FEATURE FUSION IN MULTIMODAL EMOTION RECOGNITION SYSTEM FOR ENHANCEMENT OF HUMAN-MACHINE INTERACTION

Veni S1, Anand R2, DIVYA MOHAN3 AND ELDHO PAUL4
1Department of Electronics and Communications Engineering Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
2Assistant Professor, Research Member in Sona SIPRO, Department of ECE, Sona College of Technology, Salem & Research Scholar, Amrita Vishwa Vidyapeetham University, Amrita School of Engineering, Coimbatore.
3Assistant professor, Dept of ECE, Ahalia School of Engineering and Technology, Palakkad,India & Research Scholar, Amrita Vishwa Vidyapeetham University, Amrita School of Engineering, Coimbatore
4Assistant Professor, Research Member in Sona SIPRO, Department of ECE, Sona College of Technology, Salem

Abstract. Emotion Recognition (ER) systems is very much important for interpersonal relationship. Emotions are developed by some physiological changes. The straightforward of this effort is to discover the competence of language and facemask elements to deliver the feeling exact information for enhancing the Human-Machine interaction. The techniques and systems used in emotion detection may vary depending on the features inspected. Since both these features complement each other, combining them results in higher performance in terms of accuracy of 94.734%. The proposed system was tested on ENTERFACE’05 database and real time video. For Video, Speeded Up Robust Features (SURF) and Gabor features are used.

Keywords: Emotion recognition; Support Vector Machine (SVM); prosodic features; SURF and Gabor Features; Extraction of energy contour; Extraction of pitch contour

1. INTRODUCTION

Emotion theory is related with organization of some aspect of human behavior or performance [1]. It enables us to make predictions about the behavior or performance. Components of emotion include (i) Cognition (ii) Physiological (iii) Expressive (iv) Motivation and (v) Feeling. Recently, emotion recognition in human - computer interaction studies (HCI) is one of the topics that researchers are most interested in. To have a human like interaction, machines should be trained in such a way that it can identify the faces and recognize the emotions. During the training process of a machine the main challenge comes during the natural interface with the user. ER is very important because the meaning of the message can be altered by the emotions. Emotions are transparently identified from the face, body gestures, sound, and intensity of the voice. Hence, researchers have started to shift their studies toward spontaneous facial expression by fusing the speech expressions [2-4]. The fusion of features is performed either at the decision level or before classification. Multimodal approaches by combining the mode of interactions results in enhancing the understanding level and results in an efficient ER system in terms of better performance and robustness.

The summary of this work is prearranged as follows. Section II deals a review of related works and methodology used in this work. The proposed work is also presented in this Section for facial ER system, Speech ER system and multimodal system. Section III deals the Results and discussions, and final section concludes the paper.
2 MATERIAL AND METHOD

2.1 Related Works

Emotional expression is one of the basic expressions of the human. It motivates an action by adding a meaning and richness to the human experience. Basic emotional theory states that almost all ER system works on the universal emotions. However, in majority of real-life situations the system fails as it is different from the acted emotions. Existing studies about ER systems from video are discussed in this section which includes following dimensions (i) facial system, (ii) acoustic system and (iii) combined facial and speech system.

2.1.1 Facial ER system

Few other works using facial ER system employed methods such as optical flow technique[4] [5], [6][7], appearance model [8] and local parameterized models [9]. OF is also used by Rosenblum [10] [11] [12] to measure the facial region with Radial Basis Function (RBF) network as a classifier. Otsuka and Ohya [11] [13][14] used the OF to calculate the 2D Fourier transform coefficients which is applied to Hidden Markov Model (HMM) to classify the expressions [17].

2.1.2 Multimodal ER system

Extensive research was performed based on ER by using either facial or speech features in the past, now the researchers have moved towards fusion of audio and visual data for an efficient ER system. Decision level fusion [23] separately means every unimodal structure and combines the result at the end. For predicting the emotions effectively two unimodal systems are shared at the better dimension using facial, acoustic, facial, features. Few other studies were made on emotion recognition using facial features [24-25], speech features [26-29] and with combining both features [30-33]. Based on the above studies, it is observed that design of a multimodal system has significant advantages in enhancing human machine interaction. With this motivation, fusion of features (facial and speech) is attempted in this work to enhance the human-machine interaction. The proposed system in this study with the fusion of facial and spectral features is providing better results in terms of accuracy. The details are elaborated in the next sections.

