2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)
18-19 December 2017, WIT, Dehradun, India

Audio Visual Emotion Recognition Using Cross Correlation and Wavelet Packet Domain Features

Shamman Noor [1,*], Ehsan Ahmed Dhrubo [1,†],Ahmed Tahseen Minhaz [1,‡],Celia Shahnaz[1,§], Shaikh Anowarul Fattah [1,¶]
[1]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology Dhaka-1000, Bangladesh
*shammannoorshoudha@gmail.com †ehsan.ahmed.dhrubo@gmail.com ‡tahseenminhaz92@gmail.com §celia@eee.buet.ac.bd ¶fattah@eee.buet.ac.bd

***Abstract*—The better a machine realizes non-verbal ways of communication, such as emotion, better levels of human machine interrelation is achieved. This paper describes a method for recognizing emotions from human Speech and visual data for machine to understand. For extraction of features, videos consisting 6 classes of emotions (Happy, Sad, Fear, Disgust, Angry, and Surprise) of 44 different subjects from eNTERFACE05 database are used. As video feature, Horizontal and Vertical Cross Correlation (HCCR and VCCR) signals, extracted from facial regions - eye and mouth, are used. As Speech feature, Perceptual Linear Predictive Coefficients (PLPC) and Mel-frequency Cepstral Coefficients (MFCC), extracted from Wavelet Packet Coefficients, are used in conjunction with PLPC and MFCC extracted from original signal. For both types of feature, K-Nearest Neighbour (KNN) multiclass classification method is applied separately for identifying emotions expressed in speech and through facial movement. Emotion expressed in a video file is identified by concatenating the Speech and video features and applying KNN classification method.**

***Index Terms*—Horizontal and Vertical cross correlation, Perceptual Linear Predictive Coefficient(PLPC), Mel Frequency Cepstral Coefficient(MFCC) , Viola Jones Algorithm.**

I. Introduction

Emotion Recognition is a process of identifying human emotion, usually from speeches and facial expressions. Computational methodologies, leveraging techniques from multiple areas like signal processing, machine learning and computer vision, have been developed to identify automatically expressed emotions. Recognition of emotion enhances naturalness in human machine interaction (for example: on-board car driving system, autonomous call center services, interactive movie, story-telling, E-tutoring, autonomous psychological therapy etc.), speech-to-speech translation system, medical disorder diagnosis, indexing and retrieving Speech/video files based on emotions and so on. Emotion recognition from speech and images has been widely studied. In previous studies, Constraint Local Model (CLM) for face tracking, prosodic and spectral features, cross-modal relevance calculation, pitch and intensity have been used to derive weights. For pitch contour calculation, auto-correlation algorithm was used and blur insensitive LPQ (Local Phase Quantization) was used to detect facial emotion recognition in [1]. SVM classifier was used and accuracy of 47.6% (Unweight) and 62.9% (Weighted) was obtained. In [2], MFCC feature vectors, RBM based unsupervised pre-training and discriminative pre-training were used as features and Deep Neural Network Hidden Markov Models or DNN-HMMs, were used to classify and reported accuracy was 77.92%. Local Phase Quantization (LPQ), Mel-Frequency Cepstral Coefficients (MFCC) and relative spectral features (RASTA) based on perceptual linear prediction (PLP) were used as features and SVM was used as classifier in [3] and the accuracy was 76.4%. Sparse kernel reduced-rank regression (SKRRR), SVM classifier and SR classifier were used to detect emotion in [4] and their accuracy was 87.02% for SVM and 87.46% for SR.

In this paper, an efficient feature extraction algorithm, applied on database eNTERFACE05, is proposed to extract information relating emotions present in Speech and video. Facial regions like eye, eyebrows, mouth, nose and nasolabial-fold show separable characteristics for different emotions [5]. Viola Jones Algorithm can detect and separate these parts from images [6]. Horizontal and vertical cross correlation signals help distinguish facial images by providing detailed variations in face geometry along the vertical and horizontal directions, respectively [7]. In case of Speech signals, Wavelet Packet Coefficients, Perceptual Linear Predictive Coefficients and Mel-frequency Cepstral Coefficients [8] from overlapped frames are extracted, after exploiting pre-emphasis filtering [9] [10] [11], from each of the right, left and mono channels. Application of temporal smoothing and statistical function(mean) remove noisy components in features of windowed frames and consider all frames and their temporal evaluation. [11].

