IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Multimodal Emotion Recognition using Deep Neural Network- A Survey

Haritha C. V[1]*, Pillai Praveen Thulasidharan[2]

[1]Department of Computer Science and Engineering, N.S.S College of Engineering, Palakkad
[2]Department of Computer Science and Engineering, N.S.S College of Engineering, Palakkad

*Corresponding Author: hbharithababu@gmail.com

Available online at: www.ijcseonline.org

Abstract— Emotion recognition is a process by which human emotional states can be identified. Most of the present methods make use of visual and audio information's together. With recent advancements in deep neural networking, there are several methodologies to identify human emotional states. One of the methods that detect the emotional states is based on a multimodal Deep Convolution Neural Network (DCNN), that use both the audio and visual cues in a deep model. BLSTM-RNN is another method which makes use of multimodal features to capture emotions. A much more efficient approach is using a convolutional neural network (CNN) to extract features from the speech, and for the visual modality, the features can be extracted using a deep residual network of 50 layers. To capture contextual information's a long short-term memory network can be utilized above these two models. Deep belief networks are another method which takes multimodal emotion recognition into account by first learning the features of the audio and video separately; after which it concatenates these two features. Visual features hold more importance in emotion recognition, so ResNet along with SVR for training can be used to predict emotion states effectively.

***Keywords*—** DCNN, DBN, Residual Network, LSTM, SVR

I. INTRODUCTION

Emotion recognition can be defined as a process of predicting the affective state of an individual from low-level signal cues produced by him. It is an essential component in human-computer interaction. So, it plays a major role in human intelligence as well as social interactions. Speech and facial expressions are the most natural ways of expressing one's emotion. So, integrating them for predicting emotional states has attracted extensive attention.

Emotion recognition is not an easy task. It is mainly because the accuracy of an emotion recognition system relies mainly on the ability to generate some representative features. But, this is a very tedious task. Emotion states do not have explicit temporal boundaries and different individual express their emotions in different ways.

Feature extraction can be viewed as the first step of distinguishing emotion. So far there has been a substantial number of work [1, 2] indicating feature extraction for audio-visual emotion recognition, such as Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients. Recently machine learning techniques have been gathering momentum in predicting the emotional states of a user. There are several methodologies indicating the use of the deep neural network in the field of emotion recognition. They are

- Multimodal Deep Convolution Neural Network (DCNN).
- Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN).
- Combined CNN and ResNet-50 architecture with LSTM.
- Deep Belief Network (DBN) models.
- Residual Network along with Support Vector Regression.

II. RELATED WORK

A. Deep Convolutional Neural Network for Emotion Prediction

Deep Convolutional Neural Network [3, 4] is composed of convolutional layers followed by fully connected layers, where convolutional layers learn a discriminative multi-level feature representation from raw input and the fully connected layers can be regarded as a non-linear classifier. DCNN is trained with two stages. In the first stage, two DCNN models are pre-trained over image data. In the next stage, the outputs of these two DCNNs are combined to form another network.

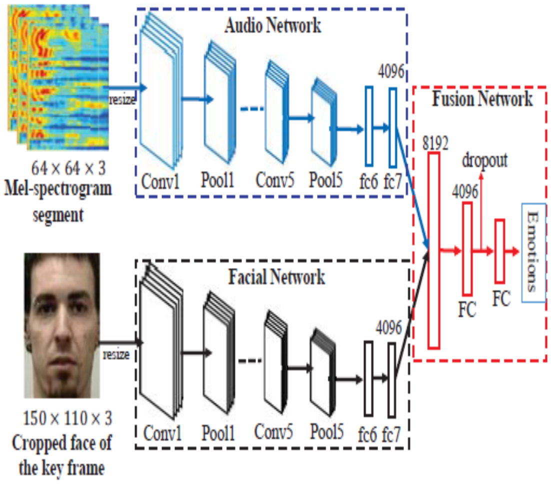

Figure 1: Multimodal Deep Convolutional Neural Network[4]

This network is an example of AlexNet model [6] which contains five convolutional layers, among which three are followed by max-pooling layers, and three fully connected layers.

As shown in figure 1 the process[5] start by first extracting the log Mel-spectrogram from audio signals, then divide the spectrogram into overlapping segments using a context widow with a fixed length. The central video frame in this context window is chosen as the key frame of visual data. Initially, the audio and facial networks are trained, and then jointly train the fusion network in the next stage. After training the fusion network, the resultant is a set of joint feature representation for each video segment. On applying average pooling on its segment features one can have its final global feature representation. Based on these representations, emotion prediction can be done with classifiers such as the linear SVM

B. *Predicting Asynchronous Dimensional Emotion Ratings*

Machine learning algorithms are an effective way to find the emotional states of a person. But they are sensitive to outliers. So, there must be a new technique which is insensitive to noises and at the same time, they must be able to identify the context. Such a solution is LSTM [7] network. LSTM is the basic component in Recurrent Neural Network.