2.2 Methodology

In this work, a multimodal scheme for programmed ER is proposed. Figure. 1 depicts the whole process of machine learning emotion recognition used in this study. The stages include (i) Obtaining emotional face and speech database (ii) Feature selection and extraction (iii) Classification. Initially, two unimodal systems are developed by exploring the facial and speech features. The description of these features is elaborated in the following sections. Datasets used is ENTERFACE’05 database. In addition, the developed system was tested with real time video data.

![Figure. 1 Machine learning approach for emotion recognition system](image)

2.2.1 Feature Extraction

2.2.1.1 Prosodic Features

One of the significant features of the speech is the energy contour. It is used to differentiate the unvoiced region and voiced region of the speech. Vowel sound mainly serves high energy contour. Since speech signal is stationary in nature, it has to be segmented by multiplying with a window function. Most preferred window for segmenting is hamming window because it gives the higher weight to middle samples. After segmentation, squaring is done which yields the energy signal. These energy values are plotted against the time domain to get the energy contour. Fundamental frequency of the speech signal is referred as pitch. Pitch contour mainly refers to a curve that track the pitch in a given speech signal. It includes variety of sounds utilizing multiple of pitches and also relate to the frequency function at one point to a later point. Resonance frequency of human vocal tract is referred to as formant.
2.2.1.2 Spectral Features

Spectral features are mainly used to capture the information based on movement of articulare, shape and size of vocal tract. In the proposed work, MFCC features are used to extract the emotional information. They are the coefficients of Mel Frequency Cepstrum which is derived from power spectrum. In this work for each audio file 13 coefficients are derived. After taking the mean and variance of these coefficients it is fed into the feature extraction part.

2.2.1.3 Gabor Features

Properties of image are mainly defined by the features extracted out of it, which is then used for the identification of the image in next phase. Gabor filters are used in this work to extract the features with 5 scales and 8 orientations [34, 35].

Gabor wavelet is defined as

\[ \Psi(z) = \frac{P_{a,b}}{\sigma^2} \exp\left[ -\frac{P_{a,b}^2}{2\sigma^2} \right] \left\{ \exp\left( iP_{a,b} Z \right) - \exp\left( -\frac{\sigma^2}{2} \right) \right\} \]  (1)

Here \( z = (x, y) \) which is the input image, \( a \) denotes the orientation and \( b \) denotes the scale of the Gabor filter. \( P_{a,b} = P_b e^{b \theta} \) Where \( P_b = P_{max} / f_b \) and \( f_b = \pi b / 8 \). \( P_{max} \) gives the maximum frequency and the spacing factor is termed by \( f \).

![Figure 2. Real and imaginary part of Gabor Wavelet](image)

Gabor filter mainly has two parts: the real part and imaginary part. The real part is transformed to get the magnitude feature and the phase feature is derived from the imaginary part. For every image there are 40 Gabor features. Computation of Gabor wavelet features is more complex and also the redundancy is high due to the convolution of Gabor wavelet with facial images. Statistical parameters are computed such as mean, standard deviation over feature vectors to reduce the problem of computational complexity. The real and imaginary part of Gabor wavelet used is shown in Figure 2.

2.2.1.4 SURF Features

The SURF detector is based on Gaussian derivative filters which is used to locate the features. By convoluting the basis image with the factor of the Hessian (DoH) matrix (Equation 2). The metric obtained by this method is further divided by the Gaussian’s variance, \( \sigma^2 \), to normalize its response:

\[ \text{DoH}(x, y, \sigma) = \frac{G_{xx}(x, y, \sigma) G_{yy}(x, y, \sigma) - G_{xy}(x, y, \sigma)^2}{\sigma^2} \]  (1)

where \( G_{ij}(x, y, \sigma) = \frac{\partial^2 N(0, \sigma)^2}{\partial i \partial j} \cdot \text{image}(x, y) \)  (2)

