EasyChair Preprint

№ 7738

A Review of Deep Learning models for Facial Emotion Detection

Deepshikha Mehta, Shweta Barhate and Mahendra Dhore

April 9, 2022

A Review of Deep Learning models for Facial Emotion Detection

1st Deepshikha Mehta
Reasearch Scholor
Departement of electronincs and Computer Science, RTMNU
Nagpur, India
deepshikhamehtsjoshi@gmail.com

2nd Shweta Barhate
Assistant Professor
Departement of electronincs and Computer Science, RTMNU
Nagpur, India
shwetab73@yahoo.com

3rd Mahendra P. Dhore
Professor and Principal
SSESA's Science College
Nagpur, India
dhoremp@gmail.com

Abstract*— In the recent years facial emotion recognition has been majorly a very powerful topic among researchers, as it has an impactful contribution in effective Human Computer Interaction. Facial emotion recognition is part of affective computing that enables computers to understand human emotions and respond accordingly. This paper provides a brief review of Deep learning models that can be used to enhance the limitations of facial emotion recognition issues. The focus is on up-to-date deep learning approaches such as Convolutional Neural Network, Recurrent Neural Network, Transfer Learning and Generative Adversarial Network. The purpose of this paper is to assist and guide researchers by providing insights and future directions in terms of enhancing this discipline.

Keywords*—Facial emotion detection, Deep Learning, Survey, Deep learning models*

1. INTRODUCTION

Emotions are an inextricable aspect of human communication that aids us in understanding others' intentions. Emotions can be expressed in a variety of ways that may or may not be visible to the human eye. Verbal component of emotion convey one-third of human communication whereas nonverbal component conveys two-thirds. Emotion recognition aims to establish effective human computer interactions by providing them a better way to recognize and understand human emotions for an operative interaction between them. This can be implemented in technologies where computers need to understand the emotional state of the user and respond accordingly. The most perceptible human emotion way is through facial expression. In general, the face provides three sorts of signals: static, slow, and quick signals. [1] Skin colour, which encompasses various long-term components of face skin pigmentation, greasy deposits, face forms, the composition of bones, cartilage and shape, and the placement and size of facial features such as brows, eyes, nose, and mouth, are all static signals. Permanent wrinkles, as well as changes in facial appearance such as muscle tone and skin texture changes that occur over time, are examples of slow signals. The face muscles movement, impermanence face aspect changes, impermanent wrinkles, and shifts in the position and form of facial features are all examples of fast signals. These signals are visible on face only for few seconds and humans are capable for hiding their genuine feeling in any situation. In early twentieth century Ekman and Fiesen [2] defined six basic emotions sad, anger, fear, disgust, happiness and surprise and also showed that humans can express their feelings in a certain way irrespective of their culture. Deep learning modals can be applied for the classification of human facial expressions.

2. RESEARCH METHODLOGY

Systematic Literature Review guidelines have been used to review existing studies by using diverse stages of identification, analysis and interpretation concerning to research questions framed in an proper manner.

A. Research Questions

The research questions provide a baseline to entire research process which includes the data collection, analysis and inclusion criteria.

RQ1: What is the state of art methods in facial emotion recognition?

RQ2: What are available Deep Learning model with image processing application which can be implemented in facial emotion recognition ?

RQ3: What are the limitations in existing facial emotion recognition systems developed?

B. Data Extraction

By considering the identified research questions the different techniques and challenges with state of art papers were selected for the systematic review.

C. Inclusion and Exclusion Criteria

Articles were included by searching keywords, and title of the papers that where related to facial emotion recognition with deep learning modals. Further non peer reviewed and duplicating, low quality papers were excluded.

3. LITERATURE REVIEW

In the recent years Deep Learning based techniques and architectures have been used for precise facial emotion recognition modals. In this section of the paper related state of art research work has been selected for review. This will help the researchers apply the appropriate methodology according to the objective framed for their work.

Khattak et al. in [3] proposed deep learning based convolutional neural network for emotion classification and detecting gender an d age from facial expressions. Layer and parameter setting were revised in the proposed CNN modal. The architecture consists of two convolutional layer with max pooling layer and a flatten layer for creation of feature vector. The first convolutional layer extracted low level features and second extracted high level features. It was suggested that other deep learning modal should also be used as extension of their work.