II. Proposed Method

A. Emotion Recognition from Visual data

1) Pre-Processing: Frames are collected from video files. As geometric shape variation is exploited in order to identify emotions present in videos, color information is not neede. So, these frames are converted from RGB scale to gray scale. Median filtering is applied to smoothen and denoise the images. Exploiting the short-term energy of speech signals windowed frames and a predetermined threshold value, silent frames, containing no speech content and therefore negligible information regarding emotion, are removed. This method of

978-1-5386-2621-4/17/$31.00 ©2017 IEEE

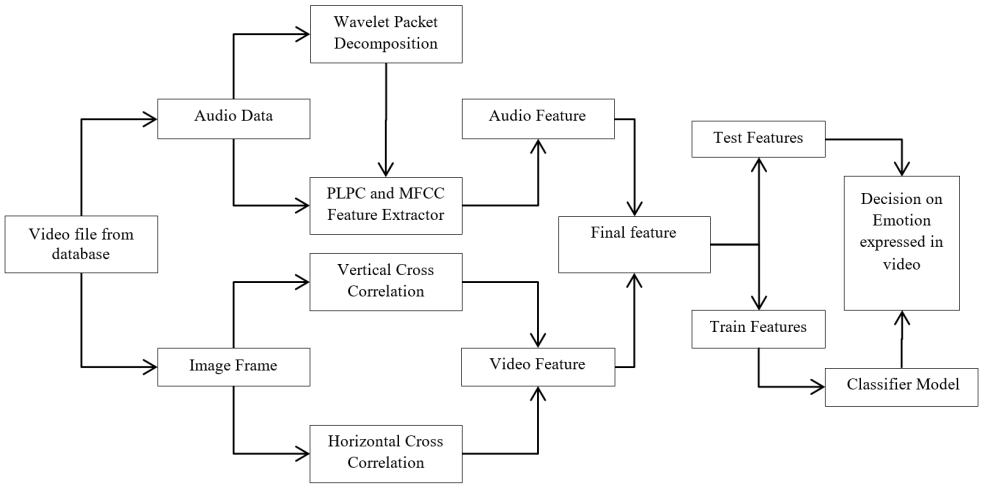

Fig. 1. Flow diagram of Audio-Visual Emotion Recognition Algorithm

silent frame removal reduces computational cost, redundancy and erroneous frames.

2) Horizontal and Vertical Cross-Correlation: As a particular type of emotion can be differentiated from others using unique shapes of eye and mouth, these facial regions are extracted. Left half of the extracted regions are selected for symmetrical property. Horizontal Cross Correlation (HCCR) and Vertical Cross Correlation (VCCR) signals of the images using 1 are computed, where R_{xy} is the cross-correlation vector, x and y are consecutive rows and columns for horizontal and vertical cross-correlation, respectively. Cross correlation is a measure of similarity between two series as a function of their relative displacement. Horizontal cross correlation signal is the cross correlation of two consecutive rows and vertical cross correlation is cross correlation of two consecutive columns.

$$R_{xy}(m) = \sum_{n=0}^{N-m-1} x_{m+n} y*_n \tag{1}$$

B. Emotion Recognition from Speech data

As the three different channels (left channel, right channel and average of left and right channel or mono channel) of Speech signal provide slightly different accuracies, all of them are used for feature extraction.