By blurring the image the Gaussian filters are removing the noisy data, hence many feature detection schemes rely on Gaussian filter. SURF beats the existing techniques in terms of fast performance, robustness, repeatability, and distinctiveness. The algorithm mainly has two parts (i) Detector and (ii) Descriptor. Key point in the input image is located by the detector whereas feature extraction and the construction of the feature vector are performed by descriptor. For detecting the key points it is necessary to construct the scale space which is commonly a pyramid scale space. For describing the features of key points SURF uses the response of Haar wavelets. These wavelets are preferred because of its robustness and less computation time. In order to extract the descriptor, first step is to construct a square region which is centered at the key point. This region is then split into smaller sub-region, and every sub-region response of Haar wavelet is computed. To obtain the first entries of the feature set, wavelet responses are summed up, and for the intensity changes absolute values of the responses are calculated. The results of the SURF features extracted for the ENTERFACE’05 database is shown in Figure 3.
For classification SVM is used in this work which is a classification prediction that use theory space of a direct roles in a extraordinary dimensional feature vector [36]. Using SVM, it is possible to map linearly non-separable features into linearly separable filters[37]. In this work, RBF kernel and polynomial kernels were used and the results are compared. The next section elaborates proposed methods for ER system using facial features, acoustic features, and combined features.

2.3 Description of proposed multimodal system

2.3.1 Facial Emotion Recognition System

The projected work for the facial feeling recognition is depicted in Figure 4. Three key points in the facial ER systems are detecting the face, extracting feature and the classification. The response given to the detection stage in the shape of an image, so that the video input data is converted into frames.

Then, this algorithm [38] is used over every framework to get the region of interest. The output of the face recognition and the ROI detection is shown in Figure 5.

In the next stage, Algebraic features cruel and standard deviation were calculated for respectively responses of the Gabor filter. Next SURF features are extracted from the entire face. To lessen the dimensionality dilemma, mean of the features were calculated which caused in the feature vector length of 64. Classification was carried out using SVM with four different emotions: anger, disgust, fear and happiness.
2.3.2 Speech ER system

The proposed system for emotion recognition using speech is depicted in Figure 6. Recognition of emotion from speech, spectral and prosodic features is employed. Studying the spectral behavior, MFCC is explored while pitch, energy and formants are utilized to find the prosodic information which is shown in Figure 7.

2.3.4 Extraction of energy contour

Next feature extracted is the energy contour, and the steps for extraction are described in Figure 8. The dissimilar steps in discovery the frequency are given below (i) Pre-emphasize of input speech signal and normalization to 1 and -1 amplitude (ii) Segmentate of voice signal by getting after applying hamming window (iii) Removal of DC values using Butterworth high pass filter (iv) Computation of the logarithm of the spectrum for the visibility of all the values. Based the proposed systems for emotion recognition using facial and speech features, a multimodal system can be designed as shown in Figure 10.

3. RESULTS AND DISCUSSIONS

Initially face ER system was developed and two features are extracted namely Gabor and SURF. The method of extraction of these features is explained in section II. As an initial step Gabor features are stemmed from left eye, right eye and mouth and fed to the classifiers. The accuracy obtained for Gabor features by polynomial kernel is 68.42% and RBF kernel is 47.36%. As a second step, the Face system is made by SURF features. It gives accuracy as 63.15% for polynomial kernel and 47.36% for RBF kernel.
As a third step, SURF and the Gabor features are combined which gives improved accuracy as 66.66% using RBF kernel and 90.47% using polynomial kernel. It was observed from the results that polynomial kernel outperforms well for the classification using combined features (Table 1). Next stage after face ER system, speech ER system was developed. The MFCC features of the speech signal is extracted which yields a system with an accuracy of 63.15% for polynomial kernel. For speech system, classification using RBF kernel is giving better results using prosodic features whereas polynomial kernel gives better results for spectral features [42]. Fusion of these features increases the overall performance of the system in terms of accuracy using polynomial kernel up to 84.21%. Table 1 gives the accuracy of different systems and its comparison by using different kernels. From the Table 2, it was observed that the data set of DEAB is giving the highest accuracy of 94.73% compared with unimodal face and speech ER systems. The details of fusion of data considered are elaborated in Figure 11.