Said et al. in [4] proposed a face-sensitive convolutional neural network for facial emotion recognition. The proposed Convolutional neural network has three main functions, localize faces in images and cropping faces image, analyzing the expressions, and recognize emotions. The architecture is composed of 15 convolution layers with different kernel sizes and strides. The detection task is performed at different feature map by reducing the size of the last feature map gradually and applying the prediction at each feature map. The datasets used are CelebA dataset which contains 202,599 RGB images of celebrities in the world and the UMD faces dataset which has images from the internet using famous search engines. Most of the existing object detection models are based on taking ground truth boxes with their associated classes but the proposed model was designed to analyze face attributes to generate emotion predictions. This technique enhanced the overall performance when compared with classical object detection models.

Minaee et al. in [5] proposed a deep learning methodology based on attentional convolutional network which is focusing on vital parts of the face. This included multiple datasets such as FER-2013, CK+, FERG, and JAFFE. The proposed methodology is based on an attentional convolutional network, which can focus on feature-rich areas of the face. Visualization techniques is used to select the most relevant region which are the parts of the image that have the strongest influence on the classifier's outcome. For recognizing facial emotions, special attention to specific areas is critical, since neural networks with fewer than ten layers may compete with (and even beat) much deeper networks in emotion identification.

Lui at al. in [6] proposed recognition of emotions based on two dimensional modal detecting valence and arousal. The convolutional neural network was designed with four convolutional layers, three max pooling layers and 3 fully connected layers. The architecture was designed to detect valence dimension with nine classifications and it was observed that there was no significance difference between adjacent valence dimensions in actual emotion. Facial expression using convolutional neural network gave significant good results but it was observed that it needed lot of data for achieving better performance.

Min Wu at al. in [7] proposed a fusion based convolutional neural network for detection of emotions using facial expressions and speech. LBP-TOP was employed to extract the low level dynamic emotion features of facial expression and deep convolution neural network was used extracting high level emotion semantic features. The deep convolution neural network comprise of convolution layer, max-pooling layer, full connection layer, and Softmax regression in stack. The focus of this modal was on speech data which were combined with facial expression result to achieve high accuracy.

Liu et al. in [8] proposed an action unit based attention mechanism for facial emotion recognition. It works on the mechanism for self-attentional based method to extract single feature from any position using action units. CK+, Oulu CASIA, MMI, and AffectNet datasets were used for training and testing. The system integrated the spatial and temporal attention modules in the proposed CNN RNN architecture. The action units attention module based on facial action coding system and attentive pooling module based spatial feature aggregation was used.

Sadeghi et al. in [9] proposed feature extraction based on Gabor filter to reduce feature dimensionality for facial expression recognition. Gabor filter are convolved with the image and the results are coded in matrix which is further divided into several blocks. The histogram of the blocks is used to create the final feature vector. Support vector machine is used for classification. The datasets used are CK+, SFEW, MMI and RAF-DB. This feature extraction method outperformed the state-of-the-art texture descriptors on these datasets.

Ghofrani et al. in [9] proposed multitask convolutional neural network along with ShuffleNet V2 architecture. It was observed that computational cost of ShuffleNet is very less than ResNeXt or Xception. Data augmentation was done to overcome asymmetric distribution. The overall accuracy of this architecture was low compared to state of art systems.

Hussain et al. in [10] proposed Convolution Neural network based architecture VGG 16 for appropriate feature extraction and classification for facial emotions. Face is detected using haar cascade classifier the face features are extracted for face recognition and then the underlining emotion is recognized from the facial expression. The architecture VGG 16 is fabricated with CNN model for large

database recognition and classification. The system was trained using KDEF dataset to classify seven emotions.

Xie et al. in [12] proposed deep multi path convolutional neural network with relevant region attention for facial expression recognition. Deep Attentive Multi-path Convolutional Neural Network (DAM-CNN) includes feature extraction and attention based SERD and multipath variation network. Feature extraction is done using VGG Face is a CNN having 16 convolutional layers, 5 pooling layer and 3 fully connected for face recognition and pre trained VGG-Face modal for feature extraction. The extracted features are forwarded into attention network by weighing the features by attention mask which will preserve spatial information. Multi-path auto encoder is used to achieve reconstruction of samples among all decoders.

