IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8837

MULTIMODAL EMOTION RECOGNITION BASED ON FEATURE FUSION FOR ENHANCEMENT OF HUMAN-COMPUTER INTERACTION

D.Saisanthiya[1], P. Supraja[2]

[1]Research scholar, Department of Computer science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu , India 603203

[2]Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India 603203

Abstract: Emotions are intrinsically part of human mental activity and play a key role in human decision handling, interaction and cognitive processes. Recognizing emotion is an essential step to have complete interaction between human and machine. Emotion Recognition (ER) frameworks is especially significant for a human personal relationship. Feelings are created by some physiological changes. The clear view of this exertion is to find the capability of language and facemask components to convey the inclination precise data for improving the Human-Machine interaction. The methods and frameworks utilized in emotion detection may differ depending on the features reviewed. The combination of features is performed either at the decision level or the previous arrangement. Multimodal approaches by consolidating the method of cooperation brings about upgrading the level and outcomes in a productive ER framework as far as better execution and power. Since both these features compete one another, consolidating them brings about better as far as precision of 94.843%. The proposed framework was tried on ENTERFACE dataset and ongoing video. For Video, Speeded Up Robust Features (SURF) and Gabor highlights are utilized.

Keywords-: Emotion recognition; Multimodal Approach; Support Vector Machine (SVM); SURF and Gabor features;

1. INTRODUCTION

Emotion hypothesis is connected with association of few part of human conduct or execution [1]. It empowers us to make expectations about the conduct or execution. Segments of feeling incorporate (I) Comprehension (ii) Physiological (iii) Motivation and (iv) Feeling. As of late, feeling recognition in human - computer interaction considers (HCI) is one of the points that scientists are most keen on. The combination of highlights is performed either at the choice level or before order. Multimodal approaches by joining the method of connections brings about upgrading the arrangement level and results in a productive ER framework as far as better execution and strength. Emotion Recognition assumes a critical part in human-computer connection and has been examined for a long time. There are a few kinds of passionate information methodology, including look, body development and motions, discourse and so forth Among these information modalities, sound video recording enjoys a benefit of non-contact which prompts its boundless application. So sound video, based Emotion recognition has drawn to numerous researches eye. Emotions are vital during the time spent on human dynamic, association and perception. With the innovative technology and the developing of our comprehension of emotions, the interest in automatic emotion recognition frameworks is likewise expanding. Nonetheless, because of the intricacy and variety of human feeling articulations, in the event that one considers a specific articulation structure and judges human feelings, the eventual outcome is uneven and not unbiased, which will prompt numerous important feelings Information is lost.

2. MATERIAL AND METHOD

2.1 Related work

Emotion recognition is the fundamental articulations of a human being. It inspires an activity by including an importance and lavishness to the human experience. Essential enthusiastic hypothesis expresses that practically all ER framework chips away at the all inclusive feelings. Nonetheless, in larger part of genuine circumstances the framework flops as it is not the same as the feeling what they acted. Existing examinations about ER frameworks from video are examined in this segment which incorporates following measurements (I) facial framework, (ii) acoustic framework and (iii) consolidated facial and speech framework.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8837

2.1.1 Facial ER Framework

Hardly any different works utilizing facial ER frameworks utilized strategies, for example, optical stream technique[3] [4], appearance model [5], [6][7], and nearby defined models [8]. OF is additionally utilized by Rosenblum [9] [10] to quantify the facial area with Radial Basis Function (RBF) network for classification. Otsuka and Ohya [11][12] utilized the OF to ascertain the 2D Fourier change coefficients are applied to Hidden Markov Model (HMM) to arrange the articulations [13]. Well's additionally utilized in blend with SVM as Serial Multiple Classifier System to get best outcomes for discourse feeling acknowledgment. As SVM straightforwardly gives a grouping all things being equal of a score, HMM's can be utilized for preparing[14] the examples and SVM for arrangement. Alongside different classifiers, boosting can likewise be utilized as a method for fostering a solid order framework where at least two feeble classifiers are consolidated to shape a solid classifier. The paper [15] likewise discusses implanting HMM, for example fostering a two-dimensional HMM, comprising of super states and implanted states. The information is demonstrated in two ways by super states and installed states. For face pictures, start to finish highlights can be super states and right - to-left highlights can be implanted states.

