INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 2)
Available online at: www.ijariit.com

Audio-Visual Emotion Recognition using 3DCNN and DBN Techniques

S. Kumari
kumari6996@gmail.com
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

U. Kowsalya
kowsalyaias97@gmail.com
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

R. Preethi
preethi19ravi@gmail.com
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

R. Theepa
theepaguru127@gmail.com
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

J. Edward Paulraj
paulgracehannah@gmail.com
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

S. JeyaAnusuya
hodece@tjsec.in
T. J. S. Engineering College, Puduvoyal, Tamil Nadu

ABSTRACT

Emotion recognition is difficult because of the emotional hole amongst emotions and varying audio-visual highlights. Propelled by the effective element learning capacity of profound neural networks, this system proposes to connect the emotional hole by utilizing a hybrid deep replication, which first creates varying audio-visual fragment highlights with Convolutional Neural Networks (CNN) and 3DCNN, at that point wires varying audio-visual section includes in a Deep Belief Networks (DBN). The point of this postulation work is to research the algorithm of discourse Emotion recognition utilizing MATLAB. Right off the bat, five most generally utilized highlights are chosen and separated from discourse flag. After this, measurable esteems, for example, mean, change will be gotten from the highlights. This information alongside their related Emotion target will be bolstered to MATLAB neural network apparatus to train and test to make up the classifier. The general framework gives a solid execution, arranging effectively over 82% discourse tests after appropriately preparing.

Keywords: *Speech Emotion Recognition, Classification, Convolutional Neural Networks, CNN, Deep Belief Networks, DBN.*

1. INTRODUCTION

Perceiving human emotions with PCs is typically performed with a multimodal approach because of the inborn multimodality normal for human emotions articulation. Discourse and outward appearance are two normal and compelling methods for communicating emotions when people speak with each other. Amid the most recent two decades, varying media emotions acknowledgment coordinating discourse and outward appearance has pulled in broad consideration inferable from its promising potential applications in human-PC collaboration. Be that as it may, perceiving human emotions with PCs is as yet a testing undertaking since it is hard to extricate the best sound and visual highlights portraying human emotions. In this proposition, the point is to characterize a cluster of recorded discourse motion in four classifications utilizing MATLAB, in particular: cheerful, dismal, irate, nature.

Before removing the highlights, these discourse signals should be pre-processed. In the first place take tests from the discourse to change over simple flag to computerized flag. At that point, the standardization ensures each is in a similar volume extend. Finally, the division is to isolate motion in outlines so discourse flag can keep up its trademark in brief term. Five regularly utilized highlights are examined and separate utilize MATLAB. Discourse rate and vitality are the most fundamental highlights of discourse flag yet despite everything they have critical diverse between emotions, for example, irate and miserable. Contribute is as often as possible utilized this subject and auto-relationship technique is utilized to identify the contribute each edge. After that measurable esteem, for example, mean, variance, max estimation of the pitch will be figured for discourse signals. The organization is another essential component of this framework. Linear Predictive Coding (LPC) strategy is utilized to extricate the main arrangement.
Like pitch, measurable esteems are figured for the first configuration. Initial three coefficients of MFCCs are taken to determine means and differences. Every one of the 15 highlights of 60 s is put into an info lattice alongside an objective framework, which shows the emotions state for each made the contribution out of neural network.

MATLAB neural network design acknowledgment apparatus is utilized to prepare and test the information and play out the classification, in the end, figures of mean square blunder and disarray will be given to demonstrate how great the execution is.