Table 1: State of Art Deep Learning Techniques

Ref. Year	**Objective**	**Techniques**	**Dataset**	**Limitation**
[3] 2022	Deep learning based Convolutional neural network for classifying age and gender	Two convolutional layer with max pooling layers and flatten layer	JAFFE and CK+	Overfitting and pre trained modal could be used.
[4] 2021	Face sensitive CNN (FS-CNN) to detect faces in high resolution images and recognize emotions	15 convolutional layers with different kernel sizes and strides	CelebA and UMD faces	Recognition based on face attributes only. Confusion between anger and fear emotion.
[5] 2021	Attentional Convolutional Neural Network to classify the underlying emotion in facial images.	4 CNN with two max pooling layer and a Rectified linear unit activation function	FER2013, CK+, JAFFE	Cross dataset study could have been done
[6] 2020	Facial expression based dimensional emotion recognition based on CNN	CNN with 4 convolutional layer. Introduced ADAM algorithm	CK+, FER2013	Improve the performance of valence. Insufficient no of images in dataset
[7] 2020	Dynamic emotion recognition by using both facial expression and speech modalities	Fuzzy fusion based two stage neural network using DCNN LBP-TOP and spectrogram	SAVEE, eNTERFACE' 05, and AFEW	confusion appears between sadness and neutral. close to real world scenarios with multiple features should be used
[8] 2020	Proposed a novel action-units attention mechanism tailored to FER task to extract spatial contexts from the emotion regions	Two dimensional modal and CNN with 4 convolutional layer with ReLu. Self-attentional mechanism used	CK+, Oulu-CASIA, MMI	Pairwise sampling strategy needed, overfitting
[9] 2019	Appearance based feature extraction method is proposed	Gabor filter convolutional coefficients of each pixel to reduce feature dimensionality	CK+, SFEW MMI, RAF-DB	Only one feature taken and intra class variation in dataset
[10] 2019	Integrated the face detection and emotion recognition to develop a combined system which can perform the facial recognition and emotion extraction in real time	Multi-Task cascaded Convolutional Neural Network and ShuffleNet architecture	FER2013	Overall accuracy of system was low.
[10] 2019	Face detection, recognition and emotion classification in real time images	CNN based VGG16 architecture is used	KDEF	Overfitting
[12] 2019	Two novel modules: an attention-based Salient Expressional Region Descriptor (SERD) and the Multi-Path Variation-Suppressing Network (MPVS-Net)	VGG-Face network for extracting features, SERD for refining CNN features and highlighting salient expressional regions	CK+, JAFFE, TFEID, BAUM-2i, FER2013	Overfitting, Need for deep network

4. DEEP LEARNING ARCHITECTURES

In this section the Deep learning modals and architectures are briefly summarized with orientation to their application in processing facial images according to the referenced papers.

Deep learning has been a very popular topic with researchers as it has achieved state of art performance in many

application areas. Deep Learning is an end to end learning method which internments high level abstractions with the help of hierarchical architecture. [13] Deep learning is a subset of machine learning which is a standard prototype that represents the functioning of human brain. This typically consists of neural network model where neurons are responsible to act as inputs and each of them are connected to outputs.

D. Convolutional Neural Network(CNN)

In the field of Deep learning , the CNN is the most renowned and commonly used algorithm. The main advantage of CNN compared to other methods used earlier is that it automatically categorizes the relevant features without any human supervision [14]. CNNs have been extensively applied in a range of different areas, including computer vision, speech processing, Face Recognition, image processing, facial emotion recognition etc.

