Unimodal to Multimodal Emotion Recognition: A Systematic Review

Prof. Komal Anadkat *1, Prof. Hiteishi M. Diwanji2
*[1]IT department, Government Engineering college, Gandhinagar, Gujarat, India
komal.anadkat282@gmail.com
2IT department, L.D. College of Engineering, Ahmedabad, Gujarat, India
hiteishi.diwanji@gmail.com

Article Info
Volume 83
Page Number: 6423 - 6427
Publication Issue:
July - August 2020

Article History
***Article Received**: 25 April 2020*
***Revised**: 29 May 2020*
***Accepted**: 20 June 2020*
***Publication**: 10 August 2020*

Abstract

As Artificial intelligence is the fastest growing field for interdisciplinary research field in which computer science techniques can be implemented to solve the problems of social science. Emotion recognition is very challenging work as every person will express his/her emotion in different ways. The primary goal of this paper is to present the detailed study and critical analysis of existing unimodal recognition techniques like expression recognition from still images or video, audio expression recognition from available speech data , text expression from available social media post and psychological signals expression from EEG,EDA and ECG data. As unimodal emotion analysis is suffering from lower accuracy and lack of reliability, the multimodal emotion fusion is essential. This paper summarized the background of the multi modal fusion with existing data sets, feature extraction methods, different deep learning models they used and various fusion techniques like feature level or decision level.

Keywords: Emotion recognition, multimodal, fusion, Survey, Unimodal.

I. INTRODUCTION

The demand for human-computer interaction are increasing now a days as people are facing more problems related to mental health than physical problems. Emotion recognition and depression detection play a key role to decrease suicide rates .If machine can observe, analyze, understand and can give feedback about the emotions of human accurately, it will be a great achievement in the field of computer vision and mental health care monitoring. Facial expression can easily observe and people mostly communicate their emotions via face but at a same time people can faked their facial expression.so it is better to combine multiple models like speech recognition, EEG signal, Social media text along with facial expression to accurately measure the emotion.

II. THE STATE OF THE ART

In this section, we describe various Unimodal system’s existing work which mainly cover expression modal, audio expression modal, and psychological modal and social media modal. After reviewing and comparing existing literature, different fusion techniques will be discussed. The next section covers multimodal emotion data fusion and recognition using different machine learning and deep learning techniques.

A. Expression Recognition Model

Facial expression are playing key role for identifying intentions or emotions of other human as humans are mostly communicate via facial landmarks like nose, eyebrows ,eyes and mouth. Ekman et al. [11], considered pioneers in this research, argued that it is possible to detect six basic emotions, e.g., Anger, Joy, Sadness, Disgust and Surprise from cues of facial expressions. Before the invention of deep neural networks, different machine learning techniques were used to extract the features from still images or sequential frames like video. But it’s very difficult to extract the features manually as it require lots of domain knowledge and effort. Deep learning based facial emotion recognition enabled to extract the features automatically. Using deep learning is like “Black box”, as neural network dose not disclose how the features are extracted.

In recent time, a deep-learning algorithms are widely used for the purpose of feature extraction, classification and many computer vision task. Many researcher have done lots of work on facial emotion classification using neural network. Kaihao Zhang, Yongzhen Huang[1] have proposed Spatial-Temporal Networks which include two kinds of networks. First network called PHRNN learned dynamic features from consecutive frames or video and second network called MSCNN extract static features from still frames. Later on, the two kinds of networks are combined to improve the performance of facial expression recognition. Recently Bilal Taha, Dimitrios Hatzinakos [2] have proposed the network which consists of a CNN model that is deep enough to automatically learn features from the facial image. Many preprocessing has been done by the author like dropout, batch normalization, and data augmentation to avoid over fitting. Then, handcrafted features like texture, shape and fusion of both have compared with CNN model. The Comparison shows that CNN outperform than handcrafted features. A. Talipu , A. Generosi, M. Mengoni [3] have merged three different datasets like CK+,FER and affectnet. The reason behind merging is that it can give higher accuracy as CK+ and FER are wild dataset. Deepak Kumar Jain [4] have used Residuals network after alternate convolution layer to increase the prediction accuracy. Table 1. Shows the comparative analysis of different literatures of facial expression recognition model.