2. BLOCK DIAGRAM

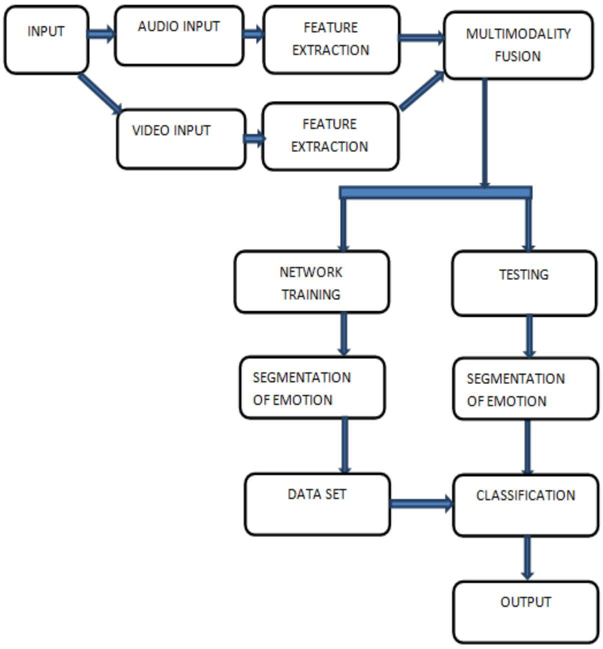


Fig.1 Proposed System Block Diagram

I. PRE-PROCESSING

Before removing the highlights, there are some essential strides to take to control discourse flag. Pre-process predominantly incorporates examining, standardization and division as appeared in Figure.2. The accompanying figure Pre-procedure of discourse flag. Discourse voice is a simple flag and it should be changed over into advanced flag to process in PC. Inspecting hypothesis gives away to change the simple flag x(t) into a discrete time flag x(n) and remains the normal for the unique flag.

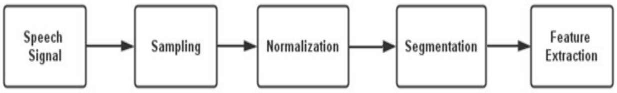


Fig.2 Pre-Processing

As indicated by inspecting theorem [5], when the examining recurrence is bigger or equivalent to 2 times of the most extreme of simple flag recurrence, the discrete time flag can reproduce the first simple flag. As showed in Figure 3, sampling is performed by gathering focuses from the simple flag in a specific rate Ts. be generally the testing frequencies for discourse signals are 8000Hz, 16000Hz and 44100Hz. In MATLAB examining is connected naturally after the chronicle work.

II. FEATURE EXTRACTION

An initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning. Convolutional Neural Network (CNN) is a powerful machine learning technique from the field of deep learning. CNNs are trained using large collections of diverse images. From these large collections, CNNs can learn rich feature representations for a wide range of images. These feature representations often outperform hand-crafted features such as HOG, LBP, or SURF. An easy way to leverage the power of CNNs, without investing time and effort into training, is to use a pre-trained CNN as a feature extractor.

III. FUSION

Two major groups of multimodality interfaces have merged, one concerned in alternate input methods and the other in combined input/output. The first group of interfaces combined various user input modes beyond input/output, such as speech. The second group of interfaces combines a visual modality with a voice modality (speech recognition for input, speech synthesis and recorded audio for output). Deep Belief Networks (DBN) consists of two different types of neural networks – Belief Networks and Restricted Boltzmann Machines. In contrast to perceptron and back propagation neural networks, DBN is unsupervised learning algorithm. Average-pooling is applied to all segment features from each video sample to form the fixed-length global video feature representation. We employ average pooling to process features extracted from segments. Based on this global video feature representation, the linear SVM classifier can be easily employed for emotion identification.

3. PAST SYSTEM-A SUMMARY

Recognizing human emotions with computers is usually performed with a multimodal approach due to the inherent multimodality characteristic of human emotion expression. Speech and facial expression are two natural and effective ways of expressing emotions when human beings communicate with each other. During the last two decades, audio-visual emotion recognition integrating speech and facial expression has attracted extensive attention owing to its promising potential applications in human-computer interaction. However, recognizing human emotions with computers is still a challenging task because it is difficult to extract the best audio and visual features characterizing human emotions. The past system contains several disadvantages, in that some of them are summarized as follows: (a) Time-consuming Process and (b) Low Accuracy in recognition.