1) Pre-emphasis: Endpoint detection method detects the starting and ending points of a speech. This method is applied on the Speech signal which removes the silent region at the starting and ending of Speech signal, minimizing redundant Speech data remaining in the silent frames. The Speech signal is then passed through a pre-emphasis filter, H (a first order high pass filter) with pre-emphasis coefficient a = 0.9785 to emphasize information of formants and remove the impact of excitation source using 2. The filtered signal is then segmented into 25 ms frames with each overlapped by 10 ms, as within 10-25 ms, speech signal can be considered quasi-stationary. In order to prevent Gibbs phenomena, which introduces high frequency noise components due to sharp cut-off edges of window, these segments are windowed by Hamming window having tapered edges.

$$H(z) = 1 - az^{-1} \tag{2}$$

2) Perceptual Linear Predictive Coefficient: Perceptual Linear Predictive Coefficients (PLPCs) of length 13 are extracted applying Bark filter bank using 3 from each windowed frame, where B is Bark scale value and f is frequency in Hertz. Bark frequency scale represents the way human ear perceives frequency ranges and is useful for emotion related information extraction.

$$B = 13tan^{-1}\frac{0.76f}{1000} + 3.5tan^{-1}\frac{f^2}{7500^2} \tag{3}$$

3) Mel Frequency Cepstral Coefficient: Mel-frequency Cepstral Coefficients (MFCCs) of length 26, using 13 filters of Mel-Frequency band are extracted from windowed frames using 4, where m is Mel scale value of Hertz value f. Mel scale represents a different entity for human-ear frequency range perception mechanism from Bark scale. Both Bark and Mel scale is utilized here for extraction of detailed information regarding emotion.

$$m = 2595log(1 + \frac{7}{700}) \tag{4}$$

4) Post-Processing: Temporal smoothing or averaging filtering of length 3 (taking into account two previous and two following frames) is applied using 5 to each frames features as it removes any sudden changes in features due to noisy speech samples. Here, x_{sma} is the smoothed feature, x is the extracted feature, W is window length. Mean of all frames features is taken because effect of all frames and their temporal evolution are taken into account as emotions are expressed for longer durations than 25ms.

$$x_{sma}(n) = \frac{1}{W} \sum_{\frac{-(W-1)}{2}}^{\frac{(W-1)}{2}} x(n+i) \tag{5}$$

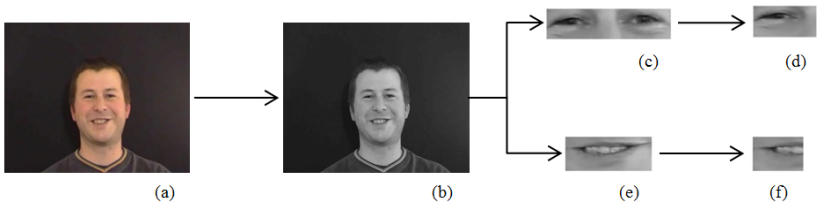

Fig. 2. Images illustrating image processing steps for extracting facial regions; (a) Original Image; (b) Gray-scale image; (c) Extracted Eye region; (d) Left half of Eye region; (e) Extracted Mouth region; (f) Left half of Mouth region

5) Wavelet Packet Decomposition: A different set of feature is extracted from the original Speech signal in this step. The Speech signal is down sampled from 48 kHz to 16 kHz for faster decomposition in wavelet domain and for less memory occupation. Wavelet Packet Decomposition up to level 3, using filters coiflet5 and daubechies10 is then applied and the signals of nodes (3,0), (3,1), (2,1), (1,1) from the decomposition tree are selected. Wavelet Packet Decomposition (WPD) presents another scale of perceptual frequency range. We have extracted PLPC and MFCC features, as before, from each of these four signals and concatenated them with the previously extracted features to obtain our final Speech feature vector. The resulting Speech feature length is 195, consisting of 13 PLPC features, 26 MFCC features and 156 PLPC and MFCC of WPD.

III. Dataset Representation

Database eNTERFACE'05, which includes 44 persons or subjects who gave demo videos of 6 classes of emotions (Happy, Sad, Disgust, Fear, Angry and Surprise) [12] is used to apply the developed algorithm and compute accuracy results. 5 sentences of each emotion from each of these subjects were recorded, providing 1320 videos for experiment.

