Comparative Review on Methods of Facial Emotion Recognition

Kuenzang Choden[1], Karma Tensel[2], Tashi Darjay[3], Sangay Yonten[4], Moujhuri Patra[5]

Department of Information and Technology, College of Science and Technology.
* E-mail: [1]0217537.cst@rub.edu.bt, [2]0217513.cst@rub.edu.bt, [3]0217527.cst@rub.edu.bt,
[4]0216526.cst@rub.edu.bt, [5]moujhuri@gmail.com

Abstract

Emotion detection using facial recognition is widely used in robotics, psychology, gaming, and security. Different models and methods yield different accuracy and performance depending on the dataset used, their construction and the hyperparameters set. This paper includes a comparative based review of two models namely Deep-Emotion and MicroExpNet based on Attentional Convolutional Network (ACN) and Knowledge Distillation (KD) respectively which are both based on Convolutional Neural Network (CNN). The paper presents the comparison of the two Facial Emotion Recognition models through a literature review. The comparative review is based on their network architecture, their results, and hyperparameters. This review provides the idea of the applicability of the models in different fields as both of the models proves to be better in their domain. The MicroExpNet model due to its size and speed is known to have a future scope in mobile deployment and provides valuable insights to the development of microarchitecture whereas Deep-Emotion using ACN detects emotion in relatively fewer layers of CNN. This paper further contributes to laying a foundation for future implementers to better understand which model to implement concerning the area of their application.

Key Words: *Facial Emotion Recognition, Convolutional Neural Network (CNN), Attentional Convolutional Network (ACN), Knowledge Distillation (KD), Deep-Emotion, MicroExpNet.*

1. INTRODUCTION

The importance of human expression was understood as early as the 17th century as it has been a studied research topic of physiology by renowned scientists such as Darwin. Mehrabian (1968) stated that humans communicate 7% using linguistic language (verbal part), 38% by paralanguage (vocal part), and 55% by facial expression (as cited in Khandait et al., 2011). After the Facial Coding System by Ekman and Friesen in 1970, many methods have been developed and examined to detect facial emotion. Among many methods for FER, CNN is a sought method as it provided high accuracy as shown by Khorrami (2015) (as cited in Minaee & Abdolrashidi, 2019).

Many contributions have been made for FER using CNN. The methods claim their accuracy on the performed dataset. CNN implemented differently by different models poses a question as to which method is better as they are not implemented on the same quantity and type of dataset. This paper provides a thorough comparison of how such two methods work and performs.

This paper presents a comparative review of models for facial recognition systems (FER). Deep-Emotion and MicroExpNet based on Attentional Convolutional Network (ACN) and Knowledge Distillation (KD) were considered in our study. Both models are based on Convolutional Neural Network (CNN). The comparison is drawn based on their network architecture, results, and hyperparameters. The rest of this paper is organized as a Literature review in section 2, methodology in section 3, and result and discussion are presented in section 4.

2. LITERATURE REVIEW

Investigations were done on Japanese Female Facial Expression(JAFFE) collection and the Automatic Facial Expression Recognition framework based on Neural (Khandait et al., 2011) yields a precision of 96.42 % for 30 tried pictures and 100% exactness for all preparation sets. In another experiment by Prasad M (2015) on the JAFFE database,

a success rate of 93.6% was achieved using the system. The precision added up to around 74 % utilizing this algorithm (on average for all emotions). Burkert et al. (2016) proposed a convolutional neural network (CNN) model for facial emotion recognition that yielded an accuracy of 99.6% for CK+ and 98.63% for MMI datasets performing slightly better than the state-of-the-art. Barsoum et al. (2016) used a deep convolutional neural network (DCNN) and achieved 85.1% accuracy on the FERPlus dataset. Minaee & Abdolrashidi (2019) presented using Attentional Convolutional Network a facial emotion recognition model. Cugu et al. (2019) conducted research using the MicroExpNet model based on the Knowledge Distillation (KD) method. Khorrami showed that CNN can attain high correctness in emotion identification with the use of zero-bias CNN on CK+ datasets and used Toronto Face Dataset TFD to achieve state of the art results (Minaee & Abdolrashidi, 2019). Deep-Emotion and MicroExpNet both implement CNN as ACN or through knowledge distillation.

A comparative study on a few techniques of mood detection was provided by Gavde and Pednekar (Gavde & Pednekar, 2018). The paper was a general comparison of papers from the past years, the latest paper being of 2016. The general comparison provided the different database used, feature extraction technique and the classification method. This did not provide a detailed comparison of the methods and also the methods compared are not the latest or in top ranks today. In a paper by N.U. Khan (2012), a comparative analysis of FER techniques was provided. The comparison provided is detailed but the techniques compared are from the year as latest as 2012. Such comparison is not present for techniques developed in recent years.

3. METHODOLOGY

3.1 Selection and Study of Methods

Various papers on facial emotion recognition were identified from Papers with Code. Papers with code is a repository of machine learning research work and codes on the new contributions made. The papers of recent years with innovative implementation methods were selected and further filtered to two for a detailed review.