2.1.2 Multimodal ER Framework

Expanded exploration was dependent on ER by utilizing either facial or speech highlights before, presently the analysts have moved to the combination of sound and eye information for an effective ER framework.

Decision level combination [15] independently implies each unimodal construction and joins the outcome toward the end. On anticipating the feelings adequately two or more unimodal frameworks are shared at the best measurement utilizing facial, acoustic, facial, highlights. Scarcely any different examinations were made on feeling acknowledgment utilizing facial highlights[16], speech highlights[17], and consolidating the two highlights. In light of the above examinations, it is seen that the plan of a multimodal framework enjoys huge benefits in upgrading human-machine collaboration. With this inspiration, a combination of highlights (facial and discourse) is endeavoured in this work to improve the human-machine collaboration. The proposed framework in this examination with the combination of facial and spectral features is giving better outcomes as far as precision. The subtleties are explained in the following segments.

2.2 Methodology

This multimodal conspire for customized ER is proposed the proposed work. Figure. 1 portrays the entire interaction of AI feeling recognition utilized in this investigation. The levels incorporate (I) Obtaining passionate face and speech data set (ii) Feature determination and extraction (iii) Classification. At first, two unimodal frameworks are created by investigating the facial representation features and speech highlights.

Fig.1. Deep learning approach for emotion recognition system

The portrayal of these highlights is expounded in the accompanying areas. Datasets utilized is ENTERFACE'05 information base. Also, the created framework was tried with continuous video information.

2.2.1 Extraction of features

Prosodic features

The important and main highlights of speech is in the form of energy. This is utilized to separate the unvoiced district and voiced area of the discourse. High energy form came from the vowel sound. As the disclosure signal is fixed in natural view, it must be fragmented by increasing of window work. The most favored window for dividing is hamming window since it gives a higher load to center examples. After division, figuring out is done which yields the signal in the form of energy. These energy esteems are plotted against the time-space to get the energy shape. The principal recurrence of the speech signal is alluded to as pitch. Throw shape for the most part alludes to a curveball that tracks the contribute a given discourse signal. It incorporates assortment of sounds using numerous of pitches and furthermore identify with the recurrence work at one highlight a later point. The reverberation recurrence of the vocal parcel human is alluded to as each of several prominent bands of frequency that determine the phonetic quality of a vowel.
Spectral highlights:

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8837

Spectral highlights are essentially used to catch the data dependent on the development of well-spoken, shape, and size of the vocal lot. In the proposed work, MFCC highlights are utilized to separate passionate data. These are coefficients of Mel Frequency Cepstrum that is gotten from the power range. In this work for every sound document, 13 coefficients are inferred. In the wake of taking the mean and change of these coefficients, it is taken care of into the element extraction part.

Gabor highlights:
Features of the picture are principally characterized by the highlights separated out for fit, this is utilized for the ID of the picture in the following stage. Gabor channels are utilized in this work to extricate the highlights having 6 scales and 9 directions [18][19].

$$\Psi(Z) = Pa,b\ 2\ \sigma 2\ \exp[-Pa,b\ 2\ Z2\ 2\sigma 2\ ]\ [\exp(iPa,bZ) - \exp(-\sigma 2\ 2\ )]$$

$z = (x, y)$- input image; a – orientation;b-scale of the Gabor filter.

$Pa,b = Pb\ e^{\phi}b$,$Pb = \frac{pmax}{fb}$ and $\phi b = \Pi\frac{b}{8}$. ,$pmax$ -max frequency,f-spacing factor.

SURF highlights:

SURF identifier depends on Gaussian subordinate channels which is utilized to find the highlights. By tangling the premise picture with the factor of the Hessian (DoH) network. The measurement got by the strategy is additionally partitioned by the difference in Gaussian function, σ2, to standardize its reaction:
DoH(a, b, σ) = Gxx (a, b, σ). Gbb (a, b, σ) − Gab (a, b, σ)2 /σ2
where Gij(a,b, σ) =∂N(0,σ) 2 / $\partial i.\partial j$. image(a,b)

Fig.2. SURF image extracted

For characterization, SVM is utilized, this is a grouping forecast which utilizes hypothesis space of direct jobs are in a remarkable dimensional component vector. Utilizing SVM, in that it is feasible to plan straightly non-distinct highlights into directly distinguishable filters. This proposed work, RBF portion and parts of polynomial are utilized and the outcomes are analyzed. The following area expounds on proposed strategies for ER framework utilizing facial highlights, acoustic highlights, and joined highlights.

