Emotion Recognition System for Detecting Multimodal Feature using Intelligent Support Vector Machine

Sujni Paul[1], Rawia Elarabi [2], Najla Elhaj Babiker[3], P. Jayasuriya[4]

[1] *Faculty, Computer Information Science, Higher college of Technology, Dubai Mens Campus, Dubai.*
[2] *Department of Computer Science, College of computer Science & Information Technology, Jazan University, Jazan, Kingdom of SaudiArabia.*
[3] *Department of Computer Science, College of computer Science & Information Technology, Jazan University, Jazan, Kingdom of SaudiArabia.*
[4] *Lecturer, Department of Computer Science, College of Computer Science & Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia*
[1]spaul@hct.ac.ae,[2]relarabi@jazanu.edu.sa ,[3]nbabiker@jazanu.edu.sa,[4]jpanchalingam@jazanu.edu.sa

Abstract *- Multimodal emotion recognition victimization Support Vector Machine (SVM) introduces a sensible system that aims to acknowledge 3 completely different categories of human emotions appreciate happy, neutral and unhappy that relies on score-level fusion of speech and visual speech modalities from pre-processed static database. Speech refers to the sound signal created by human articulation system. The term speech options are usually wont to seek advice from the acoustic features extracted from speech. Mel-Frequency Cepstral Co-efficient (MFCC) is employed as acoustic features. Visual speech refers to the movement of lips, tongue and alternative facial muscles of the speaker. The term visual speech features is usually wont to seek advice from the features extracted from mouth or lip region. For this feature extraction method Viola Jones rule is used. every modality is sculpturesque by SVM classifier. sculpturesque result from acoustic and visual speech features got combined for recognizing the emotion.*

Keywords *- Emotion Intelligence, Emotion Recognition, Visual Speech features, SVM, Confusion Matrix*

DOI Number: 10.48047/nq.2022.20.19.NQ99124 **NeuroQuantology2022;20(19): 1361-1366**

1. Introduction

Emotion recognition relies on psychological analysis. principally emotions categorized by humans are used to replicate through voice, face expressions and body postures. For the aim of communication, citizenry ought to express their message through expressions that are mentioned earlier [1]. Here facial expression takes the most role in any interaction between a handful or cluster of peoples. In human-to-human communication deciding was a straightforward task due to the information concerning perception. Biological proof is that human emotional thoughts were transferred through multiple channels appreciate speech content, face and body postures with the assistance of brain management and neural system [2][3].

In human computer communication automatic emotion recognition is a crucial task; automatic emotion recognition is done by knowledge fusion technique. knowledge fusion is that the technique wants to mix data from multiple inputs. emotion recognition in human to computer interaction gets higher result with data fusion strategies [4]. Numerous knowledge fusion techniques are developed. In any emotion recognition process, fusion operation without doubt plays a crucial role. Operation of fusion is processed at 3 levels appreciate feature level, call level and model level for speech and visual speech primarily based emotion recognition [5].

Face and voice biometry are used as input to extract acoustic speech and visual speech data. Those extracted data are processed in 2 alternative ways at a same time, which ends two confidence values. Confidence values from speech and video from someone 'A' got calculated with fusion operator that is used to create the ultimate decision. In score level fusion, acoustic speech and visual speech options got concatenated and sculpturesque by a classifier for emotion recognition [6]. At feature-level fusion, data

spatiality can get increase and will get the information meagreness problem. At decision-level fusion, data features from multiple models are modelled by corresponding classifier for each model. A recognition result from every classifier gets into the method of fusion at the end. At model-level fusion multiple data streams got modelled in mutual correlation however contributions of multiple modalities were troublesome to explore [7] [8].

2. Problem Statement

Proposed methodology is employed to acknowledge human feelings similar to happy, unhappy and neutral from static video database. This paper introduces a sensible system that aims to discover a user's emotion to talk to a computer, by considering each audio and visual cues [9][10].