### Table 1 Comparison of two kernels

| Features                  | Efficiency by RBF kernel | Efficiency by Polynomial kernel |
|---------------------------|--------------------------|---------------------------------|
| Prosodic features         | 89.47%                   | 78.94%                          |
| Spectral features         | 47.36%                   | 63.15%                          |
| Prosodic + Spectral features | 73.68%                   | 84.21%                          |
| SURF features             | 47.36%                   | 63.15%                          |
| Gabor features            | 47.36%                   | 68.42%                          |
| SURF + Gabor features     | 66.66%                   | 90.47%                          |
| Face system + speech system | 89.47%                   | 94.73%                          |

### Table 2. Accuracy results for multimodal system with various fusion combinations

| Training Data | Testing Data | Face system | Speech system | Multimodal system |
|---------------|--------------|-------------|---------------|-------------------|
| ABCD          | E            | 84.21%      | 73.63%        | 89.47%            |
| BCDE          | A            | 78.94%      | 73.63%        | 84.21%            |
| CDEA          | B            | 78.94%      | 78.94%        | 89.47%            |
| DEAB          | C            | 90.47%      | 84.21%        | 94.73%            |
| EABC          | D            | 84.21%      | 78.94%        | 90.47%            |

Table 3 shows the confusion matrix of face system. The overall performance for this system is 90.47%. Anger, fear and happiness are correctly classified whereas the disgust is often confused with anger. Although, disgust is correctly classified by 90.42% however, almost 9.523% misclassification is happening for this emotion.

![Figure 11. Combinations of Training and Testing data used in the proposed system](image-url)
Table 3 confusion matrix for face ER system

|       | Anger | Disgust | Fear  | happiness |
|-------|-------|---------|-------|-----------|
| Anger | 100%  | 0       | 0     | 0         |
| Disgust | 0   | 100%    | 0     | 0         |
| Fear  | 0     | 0       | 100%  | 0         |
| Happiness | 5.263% | 0 | 0 | 94.736% |

Table 4. Confusion matrix for speech ER system

|       | Anger | Disgust | Fear  | happiness |
|-------|-------|---------|-------|-----------|
| Anger | 100%  | 0       | 0     | 0         |
| Disgust | 9.523% | 90.47% | 0     | 0         |
| Fear  | 0     | 0       | 100%  | 0         |
| Happiness | 0 | 0 | 0 | 100% |

Table 4 shows the confusion matrix of speech system. The overall performance for this system is 84.21%. Anger, fear and disgust [43] are correctly classified whereas the happiness is often confused. Happiness is correctly classified by 84.21% but it is misclassified with anger and disgust by 10.52% and 5.26% respectively.

Table 5. Confusion matrix for multimodal ER system

|       | Anger | Disgust | Fear  | happiness |
|-------|-------|---------|-------|-----------|
| Anger | 100%  | 0       | 0     | 0         |
| Disgust | 0   | 100%    | 0     | 0         |
| Fear  | 0     | 0       | 100%  | 0         |
| Happiness | 10.52% | 5.26% | 0 | 84.21% |

Table 5 shows the confusion matrix of proposed system. The general performing for this system is 94.736%. Anger, fear and disgust are correctly classified whereas the happiness is often confused. Happiness is correctly classified by 94.736% but it is misclassified with anger by 5.623%. The three systems are then tested for real time video data set collected (Figure. 12). The two emotions happiness and anger are considered and classified by using SVM classifier. It is possible to obtain 80% accuracy using the proposed multimodal system. To increase the quality of the system, number of training data is to be increased which is the future scope of the work. The result of the real time data classification is given in Table 6. The given system is associated with the current methods and the results are given in Table 7.

Figure 12 Detection of ROI from real time Video

Table 6. Real time emotion classification

|       | Accuracy |
|-------|----------|
| Speech | Speech system | 55.55% |
| Face system | 77.7% |
| Multimodal System | 80% |
Table.7 Comparison with existing method

| Previous Studies                          | Maximum Accuracy Obtained | Proposed Method                                  | Accuracy  |
|-------------------------------------------|---------------------------|--------------------------------------------------|-----------|
| Face System by Gabor Features            | 85.71%                    | Face System by Gabor + SURF Features              | 90.47%    |
| Speech System by Spectral Features       | 80.32%                    | Speech System by Spectral and Prosodic Features  | 84.21%    |
| Multimodal system by Morphological and Spectral Features | 93.62% | Multimodal System                               | 94.73%    |

4 CONCLUSIONS

A multimodal framework on ER system is proposed in this work. Then a combined system was developed, and results reveal that there is a marginal improvement in the performance with 94.73%. It was observed that polynomial kernel outperforms well in the classification stage of all the proposed systems. The proposed system was tested on real time video and it is possible to obtain the accuracy of 80% which can be enhanced the quality and number of training data. Feature level fusion is employed in this work and it can be extended to physiological signals with real life scenario instead of staged performance. Gestures hear beat and skin conductivity analysis may also give a better system [39]. Using advanced machine learning techniques and using deep convolutional networks it will be possible to overcome the limitations mentioned in the proposed work [40, 41].