4. PROPOSED SYSTEM-A SUMMARY

In the proposed system, we introduce a hybrid deep learning framework composed by CNN, 3D-CNN, and DBN to learn a joint audiovisual feature representation for emotion classification. It is comprised of three steps: (i) we convert the raw audio signals into representation similar to the RGB image as the CNN input. Consequently, a deep CNN model pre-trained on large-scale ImageNet dataset can be fine-tuned on audio emotion recognition tasks to learn high-level audio segment features.(ii) For multiple contiguous frames in a video segment, a deep3D-CNN model pre-trained on the large-scale video dataset is fine-tuned to learn visual segment features for facial expression recognition. (iii) The audio and visual segment features learned by CNN and 3D-CNN are integrated into a fusion network built with a deep DBN, which is trained to predict correct emotion labels of video segments. Finally, we adopt the outputs of the last hidden layer of DBN as the audio-visual segment feature. Average-pooling is employed to aggregate all segment features to form a fixed-length global video feature. Then, a linear SVM is used for video emotion classification. The proposed system contains several advantages, in that some of them are summarized as follows: (a) Improved Emotion Recognition in both Audio-and-Video and (b) Increased process time.

5. LITERATURE SURVEY

In the year of 2011, the authors "M. El Ayadi, M. S. Kamel, and F. Karray" proposed a paper titled "Survey on speech emotion recognition: Features, classification schemes, and databases", in that they described such as: as of late, expanding consideration has been coordinated to the investigation of the passionate substance of discourse signals, and henceforth, numerous frameworks have been proposed to recognize the enthusiastic substance of a talked articulation. This paper is a study of discourse feeling arrangement tending to three essential parts of the plan of a discourse feeling acknowledgment framework. The first is the decision of reasonable highlights for discourse portrayal. The second issue is the plan of a fitting order conspires and the third issue is the correct readiness of a passionate discourse database for assessing framework execution.

Decisions about the execution and constraints of current discourse feeling acknowledgment frameworks are talked about in the last segment of this study. This area additionally proposes conceivable methods for enhancing discourse feeling acknowledgment frameworks.

In the year of 2016, the authors "L. Gao, L. Qi, and L. Guan" proposed a paper titled "Information fusion based on kernel entropy component analysis in discriminative canonical correlation space with application to audio emotion recognition", in that they described such as: as a data combination apparatus, Kernel Entropy Component Analysis (KECA) is acknowledged by utilizing descriptor of data entropy and enhanced by entropy estimation. Be that as it may, as an unsupervised technique, it just puts the data or highlights from various channels together without thinking about their inborn structures and relations.

In this paper, we present an improved variant of KECA for data combination, KECA in Discriminative Canonical Correlation Space (DCCS). Not just the inborn structures and discriminative portrayals are considered, yet in addition, the characteristic portrayals of information are uncovered by entropy estimation, prompting enhanced acknowledgment exactness. The viability of the proposed arrangement is assessed through examinations on two sound feeling databases. The test comes about demonstrate that the proposed arrangement beats the current techniques in view of comparable standards.

In the year of 2016, the authors "N. E. D. Elmadany, Y. He, and L. Guan," proposed a paper titled "Multiview emotion recognition via multi-set locality preserving canonical correlation analysis", in that they described such as: in this framework, we propose a novel Multi-set Locality-Preserving Canonical Correlation Analysis (MLPCCA) for multi-see learning and combination. The proposed MLPCCA catches the inborn structure of information while it takes in the ideal reason for augmenting the connection among various arrangements of information.

To confirm the adequacy of the proposed strategy, the proposed MLPC A has been connected in sound based feeling acknowledgment and visual-based feeling acknowledgment, individually. The exploratory outcomes exhibited that the proposed MLPCCA can accomplish a higher acknowledgment exactness contrasted with the current techniques including CCA, LPCCA, and MCCA.

6. EXPERIMENTAL RESULT

The following figure illustrates the Audio Emotion recognition.