TABLE I EXPRESSION RECOGNITION LITERATURE SUMMARY

Dataset	**Classifier used**	**Year**	**Preprocessing**	**Advantages**	**Challenge**
CK+, Oulu-CASIA, and MMI	RNN	2017 [1]	Normalization, Dropout	Highest accuracy achieved is 98.50% on CK+ dataset and 81.18% on MMI.	Accuracy for fear and sadness is low as they give slight variation in dynamic frames.
Bosporus	CNN +SVM	2019 [2]	Data augmentation, dropout done	To avoid over fitting the developed CNN model, several techniques are included. Highest accuracy achieved is 92.14% for happy category when CNN used with SVM.	Dataset contains still frames which can’t generate temporal features.
CK+FER +Affectnet	CNN on Wild dataset	2019 [3]	Facial alignment on merged dataset	The best accuracy is achieved by the VGG13 architecture which is 75.48%.	Facial expression accuracy for categories such as fear and disgust are low mainly .Still images
CK+,JAFFE	DNN with residual learning	2019 [4]	Gaussian normalization and standard deviation	Batch normalization is used to improve generalization and optimization. Highest accuracy achieved is 95.23% on JAFFE dataset.	Dataset contains still frames which can’t generate temporal features

B. EEG Based Recognition Model

Electroencephalography Signal analysis play vital role in brain-computer interfacing .Many bio medical application like Emotions of patient suffering from epilepsy or paralysis can be well understood using EEG signals. Emotions can be accurately recognize using EEG signals because they cannot be manipulate. But the major challenge here is this recognition can only be done in controlled environment and lots of set up is require .EEG signal contains various frequency bands named alpha, beta, gamma, delta and theta. Large activity of beta waves & low activity of alpha waves indicate high arousal. Beta waves are active when brain is busy doing higher activity. Alpha waves are active when person is relaxed. Arousal is the ration of beta waves and alpha waves. EEG signal needs pre-processing to remove artifacts which was added at the time of collection. In paper [5] author have used differential entropy for feature extraction and then five classifiers like KNN, SVM, GELM, DBN, DBN-HMM used for the analysis task. Discrete Wavelet transform method is used for feature extraction in the combination with SVM and hidden markov model classifier by author in paper [6]. To resolve the problem of long term dependency LSTM model was used in paper [7].Wavelet transform in the combination of CNN is used for feature extraction and classification respectively in paper [8].The major challenge of this model is to classify the real emotions like happy, sad, fear, anger, disgust as the prediction labels are arousal and valence most of the time.

TABLE II EEG BASED LITERATURE SUMMARY

Dataset	**Classifier used**	**Year**	**Preprocessing**	**Advantages**	**Challenge**
12 movie clips were used.	DBN+HMM	2014 [5]	Filter the noise and artifacts, band pass filter used.	Accuracy achieved is 87.62%	Only the positive and negative emotional state allocated.
DEAP dataset	SVM + HMM	2017 [6]	Filter used to remove	Good accuracy compare to single classifier	Maximum 60% valence and 62% arousal accuracy.

			noise, sampling		
Psychiatry dept of Medical college,kerala	LSTM	2019 [7]	A notch filter was used to eliminate interference	RMSE is lower 0.005 compare to CNN+LSTM which has 0.007	Prediction is done in mean data points which does not show real emotions.
DEAP dataset	CNN	2019 [8]	Filter to remove noise.	highest accuracy obtained for valence classification Which is 62.50%	Not suitable for time series data. Real time emotions not classified

C. Social Media Post Based Model

Emotion state can be recognized by analyzing the social media post, profile pictures, tags, stories, shared quotes etc. as people used to share their opinion and mood on this platform in their day-to-day life. Automated detection of human emotions can help to recognize the people suffering from mental illness and can suggest the treatment for the same. In paper [9], author considered flicker dataset and overcome the problem of "missing modality" by introducing the concept of unimodal, bimodal and tri modal classifiers. In paper [10], author analysed the twitter data by processing visual, textual and user-interaction data. In paper [11], Raddit data was used as an input and feature extraction was done through (LIWC+LDA+bigram) techniques. Then extracted features fed to the SVM classifier.

TABLE III SOCIAL MEDIA BASED LITERATURE SUMMARY

Dataset	Classifier used	Year	Preprocessing	Advantages	Challenge
Flickr	CNN	2019 [9]	Missing data handling approach used	highest accuracy obtained for Image ,Text and tags which is 94.80%	Flickr is very small dataset so it can be biased.
Twitter dataset	Random Forest	2019 [10]	Feature Selection	Good accuracy compare to other approaches	Classification done in two categories depressed and control.
Reddit	LR,SVM, MLP,RF	2019 [11]	Tokenization,Removal of URL	highest accuracy obtained which is 91.00% when MLP used with combined features	Classification done in binary which predict whether person is depressed or not.

D. Audio Recognition Model

Some words like "Humm" or "Hey" of text cannot express emotions but using audio recognition model ,we can find the state of emotions. The major application area of audio recognition is used for detecting drowsiness of driver to avoid accident caused by the mental state of him. It can also be implemented to find the depression or suicidal risk by taking user interaction on mobile communication as an input. Egor Lakomkin1, Mohammad Ali Zamani [12] have proposed the novel approach of adding external noise to the data which will improve the robustness of the model. Mingyi Chen, Xuanji He[13] author have used the attention mechanism to identify the emotion relevant frames from the audio stream as all frames are not important like silent frames. In paper [14] author have developed two independent network to extract the local and features from speech and log-Mel spectrogram. Table IV Shows the comparative analysis of different literatures of audio recognition model.