REFERENCES

[1] Paul Kleinman (2012) Psych 101 - Psychology Facts, Basics, Statistics, Tests, and More.
[2] K. Sreenivsa Rao and Shashidhar G. Koolagudi (2015) ‘Recognition of emotions from video using acoustic and facial features’ International Journal on Signal Image and Video Processing, Vol. 9, Issue. 5, pp. 1029-1045.
[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G, (2001) ‘Emotion Recognition in Human-Computer Interaction’ IEEE Signal Processing Magazine, pp. 32-80.
[4] Simon Dobrisek, Rok Gajsek, France Mihelic, Nikola Pavesic, and Vitomir Struc (2013) ‘Towards Efficient Multi-Modal Emotion Recognition’ International Journal on Advanced Robotic System, Vol. 10, pp.1-10.
[5] P. Ekman, W. Friesen, and J. Hager (2002), The Facial Action Coding System: A Technique for the Measurement of Facial Movement. A Human Face.
[6] AlMejrad A and S Masses (2010) ‘Human emotions detection using brain wave signals: A challenging’ European Journal of Scientific Research, Vol. 44, pp. 640-659.
[7] Jibi Raj and Sujith Kumar (2015) ‘Gender based Affection Recognition of Speech Signals using Spectral & Prosodic Feature Extraction’ International Journal of Engineering Research and General Science, Vol. 3, Issue.2, pp.898-905.
[8] Lanitis, C. Taylor, and T. Cootes (1995) ‘A unified approach to coding and interpreting face images’ Proc. International Conf. on Computer Vision, pp. 368-373.
[9] M. Black and Y. Yacoob (1995) ‘Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion’ Proc. International Conf. on Computer Vision, pp. 374-381.
[10] M. Rosenblum, Y. Yacoob, and L. Davis, (1996) ‘Human expression recognition from motion using a radial basis function network architecture’ IEEE Trans. on Neural Network, Vol.7, No.5, pp. 1121-1138.
[11] T. Otsuka and J. Ohya (1997) ‘Recognizing multiple person’s facial expressions using HMM based
on automatic extraction of significant frames from image sequences’  
*Proc. International Conf. on Image Processing*, pp. 546-549.

[12] Cohen, N. Sebe, A. Garg, L. S.Chen, and T. Huang (2003) ‘Facial expression recognition from video sequences: Temporal and static modeling’, *International Journal on Computer Vision and Image Understanding*, Vol.91, Issue (1-2), pp. 160-187.

[13] L. Chen (2000) ‘Joint processing of audio-visual information for the recognition of emotional expressions in human computer interaction’, Ph.D Thesis, University of Illinois at Urbana-Champaign.

[14] Chang Y, Hu C, Feris R and Turk M (2006) ‘Manifold based analysis of facial expression’, *Journal on Image and Vision Computing*, Vol. 24. No.6, 605-614.

[15] Pantic and Patras (2006) ‘Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences’, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, No.6, pp. 636-642, 1996.

[16] Siqing Wu, Tiago H. Fulk and Wai-Yip Chan (2010) ‘Automatic speech emotion recognition using modulation spectral features’, *Journal of Speech Communication*, Vol. 53, pp. 768-785.

[17] Huang, F. Thawn and L. Didaci (2009) ‘Bimodal emotion recognition by man and machine’ *International Journal on Pattern Recognition*, Vol. 42, No. 11, pp. 2807-2817.

[18] Y. Yacoob and L. S.Davis (1996) ‘Recognizing human facial expressions from long image sequences using optical flow’, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, No.6, pp. 636-642, 1996.

[19] Y. Sagisaka, N. Campbell, and N. Higuch (1997) *Computing Prosody*, Springer-Verlag, New York.

[20] Weiyi Ma and William Forde Thompson (2015) ‘Human emotions track changes in the acoustic environment’ Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol.112, No.47, 2015.

[21] C. Chiu, Y. Chang and Y. Lai (1994) ‘The analysis and recognition of human vocal emotions’ Proc. International Computer Symposium, pp. 83-88.

[22] V. Petrushin (1998) ‘How well can people and computers recognize emotions in speech?’ *Proc. AAAI Fall Symposium*, pp. 141-145.

[23] Thushara.S and S.Veni (2016) ‘A Multimodal Emotion Recognition from Video’, *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*.