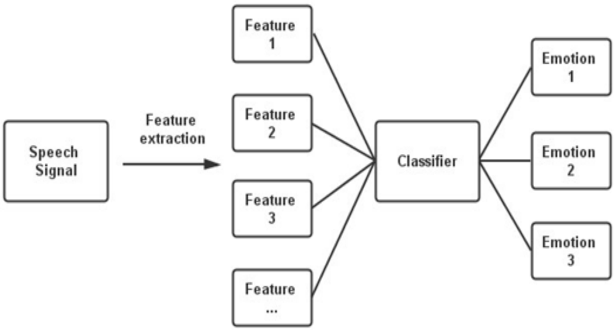


Fig.3 Audio Emotion Recognition

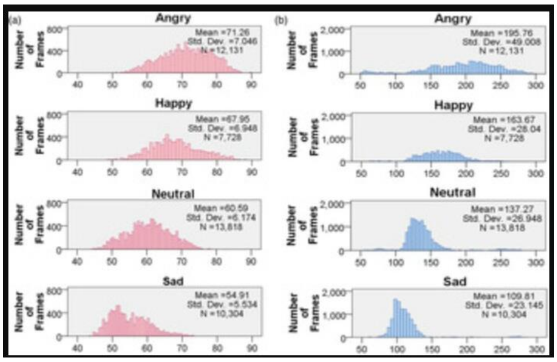
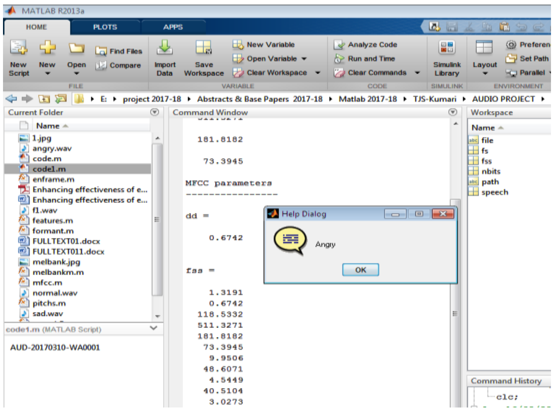


Fig.4 feature extraction of Audio Emotion recognition



```
   181.8182

    73.3945

MFCC parameters
----------------

dd =

    0.6742


fss =

    1.3191
    0.6742
  118.5332
  511.3271
  181.8182
   73.3945
    9.9506
   48.6071
    4.5449
   40.5104
    3.0273
```

Angry

Fig.5 Result of Audio Emotion recognition

The following figures illustrate the Video emotion recognition

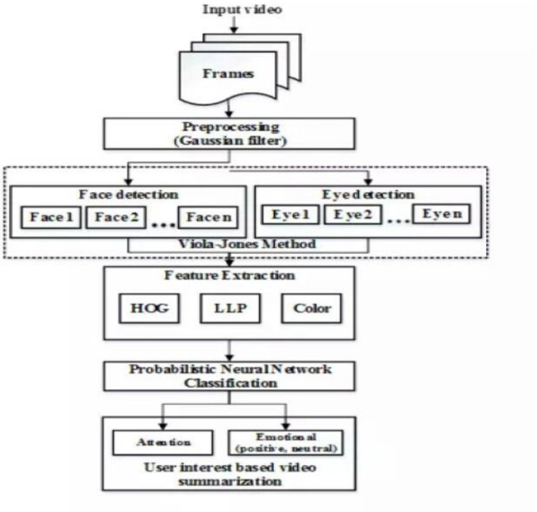

Input video
Frames
Preprocessing (Gaussian filter)
Face detection
Face1
Face2
...
Facen
Eye detection
Eye1
Eye2
...
Eyen
Viola-Jones Method
Feature Extraction
HOG
LLP
Color
Probabilistic Neural Network Classification
Attention
Emotional (positive, neutral)
User interest based video summarization