2.3 Proposed System
2.3.1 Facial ER Framework

The proposed work for facial recognition of feeling is portrayed in Figure 3. Totally 3 central issues in the facial frameworks are recognizing the face, removing highlight and the arrangement. The reaction given to the location stage looking like a picture, with the goal that the video input information is changed over into outlines.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8837

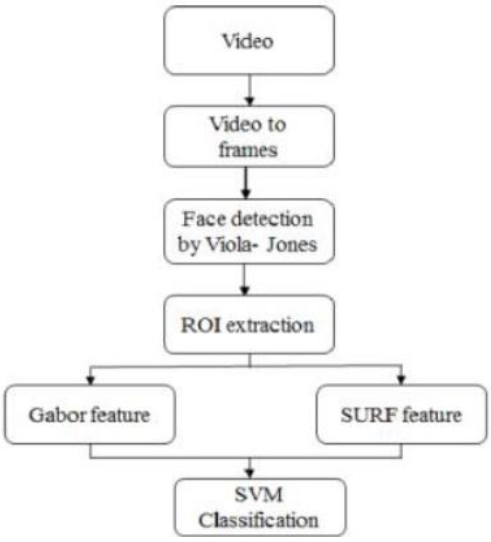

Fig .3. Multimodal ER for facial highlights

Here we find the Region of interest using the algorithm face detection by viola-jones. In the following stage, Algebraic highlights savage and standard deviation were determined for separate reactions of the Gabor channel. Next SURF highlights are removed from the whole face. To diminish the measure of height, width and length dimensions situation, the mean of the highlights was determined which was caused by the component with length of the vector 64. Grouping was done utilizing SVM with 4 unique feelings: disgust, anger, happiness, and fear.

Speech ER framework

Our work in this proposed system framework for feeling acknowledgment utilizing discourse is portrayed in Figure 4. Acknowledgment of emotion from spectral, speech, and prosodic highlights is utilized. Examining the spectral conduct, MFCC is investigated while pitch**,** energy, and formants are

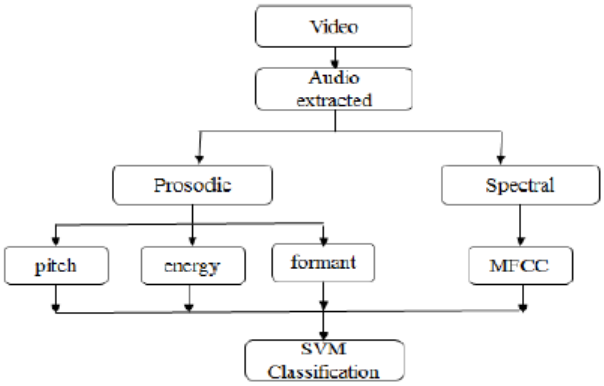

Fig. 4. Emotion recognition using speech framework

used to track down the prosodic data that is described using the figure 5.