Here, mouth expressions (visual) and vocal expressions (speech) are wont to recognize somebody's emotion. Speech recognition includes speech processing, that converts audio wave into a sequence of feature vectors and people sequence of feature vectors got decoded into sequence of words for recognition purpose [11]. Speech got recorded by victimisation mike or telephone. Recognised words will be used as commands to regulate any robotic system, knowledge entry to manage a information and for document preparation. Visual speech considers the input supply as a video. This uses muscles activity signals [12][13].

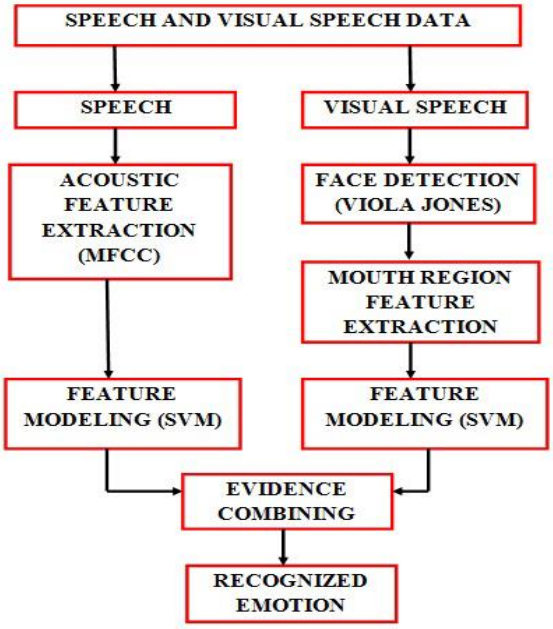


Fig. 1. Emotion Recognition System – Process

For speech recognition, acoustic feature extraction technique is employed to extract speech signal from a video database. Here, Mel frequency cepstral coefficients (MFCC) algorithmic rule is used for the method of speech feature extraction. when this, modelling of options are going to be in deep trouble speech part. For visual-speech, VIOLA JONES algorithm is used to discover face and mouth of a sample from video data. Then mouth region extraction half are going to be done.

This results a HAAR like feature of extracted mouth part. when this, modelling of extraction can be in deep trouble visual-speech part. to seek out the proper output sample, method of mixing will takes place. diagram for projected work shown in Figure 1. projected methodology is split into 5 modules • Speech feature extraction. • Face and mouth detection. • Visual speech feature extraction. • Feature modelling. • Combining evidences.

3. Feature Selection

Database contains forty-four subjects, expressed 3 completely different emotions in face and voice. This static database is employed to extract speech and visual speech features. For classification method (i.e., coaching and Testing) two different videos of all subjects from each category are taken. info is referred from customary "The enterface'05" project. Basic definition of feature extraction is that the process of spatial property reduction.

For any algorithmic rule great deal of input file could be a problematic issue to try and do the task. If data input is simply too huge for any calculation it'll be a redundant process. So, rather than massive amount of knowledge reduced set of data options are take into account as input, those reduced set of data features have the spare accuracy for modelling. mistreatment the MFCC algorithmic rule acoustic feature extraction method is going to be done. regulation of MFCC is shown in Figure 2. Extracted acoustic speech features for a sample shown in Figure three