Fig. 6 Video Emotion Recognition

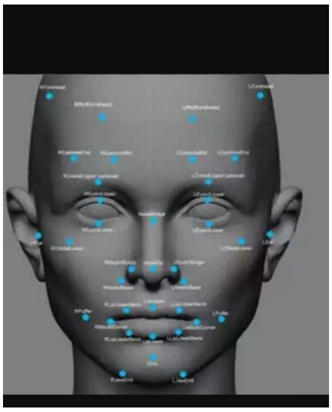


Fig. 7 feature extraction of Video Emotion Recognition

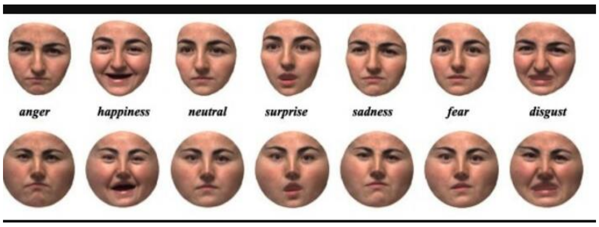


Fig.8 Result of Video Emotion Recognition

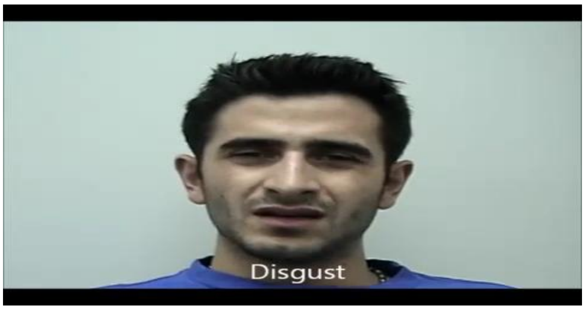


Fig.9 Video Emotion Recognition – Disgust

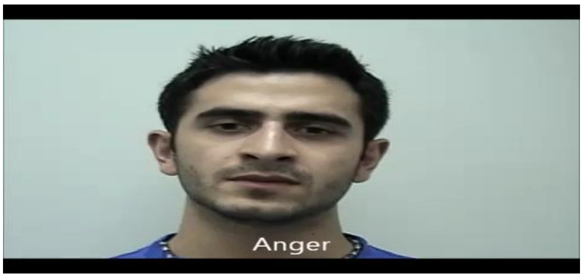


Fig.10 Video Emotion Recognition – Anger

7. CONCLUSION

MATLAB neural network is a powerful tool for pattern recognition and classification. With its simple user interface, the chosen features of speech signals could be easily loaded into the system and training for the target emotions. After suitable times of training process, extra test signals are load into the system for emotion recognition. With the desired result of 85% classification rate those selected features (speech rate, energy, pitch, format, and MFCC) are proven to be good representations of emotion for the speech signal. For the further work, the system could be improved by increase the accuracy of extracted features to classify more complicate speech samples, that is, for multiple speakers and more emotions.

8. ACKNOWLEDGEMENT

We would like to thank all those who provide us the possibility to propose this project. A special gratitude to our Project Guide Mr. J. Edward paulraj, Associate professor and Project coordinator Mrs. S. Jeya Anusuya, Associate professor, whose contribution in giving suggestions and encouragement helped us for this project to complete.

9. REFERENCE

[1] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, “Audio–visual emotion-aware cloud gaming framework,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 12, pp. 2105–2118, 2015. 1

[2] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, “Multimodal prediction of affective dimensions and depression in human-computer interactions,” in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC), Orlando, FL, USA, 2014, pp. 33–40. 1

[3] Y. Wang and L. Guan, “Recognizing human emotional state from audiovisual signals*,” IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936–946, 2008. 1, 3, 6, 7

[4] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, “Audio-visual affective expression recognition through multistream fused hmm,” IEEE Trans. Multimedia, vol. 10, no. 4, pp. 570–577, 2008. 1, 3

[5] M. Mansoorizadeh and N. M. Charkari, “Multimodal information fusion application to human emotion recognition from face and speech,” Multimedia. Tool. Appl., vol. 49, no. 2, pp. 277–297, 2010. 1, 3, 8, 11

[6] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. K¨achele, M. Schmidt, H. Neumann, G. Palm et al., “Multiple classifier systems for the classification of audio-visual emotional states,” in Affective Computing and Intelligent Interaction (ACII). Springer, 2011, pp. 359–368. 1

[7] M. Soleymani, M. Pantic, and T. Pun, “Multimodal emotion recognition in response to videos,” IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 211–223, 2012. 1

[8] J.-C. Lin, C.-H. Wu, and W.-L. Wei, “Error weighted semi-coupled hidden markov model for audio-visual emotion recognition,” IEEE Trans. Multimedia, vol. 14, no. 1, pp. 142–156, 2012. 1, 3

[9] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, “Exploring fusion methods for multimodal emotion recognition with missing data,” IEEE Trans. Affect. Comput., vol. 2, no. 4, pp. 206–218, 2011. 1

[10] A. Metallinou, M. W¨ollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification,” IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 184–198, 2012. 1

[11] D. Gharavian, M. Bejani, and M. Sheikhan, “Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks,” Multimed. Tool. Appl., pp. 1–22, 2016. 1

[12] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “Baum-1: A spontaneous audio-visual face database of affective and mental states,” IEEE Trans. Affect. Comput., 2016. 1, 3, 6, 7, 8, 11