TABLE IV AUDIO RECOGNITION LITERATURE SUMMARY

Dataset	Classifier used	Year	Advantages	Challenge
IEMOCAP	RNN ,CNN	2018 [12]	RNN and CNN both model were implemented where RNN approach for iCube robot achieve 83.2% accuracy.	Noise added externally which can't give same effect like real noise generated
Berlin EmoDB and IEMOCAP	3D CNN ,LSTM	2018 [13]	Model achieves an overall 86.99% accuracy. Attention mechanism used.	Quite Complex Testing should be done on other dataset.
Berlin EmoDB and IEMOCAP	CNN ,DBN, LSTM	2019 [14]	Deep 1D and 2D CNN LSTM to achieve 91.6% and 92.9% accuracy. Emotions are classified in real time category.	"Black box" approach as deep learning used. Over fitting of data still exist.

III.MULTIMODAL EMOTION RECOGNITION

With the advent use of big data many modes of data available online for the analysis task which can be collected from product review, customer feedback, video posted online, social media forum discussion and many more. This multimodal data is more accurate than signal modal data as many times single model information is unavailable and not so accurate. Normally people express their emotions in a multimodal ways like they speak, they act, they share opinion etc.so multimodal fusion is essential to accurately analyse the emotions of human being. Many research have been done in multimodal emotion recognition. Fig.1 shows the chart of different modalities used by authors over years where A stands for Audio analysis, V stands for Visual expression recognition and T for text analysis.

A. Multimodal Fusion Techniques

As the result obtain by single classifier is almost less robust and not so accurate, there is a need to include different feature extraction techniques or fuse the results obtain by different classifier. Multimodal emotion fusion is the process combining results of various modalities like video, audio, social media post or EEG signal. There are mainly two techniques available for data fusion:- Feature Level fusion(early fusion) and decision Level fusion (late fusion). In feature level fusion, features extracted from different emotion modals map to a new high dimension feature vector. As the new feature vector is of high dimension, dimensionality reduction method is used to reduce the redundant features from a feature set. This reduce feature set is then given to classifier to identify the emotion.

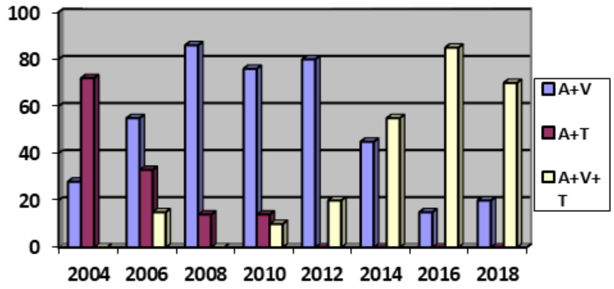

Fig.1 Chart of Different modalities over years

In decision level fusion, Most appropriate classifier is chosen for each modality and the output of different classifiers are combined to identify accurate emotions. Different classifiers are expert in different domain and thus ensemble method is chosen to accurately identify the emotion class. Fig.2 shows the chart of fusion techniques used over years.

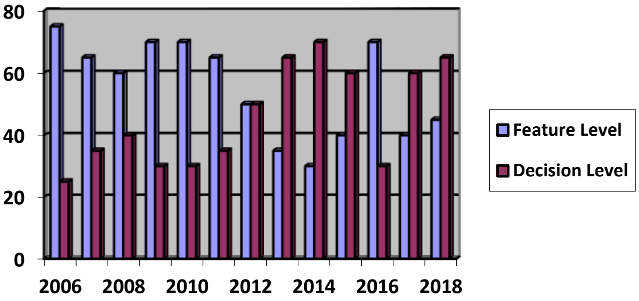

Fig.2 Chart of fusion methods over years

B. Multimodal Emotion Dataset

Many multimodal emotion datasets are available .Table V shows the list of available dataset.

Sr.no	Dataset	Modality	Emotion Label
1	RECOLA	Audio, visual ,ECG, EDA	Arousal and Valence
2	CMU-MOSEI	Text(sentiment), Audio ,visual	Anger Disgust, Fear, Happy, Sad ,Surprise
3	TFD and custom	Audio, video,	Anger Disgust, Fear, Happy, Sad ,Surprise, neutral
4	DEAP	EDA,PPG,EMG psychological signal	Happy, Relaxed, Disgust, sad, neutral
5	AFEW	Audio, video	Happy, sad, anger, fear, disgust, surprise and neutral
6	EMOEEG	EEG,EOG,EMG,ECG,EDA	Valence and arousal
7	IEMOCAP	Audio, visual	anger, happiness, sadness, neutrality

TABLE V LIST OF MULTIMODAL DATASET

IX. CONCLUSION

This paper, carried out a systematic review of available modalities of emotion recognition. We started by discussing an state-of-art in video, audio, EEG and social media based emotion recognition. To develop Efficient multimodal emotion recognition system, each unimodal classifier should be accurately develop. So that we provide the overview of different unimodal emotion recognition. Then we discussed multimodal emotion recognition and different fusion techniques like feature level and decision level followed by overview of available datasets.

