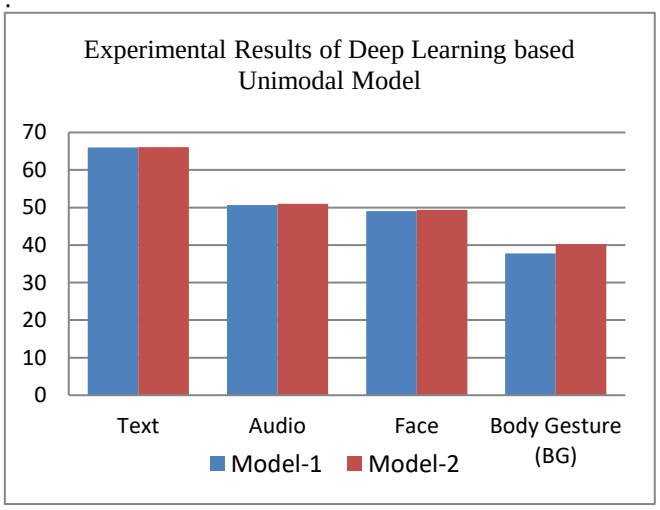

Figure 3 Experimental Results of Deep Learning based Unimodal Model

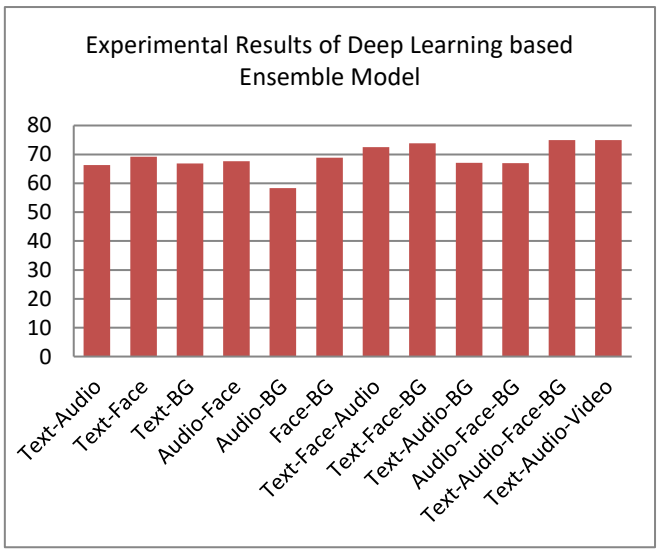

Figure 4 Experimental Results of Deep Learning based Ensemble Model

VI. CONCLUSION AND FUTURE WORK

One of the issues in multimodal emotion recognition is to know the type and number of modalities to consider, also their importance in overall classification accuracy. Different machine learning and deep learning based models were designed using one modality, two modalities, three modalities, and all modalities at a time. Similarly, the performance of bimodal models and trimodal models is better than unimodal and bimodal models respectively. The models including body gesture and facial expressions along with other modalities have performed significantly well compared to speech and text-based emotion recognition models. In the future, we explore the fusion techniques to merge emotional information extracted from multiple modalities and extracting contextual information among the utterances of a video.

REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, p. 98–125, 2017.

[2] M. Sreeshakthy and J. Preethi, "Sreeshakthy, M., and J. Preethi. "Classification of human emotion from deep EEG signal using hybrid improved neural networks with cuckoo search," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 6, no. 3-4, pp. 60-73, 2016.

[3] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] V. Chernykh, G. Sterling and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[6] K. Han, D. Yu and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[7] H. Beigi, "Beigi, H. (2011). Speaker recognition," *Springer US*, pp. 1232-1242, 2011.

[8] E. Lakomkin, C. Weber, S. Magg and S. Wermter, "Reusing neural speech representations for auditory emotion recognition," *arXiv preprint arXiv:1803.11508*, 2018.

[9] V. Perez-Rosas, R. Mihalcea and L.-P. Morency, "Utterance-Level Multimodal Sentiment Analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[10] J. G. Ellis, B. Jou and S.-F. Chang, "Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News," in *Proceedings of the 16th International Conference on Multimodal Interaction*, Istanbul, Turkey, 2014.

[11] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae and L.-P. Morency, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46 - 53, 2013.

[12] L. S. Chen, T. S. Huang, T. Miyasato and R. Nakatsu, "Multimodal Human Emotion/Expression Recognition," in *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, Washington, DC, USA, 1998.

[13] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60 - 75, 2017.

[14] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, 2013.

[15] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, p. 7–19, 2010.

[16] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, 2016.

[17] A. Zadehi, C. M, S. Poria, E. Cambria and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017.

[18] P. Effrey, R. Socher and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empir- ical methods in natural language processing (EMNLP)*, 2014.

AUTHORS PROFILE

Mahesh Huddar is Assistant Professor in Department of Computer Science and Engineering at Hirasugar Institute of Technology, Nidasoshi, Belagavi, India and he is currently pursuing Ph.D. at the Visvesvaraya Technological University, Belagavi, India in the Department of Computer Science and Engineering. He received his Master and Bachelor degrees from the Visvesvaraya Technological University, Belagavi, India in 2014 and 2008, respectively. He has published a good number of papers in journals, International, and National conferences.His main research interests include Machine Learning, Deep Learning, Multimodal Sentiment Analysis, Multimodal Emotion Recognition.He is a member of the IEEE.

Dr. Sanjeev S Sannakki has completed his Ph.D.degree in Image processing & Data Miningfrom VTUBelagavi. His career spans over a period of two decadesin the field of teaching,research and other diversifiedin-depth experience in academics. He is currentlyworking as a Professor in the Department of Computer Science and Engineering, GogteInstitute of Technology,Belgaum. Currently, he is shouldering the responsibility of Head of the Research center.He has published severalpapers in reputed national/international conferences and journals. He is alsoguiding the research scholars & UG/PG students of VTU.

Dr. Vijay S Rajpurohit is working as a Professor in the Department of Computer Science and Engg at Gogte Institute of Technology, Belgaum, Karnataka, India. He pursued his B.E. in Computer Science and Engg from Karnataka University Dharwad, M.Tech from N.I.T.K Surathkal, and Ph.D. from Manipal University, Manipal in 2009. His research areas include Image Processing, Cloud Computing, and Data Analytics. He has published a good number of papers in journals, International, and National conferences. He is the reviewer for a few international journals and conferences. He is the associate editor for two international journals and a Senior Member of the International Association of CS and IT. He is also the life member of SSI, ISC, and ISTE associations.