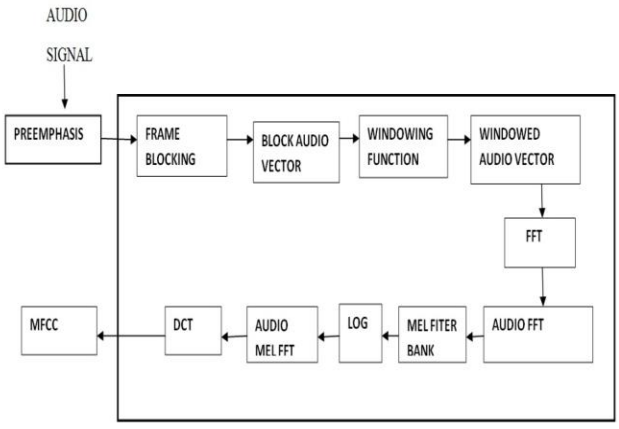


Fig. 2. Extraction of Multimodal Emotion Systems

MFCC computes the cepstral coefficients at the side of delta cepstral energy and power spectrum deviation. to induce the result as MFCC's, audio signal got processed underneath segmentation and windowing functions which ends into one hundred sixty samples of short frames. Magnitude spectrum was calculated to every frames mistreatment quick Fourier remodel (FFT) this results into a collection of Mel scale filter bank, resulted filter bank was processed underneath exponent and separate trigonometric function transformation for the output as MFCC's.

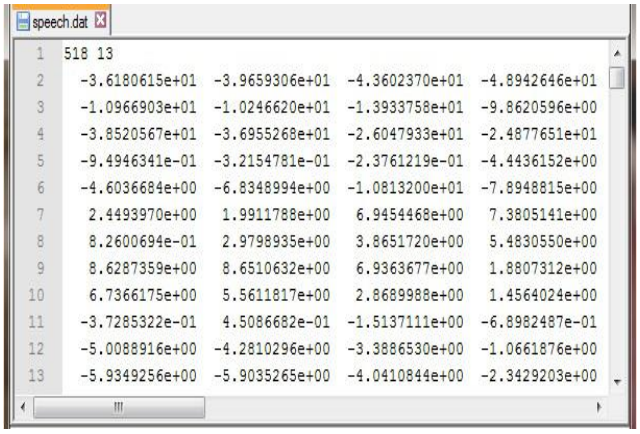


Fig. 3. Speech and Visual features Extraction using Matlab

4. Visual Speech Feature Extraction

Visual speech feature extraction defines the method of feature extraction from specific determination of an oblong mouth half from face. to seek out horizontal lines that belong to each components of a face, a footing filter is initial applied to the face region followed by a summation. mistreatment those horizontal lines illustration of mouth part distributed and correlation of localization of mouth part and mouth model achieved. to seek out the higher and lower lip part, color image is probe for the perimeters ranging from the very best correlation point. variety of pixels depends the resolution are accustomed mark the ROI parallelogram boundaries

Detection of face and mouth mistreatment Viola Jones algorithmic rule was shown in Figure 4. Mouth region extraction leads to kind of HAAR LIKE feature. mistreatment Viola-Jones object detection framework, Haar-like visual speech options got calculated. Haar-like features are used for beholding that is otherwise known as as digital image features. Calculation for detection part processes the input image with a window of result size image. for each segment of input image horizontal Haar-like options got calculated. Haar like visual speech features are shown in Figure 5.

Distinction between target objects associate degreed non-target objects got compared with the assistance of learned threshold, that separates non-target objects from the complete image. massive varietys of Haar-like features are required to state an object with spare accuracy as a result of tiny number of Haar-like feature could be a weak learner or classifier. Hence, Haar-like features are organized to create a robust learner or classifier which is termed a classifier cascade in Viola-Jones object detection framework.