X. REFERENCES

[1] Zhang, K., Huang, Y., Du, Y., & Wang, L. (2017)," Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," IEEE Transactions on Image Processing, 26(9), 4193–4203.

[2] Talipu, A., Generosi, A., Mengoni, M., & Giraldi, L. (2019)," Evaluation of Deep Convolutional Neural Network architectures for Emotion Recognition in the Wild," 2019 IEEE 23rd International Symposium on Consumer Technologies, ISCT 2019, 25–27.

[3] Taha, B., & Hatzinakos, D. (2019),"Emotion Recognition from 2D Facial Expressions", 2019 IEEE Canadian Conference of Electrical and Computer Engineering, CCECE 2019, 1–4.

[4] Jain, D. K., Shamsolmoali, P., & Sehdev, P. (2019),"Extended deep neural network for facial emotion recognition", Pattern Recognition Letters, 120, 69–74.

[5] Zheng, W. L., Zhu, J. Y., Peng, Y., & Lu, B. L. (2014),"EEG-based emotion classification using deep belief networks. Proceedings", IEEE International Conference on Multimedia and Expo, 2014-Septe(Septmber).

[6] Guo, K., Candra, H., Yu, H., Li, H., Nguyen, H. T., & Su, S. W. (2017),"EEG-based emotion classification using innovative features and combined SVM and HMM classifier", Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 489–492.

[7] Kumar, S. D., & Subha, D. P. (2019),"Prediction of depression from EEG signal using long short term memory(LSTM)",Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, 2019-April(Icoei), 1248–1253.

[8] Pallavi Pandey and K. R. Seeja(2019)," Subject-Independent Emotion Detection from EEG Signals Using Deep Neural Network", S. Bhattacharyya et al. (eds.), International Conference on Innovative Computing and Communications, Lecture Notes in Networks and Systems 56, © Springer Nature Singapore Pte Ltd. 2019.

[9] Mathieu Pagé Fortin and Brahim Chaib-draa(2019)," Multimodal Multitask Emotion Recognition using Images, Texts and Tags", WCRML '19- Proceedings of the ACM Workshop on Crossmodal Learning and Application.

[10] Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Monadjemi, A., Thirunarayan, K. Pathak, J. (2019). Fusing Visual, Textual and Connectivity Clues for Studying Mental Health. Retrieved from http://arxiv.org/abs/1902.06843.

[11] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019)," Detection of depression-related posts in reddit social media forum", IEEE Access, 7, 44883–44893. Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., & Wermter, S. (2018),"

[12] On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks". IEEE International Conference on Intelligent Robots and Systems, 854–860.

[13] Chen, M., He, X., Yang, J., & Zhang, H. (2018)," 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition", IEEE Signal Processing Letters, 25(10), 1440–1444.

[14] Zhao, J., Mao, X., & Chen, L. (2019)," Speech emotion recognition using deep 1D & 2D CNN LSTM networks", Biomedical Signal Processing and Control, 47, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035.

[15] Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019)," Human emotion recognition using deep belief network architecture", Information Fusion, 51(November 2018), 10–18.

[16] Sahay, S., Kumar, S. H., Xia, R., Huang, J., & Nachman, L. (2019)," Multimodal Relational Tensor Network for Sentiment and Emotion Classification", 20–27..

[17] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks", IEEE Journal on Selected Topics in Signal Processing, 11(8), 1301–1309.

[18] Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2015).,"Long short term memory recurrent neural network based multimodal dimensional emotion recognition", AVEC 2015 - Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Co-Located with MM 2015, 65–72.

[19] Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K. Bengio, Y. (2016). "EmoNets: Multimodal deep learning approaches for emotion recognition in video", Journal on Multimodal User Interfaces, 10(2), 99–111

[20] Yu Kong, Member, IEEE, and Yun Fu, Senior Member, IEEE, "Human Action Recognition and Prediction: A Survey",2018 JOURNAL OF LATEX CLASS FILES, VOL. 13, NO. 9, SEPTEMBER 2018-IEEE.

[21] Egger Maria1, Ley Matthias1, Hanke Sten," Emotion Recognition from Physiological Signal Analysis: A Review", 1571-0661/© 2019 Published by Elsevier B.V. Electronic Notes in Theoretical Computer Science 343 (2019) 35–55.