[24] Lee C.M and Narayanan S.S (2005) ‘Coupled Gaussian process regression for pose-invariant facial expression recognition’, *Proceedings of 11th European Conference on Computer Vision*.

[25] Filko and Yegnanarayana (2008) ‘Emotion recognition system by a neural network based facial expression analysis’, *AUTOMATIKA 54*, No. 109, pp. 44-55.

[26] Nicholson H and Pantic M (2010) ‘Emotion recognition in speech using neural networks’, *6th International Conference on Neural Information Processing (ICONIP)*.

[27] New O and Selvi M (2010) ‘Speech emotion recognition using hidden Markov models’, *IEEE Trans. Speech Audio Processing*, Vol. 18, pp. 965-973.

[28] J. Nicholson, K. Takahashi and R. Nakatsu (2000) ‘Emotion recognition in speech using neural networks’, *Journal of neural computing, springer Verilog*, Vol. 9, Issue 4, pp 290-296.

[29] Go H.J, Kwak KC, Lee DJ, and Chun MG (2003) ‘ Emotion recognition from facial image and speech signal’, *Int. Conf. of the Society of Instrument and Control Engineers, Fukui*.

[30] Chen. C, Deng. Z, Yildirim. S, Bulut. and Lee. C.M.(2004) ‘Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information’, *Int. Conf. Multimodal Interfaces, pp. 205-211*.

[31] Yoshitomii H, Kim P, Suteau A, and Allain P, (2010) ‘Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face’, IEEE International Conference, pp. 1-8.

[32] Datcu D and Rothkrantz L (2008) ‘Semantic audio-visual data fusion for automatic emotion recognition’, *Proceedings of Euromedia*.

[33] Mina Navran and Nasrollah Moghadam Charkari, (2014) ‘Fusion of Feature Sets for Facial Expression Recognition’, IEEE Transactions on Telecommunication, Vol. 62, No. 6.

[34] Mohamed Dahmane and Jean Meunier (2012) ‘Sift-flow registration for facial expression analysis using Gabor wavelets’, *11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp.175-180.
[35] Vinay.A, Vinay.S Shekhar, K.N BalaSubramanya Murthy, S. Natarajan (2015) ‘Face Recognition using Gabor Wavelet Features with PCA and KPCA - A Comparative Study’, 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), Procedia Computer Science, Vol. 57, pp. 650 – 659.

[36] Tom Mitchell (1997), Machine Learning. McGraw-Hill Computer science series.

[37] Chih-Feng Chao and Ming-Huai Horng (2015), "The Construction of Support Vector Machine Classifier Using the Firefly Algorithm, Computational Intelligence and Neuroscience, Article ID 212719.

[38] Yi-Qing Wang (2014) ‘An Analysis of the Viola-Jones Face Detection Algorithm’, Journal of Image Processing On Line (IPOL), pp.128-148.

[39] W.O.A.S. Wan Ismail, M. Hanif, S.B. Mohamed, Noraini Hamzah, and Zairi Ismael Rizman, “Human Emotion Detection via Brain waves Study by Using Electroencephalogram (EEG),” International Journal Advance Science Engineering Information Technology (IJASEIT), Vol. 6, No. 6, 2016.

[40] Poria, S, Cambria, E. and Gelbukh, A (2015) ‘Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis’, Proceedings of EMNLP, pp. 2539-2544.

[41] S. Veni and S. Thushara, “Multimodal Approach to Emotion Recognition for Enhancing Human Machine Interaction - A Survey”, International Journal Advance Science Engineering Information Technology (IJASEIT), Vol. 7, No. 46, 2017.

[42] Veni S., Anand R., Santosh B. (2021) Road Accident Detection and Severity Determination from CCTV Surveillance. In: Tripathy A., Sarkar M., Sahoo J., Li KC., Chinara S. (eds) Advances in Distributed Computing and Machine Learning. Lecture Notes in Networks and Systems, vol 127. Springer, Singapore. https://doi.org/10.1007/978-981-15-4218-3_25

[43] Anand R., Shanthi T., Nithish M.S., Lakshman S. (2020) Face Recognition and Classification Using GoogleNET Architecture. In: Das K., Bansal J., Deep K., Nagar A., Pathipooranam P., Naidu R. (eds) Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol 1048. Springer, Singapore. https://doi.org/10.1007/978-981-15-0035-0_20