The structure of CNNs was inspired by neurons in human and animal brains, similar to a conventional neural network [14] Convolutional Neural Network includes two main parts Feature extractor and a classifier [15] In Feature Extractor each layer passes its output to the input of next layers and this process selects the features in depth that required for the system. The CNN architecture consists of three layers: convolutional layer, max-pooling layer and classification layer. The output extracted from convolutional layer is a crucial step for feature extraction. Convolutional layers are always even numbered and max pooling layer is odd numbered. The output of convolutional layer and max pooling layer are called as feature mapping. [15] [16] Higher features are extracted from the lower layers of CNN. To ensure classification accuracy, the number of feature maps is frequently increased to reflect better characteristics of the input images. The classification layer takes the input from the output of the last layer of the CNN. It takes the features extracted from the layers as input with respect to the dimension of the weight matrix. The final layer feature maps are in form of vectors which are passed to the classification layer also known as fully connected layer.

CNNs can be used to avoid overfitting and add generalization using weight sharing feature. The major advantage of using CNN is unlike machine learning algorithms it can perform both feature extraction and classification in single architecture. To enhance the accuracy of the architecture one can adjust the number of layers and other hyper parameters. Model architecture is a contributing factor in enhancing performance of different applications. [14] [15] [16]. There are many CNN architectures like AlexNet, NiN, ZefNet, GoogleLeNet etc. [16]which can be used by researchers in various applications depending upon the requirement of parameters.

E. Recurrent Neural Network(RNN)

Unlike Convolutional Neural Network and Deep Neural Network, RNNs operate on a sequence of vectors. RNNs are is capable of learning long-term dependencies and it a type of long short-term memory. [15] RNN form a chain like structure having four module which are repeating. It helps to transfer information from one step to another in the network that permits information to persevere.

RNN has straightforward fine tuning after combining with other models. It also supports fixed length and variable length input. [15] Grated Recurrent Network is a simple form of Long short term memory, which is very popular in terms of cost for computation. They are also known as lighter version of RNNs as they require less number network parameters. Attention mechanism works on the concept of ignoring irrelevant section of the input data while focusing on features required and this mechanism can be easily combined with RNNs. The initiative of RNNs learns the content of an image. RNNs can be combined with CNNs for information propagation using continuous representation of hidden layers. Each feature extracted using CNN can be passed on to RNN which forms a variable or fixed length vector form. They can also be used in case of temporal dependency in an image. This arrangement of network can give significantly better performance.

F. Tranfer Learning

Transfer Learning model learns the weight and bias after training the network with large amount of data. The weights can be further used to retrain and can even be transferred to other networks. Using this, there is no more need to train the network again which reduces the computational cost drastically. [15] Pre trained modals can be used on new dataset which is similar or different from the one used while training. The Training and testing data can have two types of domain: target domain and source domain. The target domain comprises of the testing instance and source domain comprises of the training instances [18] For applying transfer learning one needs to address the issue of when to transfer, what to transfer and how to transfer. [19]

With Transfer learning one can overcome challenges such as limited data samples, cross-domain learning, scare label and mismatch.

5. CONCLUSION

This study provided a broad overview of the state of art methods in facial emotion recognition. The study of facial emotion recognition is very important in affective computing and human computer interaction, so there is vital need of inculcating recent deep learning methodologies into it. Recent trends in Deep learning show that collection of different

architectures will help in improving the accuracy of the system. Multistage architecture like combination of CNN and RNN should be implemented for feature extraction and classification. There is lot of scope for improvement in designing an efficient architecture that includes the fusion based method for facial emotion recognition. Deep learning modal require large datasets to train them but there is no large dataset available. So researchers have focused on small scale networks. Data augmentation can be done to avoid overfitting and increasing the size of the available datasets. The pre trained modals have produced state of art results in different areas, hence they should be used to train facial emotion detection systems. Transfer learning techniques can solve major issues of existing systems as it has lot to offer in this domain.

6. REFERENCES

[1] I. Revina and W. S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques," *Journal of King Saud University - Computer and Information Sciences,* vol. 33, no. 6, pp. 1319-1579, 2018.

[2] P.Ekman, "Constant across cultures in the face and emotion," *Journal of personality and social psychology* , vol. 17, no. 2, pp. 124-129, 1971.

[3] A. Khattak, M. Z. Asghar, M. Ali and U. Batool, "An efficient deep learning technique for facial emotion," *Multimedia Tools and Applications,* vol. 81, no. 1, pp. 1649-1683, 2022.