Acoustic features can be extracted using openSMILE. A facetracker based on supervised descent method (SDM) [8] can be used for detecting the face region, which is the initial step of visual features extraction. Landmarks for the face are returned by the facetracker. The optical flow around the head region is computed using Farneback's [9] algorithm. Figure 2 shows facetracking and optical flow [15].

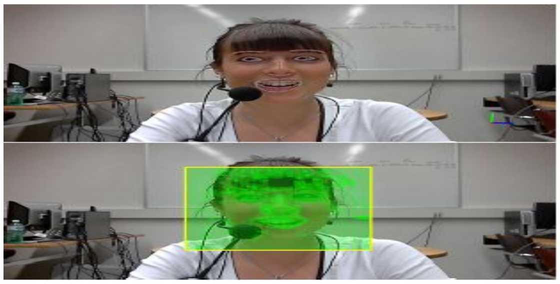

Figure 2: Top. Face tracking; Bottom. Optical flow calculated around the head region[15].

LSTM has a memory cell and three multiplicative gates: the input gate, the output gate and the forget gate. Input gate biases the value of memory cell and forget gate controls the decay of stored input. In this case, the memory blocks are bidirectional LSTM-RNN (BLSTM-RNN)[10]. Bidirectional RNN has access to all past and all future inputs. This property is made possible by processing the data in both directions within two hidden layers: one processes the data sequence forwards, the other one backward. Finally, the outputs are connected to the same output layer.

C. *Multimodal Emotion Recognition*

Multi-modal emotion recognition with the help of deep neural network is an automatic affect sensing system that uses both speeches as well as visual information in an end-to-end manner. Features for speech network are extracted using CNN and for visual network, ResNet [11] is used. ResNet's are efficient than CNN for image classifications. The output of these networks is combined and fed as input to an LSTM network which can determine the emotional states by specifying two numeric values namely arousal and valence. Arousal indicates the alertness or awakeness while valence indicates the goodness or badness of an event.

Pixel intensities of the video frame are fed as input to the visual network. ResNet consists of a series of layers. The first layer is a 7 × 7 convolutional layer. Then there are 3 convolutional layers with 1 × 1, 3× 3, and 1 × 1 sizes for each residual function. The input to the audio network is a segmented raw waveform on which a temporal convolution is performed to capture finer-level spectral information's followed by a max-pooling operation. Again a convolution operation is performed to extract long-term characteristics and roughness of the speech signal.

Features extracted from these two networks are fed to a 2 layer LSTM which is trained using backpropagation algorithm. The final result will be two numerical values, the arousal, and valence. Figure 3 shows the overall architecture of the system [16].

This network is evaluated over multimodal dataset RECOLA (REmote COLlaborative and Affective) [12] which contain multimodal data's such as audio, video, electrocardiogram (ECG) and electro-dermal activity (EDA).

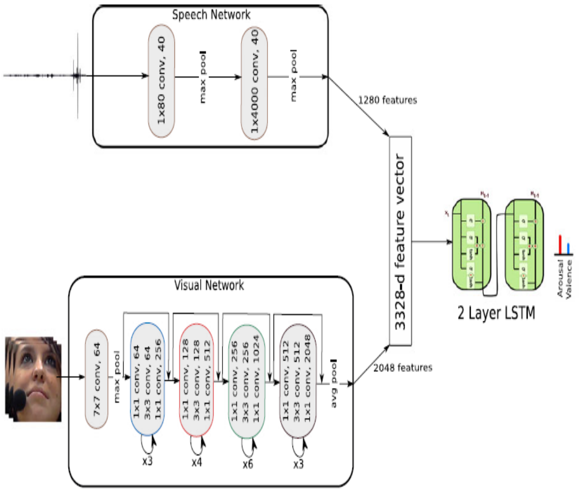

Figure 3: End to End multimodal emotion recognition network[16]

D. *Deep Belief Networks on Emotion Prediction*

A Deep Belief Network (DBN) [13] can be a graphical model, which is build-up with a number of layers of hidden units, having connections between the layers but not between units within each layer. This paper illustrates the use of four different DBN models to explore non-linear dependencies between audio and video features. DBN [14] model works by using the audio and video features. At first, the learning is performed separately and then concatenates them together. This is given as input to the second hidden layer, DBN2 model.

The variants of DBN2 model are [17].

- FS-DBN2 - Two-layer DBN with feature selection before training.
- DBN2-FS - Two-layer DBN with feature selection on the final RBM nodes.
- DBN3 - Three-layer DBN that uses an additional RBM.