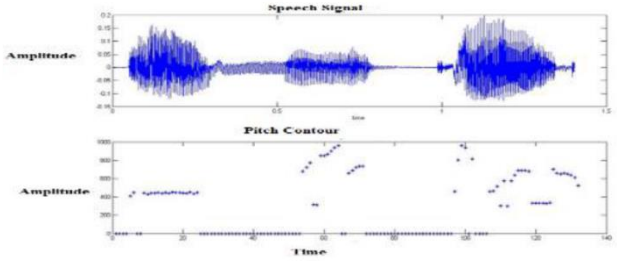

Fig.5. Pitch contour Extraction

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8837

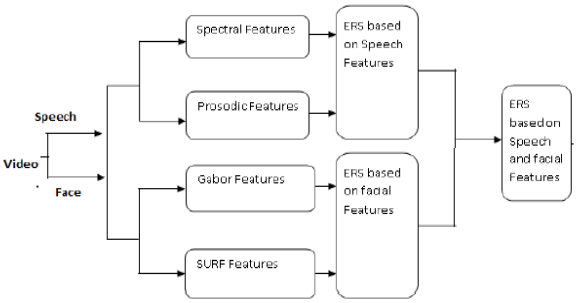

Fig.6. Proposed multimodal emotion recognition system

3. RESULTS AND DISCUSSIONS

At first face ER framework was created and some two highlights are separated in particular Gabor and SURF. The technique for extraction of these highlights is clarified in related works. As an underlying advance Gabor highlights are originated from left visual site, right site and mouth and took care of to the classification. The precision acquired for Gabor highlights by polynomial portion is 69.52% and RBF part is 48.37%. As a subsequent advance, the Face framework is constructed using SURF highlights. It gives precision as 64.16% in polynomial portion and 48.37% in the kernel of RBF bit.

Table.1. Multimodal comparison by two kernels

Features	Accuracy by RBF	Accuracy by Polynomial
Prosodic + Spectral	74.86%	85.21%
SURF + Gabor	65.65%	91.74%
Multimodal (face+speech)	89.74%	94.84%

Table. 2. Unimodal comparison by two kernels

Features	Accuracy by RBF	Accuracy by Polynomial
Prosodic	89.47%	78.94%
spectral	48.37%	64.16%
SURF features	48.37%	63.51%
Gabor features	48.37%	69.52%

In the step three, SURF and the Gabor highlights are joined which will give further developed exactness as 65.65% utilizing RBF portion and 91.74% utilizing polynomial bit. It was seen from the table 1 outcomes that the polynomial portion beats well for the characterization utilizing consolidated highlights. After that in the face ER framework, the speech ER framework was created. The MFCC highlights of the speech signal are extricated which yields a framework with a precision of 63.51 % for a polynomial bit. For speech framework, characterization utilizing RBF portion is giving better outcomes utilizing prosodic highlights though polynomial bit gives better outcomes for otherworldly highlights [42]. The combination of these features builds the general presentation of the framework as far as precision utilizing polynomial piece up to 85.21%. Table. 1 gives the precision of various frameworks and their correlation by utilizing various parts.

4. CONCLUSION

A multimodal structure on ER framework is the main research work of this paper. Then a joined framework was created, and the results uncover that there is an improvement in accuracy of 94.84%. It was seen that polynomial bit beats well in the grouping phase of the relative multitude of proposed frameworks. The proposed framework was tried on ongoing video and it is feasible to acquire the precision of 81% it can be upgraded the quality and number of preparing information. Feature step combination is utilized in this paper and it very well may be reached out to the signals of physiological things with the real situation rather than organized execution.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8837

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. P. Ekman, W. Friesen, and J. Hager (2002), The Facial Action Coding System: A Technique for the Measurement of Facial Movement. A Human Face.
4. 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.
5. Jibi Raj and Sujith Kumar (2015) 'Gender based Affection Recognition of Speech Signals using Spectral & Prosodic Feature Extraction' International Journal of Engineering Research Engineering Research and General Science, Vol. 3, Issue.2, pp.898-905.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Pantic and Patras (2006) 'Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences', IEEE Trans. Syst. Man, Cybernatics society, Vol. 36, Issue. 2, pp.433-449.
13. Siqing Wu, Tiago H. Falk and Wai-Yip Chan (2010) 'Automatic speech emotion recognition using modulation spectral features', Journal of Speech Communication', Vol.53, pp. 768-785.
14. 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.
15. Thushara.S and S.Veni (2016) 'A Multimodal Emotion Recognition from Video', International Conference on Circuit, Power and Computing Technologies (ICCPCT)
16. 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.
17. 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.
18. Mina Navran and Nasrollah Moghadam Charkari, (2014) Fusion of Feature Sets for Facial Expression Recognition', IEEE Transactions on Telecommunication, Vol. 62, No. 6.
19. 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.
20. Pujah Balasubramaniam; Gokilavani Sagadevan. "Converging Blockchain and AI technology-based Automated and Decentralized (A&D) Trust Management System using Face Detection". International Research Journal on Advanced Science Hub, 3, Special Issue ICITCA-2021 5S, 2021, 11-15.
21. Madhavi G; Jhansi Rani A.; Srinivasa Rao S.. "Pest Detection for Rice Using Artificial Intelligence". International Research Journal on Advanced Science Hub, 3, Special Issue ICITCA-2021 5S, 2021, 54-60