IV. Results

The final feature, with length 429, is obtained from each video by concatenating the Speech and visual feature sets. 80%-20% hold-out validation method is used for training classification models with training data and testing with test data using K-Nearest Neighbor (KNN) multiclass classification method. In Speech, visual and Speech-visual emotion recognition, accuracy results of 69.83, 87.6% and 96.67% are obtained, respectively and are shown in Fig. 3. Comparison with previous works Datcu et al [13], Paleari et al. [14], Mansoorizadeh et al. [15], Gajsek et al. [16], Wang et al. [17], Jiang et al. [18], Huant et al. [19], Zhalehpour et al. [3] is shown in Fig. 4. Confusion matrices with accuracies of Speech, visual and Speech-visual emotion are shown in Table I, II, III, respectively.

TABLE I
CONFUSION TABLE OF EMOTION RECOGNITION RESULT: SPEECH EMOTION RECOGNITION (IN PERCENTAGE)

	Angry	Disgust	Fear	Happy	Sad	Sur
Angry	81.80	0.00	0.00	9.10	0.00	9.10
Disgust	11.11	77.78	0.00	11.11	0.00	0.00
Fear	12.5	12.5	50	12.5	0.00	12.5
Happy	0.00	8.33	0.00	83.3	0.00	8.33
Sad	0.00	7.70	0.00	15.4	69.23	7.70
Sur	0.00	8.57	0.00	14.28	0.00	57.14

TABLE II
CONFUSION TABLE OF EMOTION RECOGNITION RESULT: VISUAL EMOTION RECOGNITION (IN PERCENTAGE)

	Angry	Disgust	Fear	Happy	Sad	Sur
Angry	90.00	10.00	0.00	0.00	0.00	0.00
Disgust	0.00	90.00	0.00	10.00	0.00	0.00
Fear	0.00	5.00	80.00	5.00	0.00	10.00
Happy	0.00	0.00	0.00	90.00	0.00	10.00
Sad	0.00	10.00	0.00	10.00	80.00	0.00
Sur	0.00	0.00	10.00	0.00	0.00	90.00

TABLE III
CONFUSION TABLE OF EMOTION RECOGNITION RESULT: AUDIO-VISUAL EMOTION RECOGNITION (IN PERCENTAGE)

	Angry	Disgust	Fear	Happy	Sad	Sur
Angry	99.00	0.00	0.50	0.00	0.50	0.00
Disgust	0.00	98.00	0.00	0.70	0.00	1.30
Fear	2.20	2.20	97.00	0.00	0.80	0.00
Happy	2.50	0.30	0.00	95.00	2.20	0.00
Sad	0.00	0.60	0.27	0.13	99.00	0.00
Sur	3.30	0.00	1.70	0.00	3.00	92.00

V. Conclusion

In case of Speech features, combining three different non-linear frequency scales - Bark scale, Mel Frequency Scale and Perceptual scale of Wavelet Packet Decomposition - provides information of emotions in greater and finer details. In case of visual features, the extraction method of features used requires less computational cost and the features preserve geometric shape information of facial regions. These advantages allow gaining high accuracy in visual emotion recognition, which is greater than the state-of-art methods' visual emotion recognition accuracy using the same dataset [20].

References

[1] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1543–1552, 2013.

[2] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 312–317.