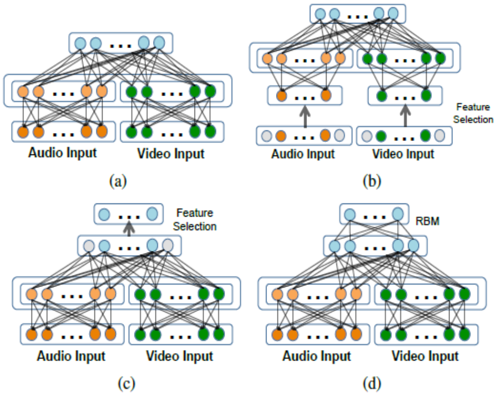

Figure 4: (a) DBN2, (b) FS-DBN2, (c) DBN2-FS, and (d) DBN3[17].

E. *Continuous Affect Recognition using Visual Features*

According to the studies in psychology, there are two major emotion computing models: discrete theory and dimensional theory. In discrete theory, the emotional states are described as discrete labels such as "surprise", "sad", "happy" etc. while in dimensional theory emotion state are considered as a point in a continuous space. Here the focus is on dimensional emotion recognition. Visual features have more influence in detecting emotional states of an individual. So besides the basic features, multi-scale Dense SIFT features (MSDF), and some of the Convolutional neural networks (CNNs) features are also used. Training is done by linear Support Vector Regression (SVR). ResNet is used to capture visual features [18].

III. COMPARISON

Table 1: Comparison of different methods

No	Method	Advantages	Disadvantages
1	Deep Convolutional Neural Network	First method that use deep learning to bridge the emotion gap	Needs further improvement in performance
2	BLSTM-RNN	Takes contextual information into consideration .	Complex network may be needed to cope up with the increase in complexity of task.
3	CNN+ ResNet	Not effected by outliers and able to consider the context	Works on French database, RECOLA
4	DBN	Good for capturing non-linear feature interactions	Does not consider dynamic features
5	ResNet + SVR	Good for complex affective states	Needs to improve recognition performance

IV. CONCLUSION

Emotion recognition can be defined as a process of identifying a person's emotional state. But this is a difficult task because usually emotions lack temporal boundaries and every individual expresses their emotions in their own ways. Recently, deep neural networks are used extensively in this field with great success. Emotions can be determined with much accuracy based on combined audio and visual effects. This kind of audio-visual emotion recognition can be achieved using multimodal DCNN, BLSTM-RNN, Combination of CNN and ResNet DBN and also ResNet with SVR.

ACKNOWLEDGEMENT

Thanking the Department of Computer Science and Engineering, N. S. S. College of Engineering, Palakkad for providing required facilities.

REFERENCES

[1].Y. Wang and L. Guan, " Recognizing human emotional state from audio-visual signals", *IEEE Trans. Multimedia.*, pp:936–946, 2008.

[2].A. Hanjalic and L. Xu, "Affective video content representation and modelling", *IEEE Trans. Multimedia.*, pp: 143–154, 2005.

[3].Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition", *Int. J. Comput. Vis.*, pp:54–66, 2015

[4].W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, et al., "Deepid-net: Deformable deep convolutional neural networks for object Detection", In *CVPR*, 2015.

[5].S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proc. Int. Conf. Multimedia Retrieval*, pp. 281–284, 2016.

[6].A. Krrizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", In *NIPS*, 2012.

[7].S. Hochreiter and J. Schmidhuber, "Long short-term memory," *NeuralComput*., pp. 1735-1780, 1997.

[8].Xiong, X., De la Torre, F., "Supervised descent method and its applications to face alignment", *in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* pp. 532–539, 2013

[9].Farneb¨ack, G, " Two-frame motion estimation based on polynomial expansion, in: Image Analysis*", in Springer*, pp. 363–370, 2003.

[10].Schuster, M., Paliwal, K.K., "Bidirectional recurrent neural networks" *IEEE Trans. on Signal Processing* 45, pp:2673–2681, 1997.

[11].K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. Conf. Comput. Vis. Pattern Recognit,* pp. 770–778, 2016.

[12].F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions", *IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, pp. 1–8, 2013.

[13].Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, pp. 1–127, 2009.

[14].J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng, "Multimodal deep learning," *in Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011

[15]. F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audio visual and physiological data," *Pattern Recognit. Lett.*, pp. 22–30, 2015.

[16].Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Bjorn W.Schuller, and Stefanos Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural networks", in IEEE Journal of Selected Topics in Signal Processing, pp: 1301 - 1309, 2017.

[17].Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 3687–369, 2013

[18].B. Sun, S. Cao, L. Li, J. He, and L. Yu, "Exploring multimodal visual features for continuous affect recognition," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Amsterdam, pp. 83–88, 2016.