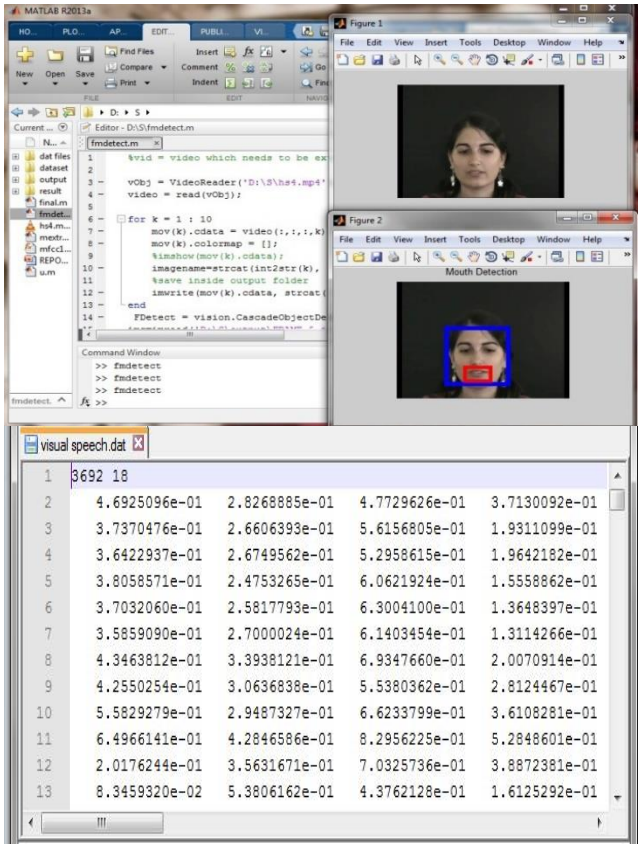


Fig. 4. Face and Mouth detection

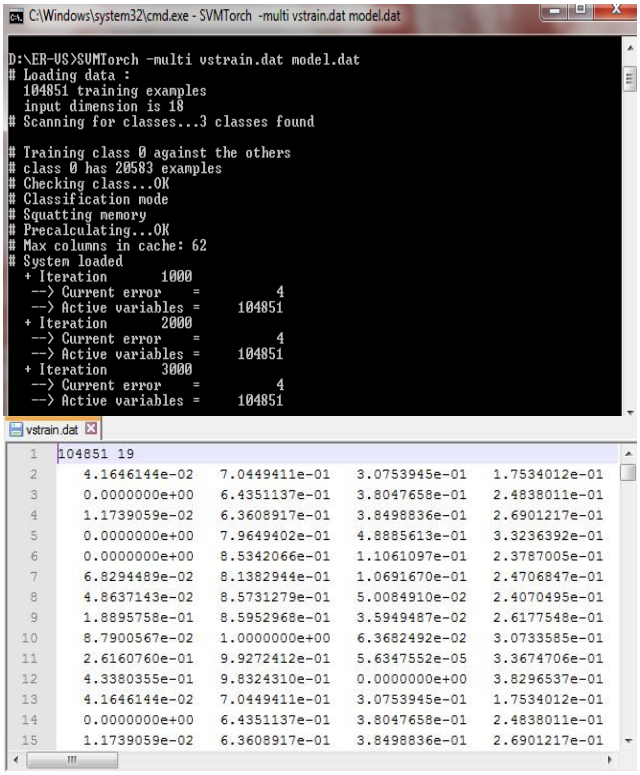
5. Support Vector Machine for Feature Modeling

A support vector machine (SVM's) may be a supervised learning rule that is employed to research info from input and acknowledge the input's pattern among over one category of patterns which is additionally known as support vector networks. Thus, SVM is principally used for classification purpose. AN SVM coaching algorithm builds a model that contains the sculptural knowledge belongs to more than one class.

AN SVM testing algorithm is used to check an example sample's modeled data with trained model, which provides the result as sample's class name in major among all classes. Figure half-dozen shows the training method and

Figure 5 shows the testing method of extracted feature victimisation SVM. Basically, an SVM model offers the illustration of samples as dots in area, this helps to examine the separation between any two samples belongs to 2 completely different categories. Prediction concerning new sample is a straightforward task once that new sample is mapped into an equivalent space to understand its category. AN SVM constructs a hyper plane which may be used for the aim of categorification.

In coaching part, numbers of sample data's feature values use to induce trained individually using SVM multi class classifier rule for speech and visual speech knowledge from database. Here, commonplace code for SVM technique is used. Once all data got trained victimisation standard procedure, all prepared for checking method. Figure seven shows the feature vectors for trained sample. In testing part, a replacement sample that one ne'er gets trained individual to trained samples was used for testing process. This process offers excellent result for objective of this project. Figure nine shows the score price from SVM for test sample, this score value shows the category name "2" in major numbers. as a result of the sample input used for this process belongs to class two among three classes.