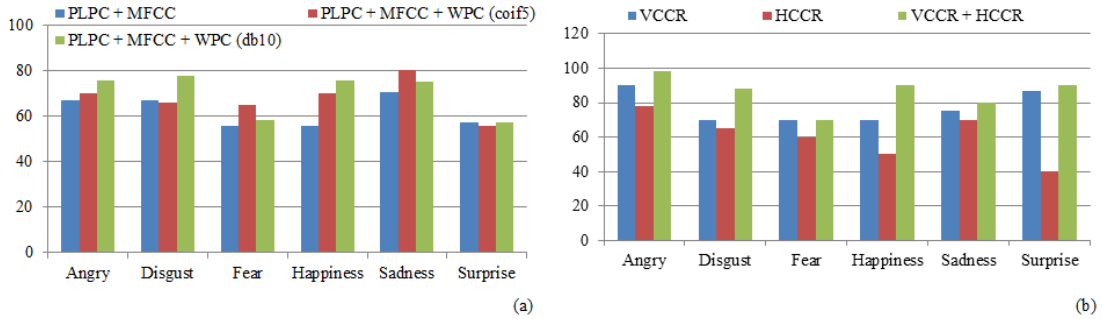

Fig. 3. Results of Audio-visual Emotion Recognition (a) Accuracy of Speech Emotion Recognition (b) Accuracy of Visual Emotion Recognition

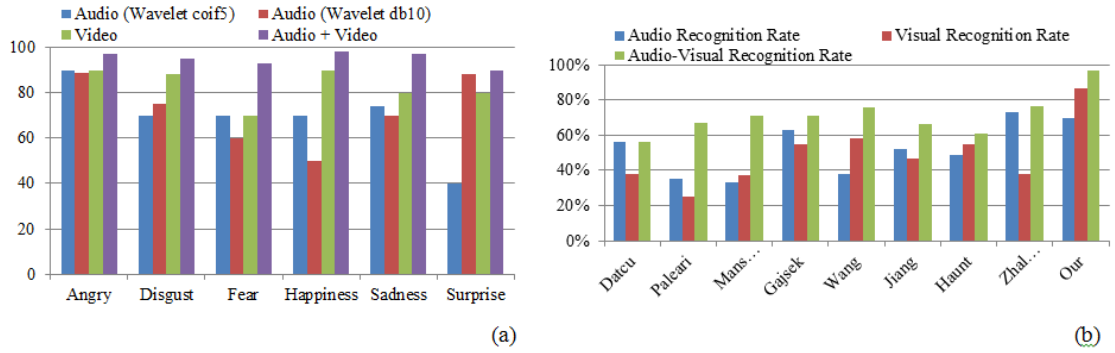

Fig. 4. (a) Accuracy of Audio Visual Emotion Recognition (b) Comparison to previous approaches

[3] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on*. IEEE, 2014, pp. 116–121.

[4] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319–1329, 2016.

[5] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[6] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, p. 14, 2003.

[7] S. A. Fattah, M. R. Khan, A. Sharin, and H. Imtiaz, "A face recognition scheme based on spectral domain cross-correlation function," in *TENCON 2011-2011 IEEE Region 10 Conference*. IEEE, 2011, pp. 10–13.

[8] A. Neustein, "Springerbriefs in electrical and computer engineering," 2001.

[9] E. Pavez and J. F. Silva, "Analysis and design of wavelet-packet cepstral coefficients for automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 814–835, 2012.

[10] C. Turner and A. Joseph, "A wavelet packet and mel-frequency cepstral coefficients-based feature extraction method for speaker identification," *Procedia Computer Science*, vol. 61, pp. 416–421, 2015.

[11] F. Eyben, "Acoustic features and modelling," in *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, 2016, pp. 9–122.

[12] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[13] D. Datcu and L. Rothkrantz, "Multimodal recognition of emotions in car environments," *DCI&I 2009*, 2009.

[14] O.-A. Schipor, S.-G. Pentiuc, and M.-D. Schipor, "Towards a multimodal emotion recognition framework to be integrated in a computer based speech therapy system," in *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*. IEEE, 2011, pp. 1–6.

[15] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[16] V. Štruc, F. Mihelic *et al.*, "Multi-modal emotion recognition using canonical correlations and acoustic features," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 4133–4136.

[17] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[18] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic bayesian network models," *Affective Computing and Intelligent Interaction*, pp. 609–618, 2011.

[19] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[20] H. Jin, Q. Chen, Z. Chen, Y. Hu, and J. Zhang, "Multi-leapmotion sensor based demonstration for robotic refine tabletop object manipulation task," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 104–113, 2016.