[4] M. B. Yahid Said, "Human Emotion recognition based on facial expression via deep learning on High Resolution Images," *Multimedia tools and Applications, Spinger,* vol. 80, no. 1, pp. 41-53, 2021.

[5] S. Minaee, M. Minaei and A. Abdolrashidi, "Deep Emotion: Facial Expression Recognition Using Attention Convolutional Network," *Sensors,* vol. 21, no. 1, pp. 1-16, 2021.

[6] S. Liu, D. Li, Q. Gao and S. Y, "Facial Emotion Recognition Based on CNN," *Chinese Automation Congress (CAC), Shanghai,* pp. 398-403, 2020.

[7] M. Wu, W. Su, W. Chen, W. Pedrycz and K. Hirota, "Two-stage Fuzzy Fusion based-Convolution Neural Network for Dynamic Emotion Recognition," *IEEE Transactions on Affective Computing,* pp. 91-114, 2020.

[8] D. Lui, X. Ouyang, S. Xu, P. Zhou, K. He and S. Wen, "SAANet: Siamese Action-units Attention Network for Improving Dynamic Facial Expression Recognition," *Neurocomputing, Elsevier,* vol. 413, pp. 145-157, 2020.

[9] S. Hamid and R. Abolghasem, "Human vision insprised feature extraction for facial expression recognition," *Multimedia Tools and Apllications,* vol. 78, no. 1, pp. 30335-30353, 2019.

[10] A. S. Hussain and A. S. A. Al Balushi, "A real time face emotion classification and recognition using deep learning modal," *Journal of Physics: Conference Series,* vol. 1432, p. 13, 2019.

[11] S. Xie, H. Hu and Y. Wu, "Deep multi-path convolutional neural network joint with salient," *Pattern Recognition,* vol. 92, no. 1, pp. 177-191, 2019.

[12] A. Ghofrani, R. M. Toroghi and S. Ghanbari, "Realtime Face-Detection and Emotion Recognition Using MTCNN and miniShuffleNet V2," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, Iran University of Science and Technology, Tehran, Iran, 2019.

[13] S. Li and W. Dong, "Deep facial Expression Recognition: A Survey," *IEEE Transactions on Affective compute,* pp. 1949-3045, 2020.

[14] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, S. J, F. Mohammed A, M. Al-Amidie and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data,* vol. 8, p. 53, 2021.

[15] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. V. Essen and A. A. S. Awwal, "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics,* Vols. 8,292, pp. 1-67, 2019.

[16] Z. Li, W. Yang, S. Peng and F. Liu, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE transactions on neural networks and learning systems,* vol. 1, pp. 1-22, 2020.

[17] B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors* , vol. 2018, no. 1, p. 20, 2018.

[18] L. Shao, F. Zhu and X. Li, "Transfer Learning for Visual Categorization: A survey," *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEM,* vol. 26, no. 5, pp. 1019-1034, 2015.

[19] S. J. Pan and Q. Yang, "A Survey on transfer learning," *IEEE Transaction Kowledge Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[20] S. Agianpuye and J.-L. Minoi, "3D Facial expression synthesis: A survey," in *2013 8th International Conference on Information Technology in Asia (CITA)*, Kota Samarahan, Malaysia, 2013.

[21] K. Wang, X. Peng, J. Yang, D. Meng and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE TRANSACTIONS ON IMAGE PROCESSING,* vol. 29, no. 1, pp. 4057-4069, 2020.

[22] M. Kaur and A. Mohta, "A Review of Deep Learning with Recurrent Neural Network," in *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2019.

[23] C. Srinidhia, O. Cigab and A. Martel, "Deep neural network models for computational histopathology: A

survey," *Medical Image Analysis, Vol. 67, 2021,* vol. 67, pp. 1-46, 2021.

[24] L. Shao, F. Zhu and X. Li, "Transfer learning for visual categorization: A Survey," *IEEE Transaction on Nueral Network and Learning Systems,* vol. 26, no. 5, pp. 1019-1034, 2015.

[25] S. Shaees, H. Naeem, M. Arslan and M. R. Naeem, "Facial Emotion Recognition Using Transfer Learning," in *International Conference on Computing and Information Technology (ICCIT-1441)*, Tabuk, Saudi Arabia, 2020.