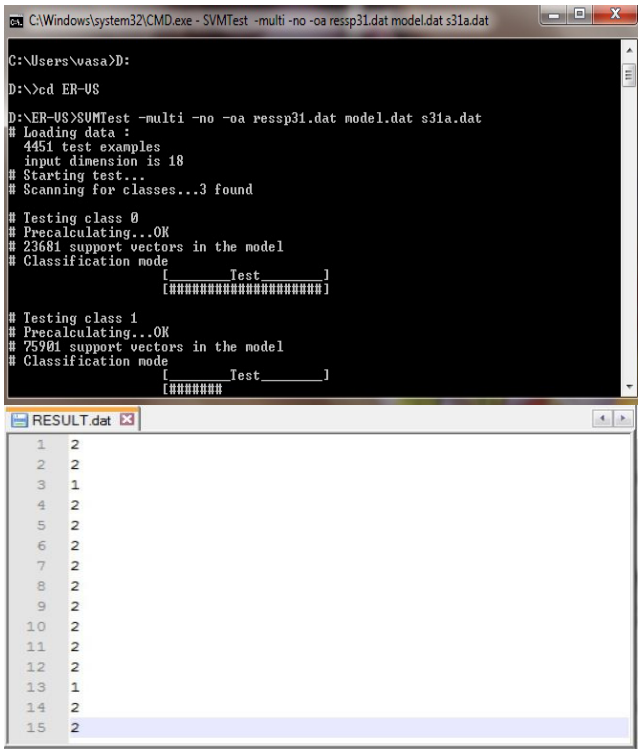


Fig. 5. Score from SVM for the test sample

Confusion Matrix

In this module, sculptural speech and visual speech results got calculated for score level fusion. Here weighted total model is employed for combining process. Weighted sum rule provides the result by summation speech and visual-speech output score. In call theory, the weighted sum model (WSM) is that the best proverbial and simplest multi-criteria decision analysis (MCDA) or otherwise called multi-criteria higher cognitive process methodology for evaluating variety of alternatives in terms of a number of decision criteria. it's important to state here that it is applicable only if all the info are expressed in mere the same unit. Given MCDA downside is outlined on m alternatives and n call criteria.

$$SV_{ij} = (W_x \times S_{ij}) + (W_y \times V_{ij})$$

wherever SVij represents the combined worth from sculptural speech and visual speech options (i represents the instances of foretold category and j represents the instances of the particular class), Sij denotes the modeled value of speech feature (i represents the instances of predicted class and j represents the instances of the actual class), Vij denotes the modeled value of visual speech feature (i represents the instances of predicted class and j represents the instances of the actual class), Wx and Wyoming denotes the constant worths used for weighted total model. Here constant values of Wx = 0.1 and Wyoming = 0.9 provides the weighted sum value as by

summing Wx and Wyoming that is perfect value (i.e., 1). Weighted summation for performance results from sculptural speech and visual speech values gives output in combined type which shows the recognized feeling in confusion matrix table. Overall performance for this categoryifier is measured by averaging the confusion matrix diagonal values, trained values and take a look at samples from each same class got intersected.

6. Experimental Results

In this work, static videos from information are accustomed do experiments whereas they're acting, to specific their emotions for act some message to the alternative party through face and voice. For experiment, forty four different samples spoke in 3 different expressions are trained victimization SVM Torch application. By using a similar samples check data's got recorded in several time for the mentioned expressions. SVM Test application is employed for testing process; it shows the result because the several sample's category name in major range among all class names. Similarly, experiments are conducted for visual speech to induce the results. Table 1 and Table two represents the performances of speech and visual speech victimisation SVM classifier. Table three shows the combined evidences from the higher than speech and visual speech performance tables, which ends in confusion matrix. Overall performance for this classifier is 91.12%.

Table 1. Emotion recognition performance using speech

TRAINED VALUES/TEST SAMPLES	HAPPY (%)	SAD (%)	NEUTRAL (%)
HAPPY	90.92	7.08	4.01
SAD	6.81	91.15	5.69
NEUTRAL	2.27	1.77	90.3

Table 2. Emotion recognition performance using visual speech

TRAINED VALUES/TEST SAMPLES	HAPPY (%)	SAD (%)	NEUTRAL (%)
HAPPY	95.45	6.81	2.36
SAD	3.54	91.90	10.61
NEUTRAL	1.01	1.29	87.03

Table 3. Emotion recognition performance using speech and visual speech

TRAINED VALUES/TEST SAMPLES	HAPPY (%)	SAD (%)	NEUTRAL (%)
HAPPY	94.99	6.83	2.52
SAD	3.86	91.82	10.11
NEUTRAL	1.13	1.33	87.35

7. Conclusion and Future work

In this paper, a multimodal feeling recognition system was implemented. Speech and Visual Speech options accustomed acknowledge several emotions of somebody's sample. MFCC feature got extracted for speech feature extraction technique. Viola Jones rule is employed for face and mouth detection. SVM was made using the speech features for every emotion. Similarly, visual speech options are accustomed construct associate SVM for every feeling. Overall performance was measured by combining the speech and visual speech primarily based SVM's for multi modal emotion recognition. Future work is to form comparison of this classifier's performance with another economical classifiers and to try to to this work with real time information set.

References

[1] Manikandan Sridharan, Delphin Carolina Rani Arulanandam, Rajeswari K Chinnasamy, Suma Thimmanna, Sivabalaselvamani Dhandapani, "Recognition of Font and Tamil Letter in Images using Deep Learning", Applied Computer Science, vol. 17, no. 2, pp. 90–99, 2021, doi: 10.23743/acs-2021-15

[2] Metallinou, S. Lee and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in Proc. of ICASSP, Dallas, Texas, 2019, pp.2462–2465.

[3] Metallinou, C.Busso, S. Lee and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in Proc. of ICASSP, Dallas, Texas, 2019, pp. 2474–2477.

[4] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, vol.1, pp.6-9, February 2019.

[5] An X, Zhang X, "Speech Emotion Recognition Based on

LPMCC", Sciencepaper Online.2019.

[6] PeipeiShen, Zhou Changjun, Xiong Chen. "Automatic Speech Emotion Recognition using Support Vector Machine," Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference, vol.2, no., pp.621-625, 12-14 Aug. 2021.

[7] M. E. Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2021.

[8] Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," IEEE Trans. Affective Comput., vol. 99, 2020.

[9] S. Manikandan, P. Dhanalakshmi, K. C. Rajeswari and A. Delphin Carolina Rani, "Deep sentiment learning for measuring similarity recommendations in twitter data," Intelligent Automation & Soft Computing, vol. 34, no.1, pp. 183–192, 2022. doi:10.32604/iasc.2022.02346

[10] Yixiong Pan; Peipei Shen and Liping Shen. "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home, 2012.

[11] Narayanan, Shrikanth, and Panayiotis G. Georgiou. "Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language", Proceedings of the IEEE, 2013.

[12] Zhi-Hang Tang, Bei-Ping Tang, Ying Han, Yi-Jie Lu, Xiang-Ling Luo, Wen-Bin Tian and Hai-Bin Wang. "A New Pattern Recognition Method Based on Nonlinear Support Vector Machine", International Review on Computers & Software, 2013.

[13] Marrero-Fernandez, Pedro, Arquimedes Montoya-Padron, Antoni Jaume-i-Capo, and Jose Maria Buades Rubio. "Evaluating the Research in Automatic Emotion Recognition", IETE Technical Review, 2014.

eISSN 1303-5150