Two papers taken for the comparative study through the literature review, Deep-Emotion and Micro ExpNet were due to the following reasons:

- Both methods are based on CNN.
- The Deep-Emotion model is based on Attentional Convolutional Network (ACN) and had implemented a challenging dataset; FER2013.
- MicroExpNet is based on Knowledge Distillation (KD), a well-known model compression method.
- MicroExpNet is an extremely small and fast model that has potential applications in mobile devices (Cugu et al., 2019)

The methods are then compared based on the following aspect:

3.2 Examining Network Architecture

The basic network architecture in a CNN comprises convolution layers, activation layers, max-pooling layers, and fully connected layers. The models chosen for comparison are based on the ACN method and KD technique and their network architecture varies and hence compared.

3.3 Inspecting Hyperparameters

Hyperparameters present a crucial role in the determination of the accuracy of a model. The models compared vary from one another in their hyperparameters. Hyperparameters are different for different models depending on how they are designed to be trained.

3.4 Inspecting Dataset

The models in the papers use five different datasets in total; FERG, FER2013, JAFFE, Oulu-CASIA, and CK+. The first model uses four datasets and the second uses two different datasets, only one dataset used is common in both.

3.5 Inspecting and analyzing the output of the model on each dataset

The result of each model was reflected in the paper as well as on the website. The result from the implementation carried out by the researchers of the papers are reflected in this paper and was analyzed.

3.6 Determining Factors that influence the methods

The performance and accuracy of the methods differ due to a few factors. The appropriate factors are identified.

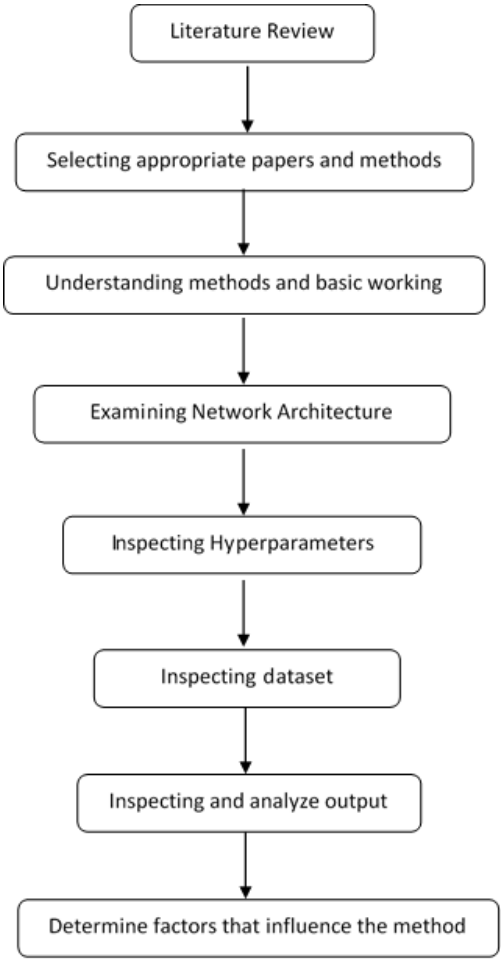

Figure 1 Methodology flow chart

4. RESULT AND DISCUSSION

4.1 Models

4.1.1 Deep-Emotion

Deep-Emotion is the model from the first selected method for comparative study. The model implements the Attentional Convolutional Network. The vitality of regions of the face in emotion recognition; the mouth and the eyes being more important than others such as the ear gave rise to Attention in Convolutional Neural Network in machine learning. This research work was conducted on four different datasets; CK+, FER2013, FERG, and JAFFE. This model is the state-of-the-art for the FERG dataset. The model was trained and tested on the FER2013 dataset which is a challenging dataset. The dataset set contains different number of images on all emotions.

Attentional Convolutional Network is a Convolutional Neural Network that focuses on important regions to carry out recognition or classification. To determine the significant regions in ACN, the trained images are subjected to a square region of size n*n and zeroed out (Minaee & Abdolrashidi, 2019). If the absence of that region does not make a correct prediction, that region is considered an important region for the particular emotion recognition. Carrying out this throughout the image provides a saliency map showing the important regions in emotion recognition. In the Facial Emotion Recognition system, Attention in CNN allows the recognition of emotion to be fulfilled in lesser layers on the neural network as it focuses on a more important region as per the given task.

4.1.2 MicroExpNet

MicroExpNet is the model trained from the second selected method for comparative study. MicroExpNet is a student network model created in the research work to implement facial emotion recognition of two datasets namely CK+ and Oulu CASIA. Knowledge distillation is applied in this research work to create student networks for emotion recognition. MicroExpNet is the smallest student network which is recognized as the state of the art for the Oulu CASIA dataset.

Knowledge distillation is a method in which a smaller neural network is trained from a pretrained larger network. Hinton et al. expressed that KD is the first and the most well-known teacher-student style compression method (Cugu et al., 2019). The teacher network used is Inception_v3 which the MicroExpNet mimics.

4.2 Network Architecture

4.2.1 Deep-Emotion

The model is composed of two different parts; the feature extraction part and the spatial transformer module. The feature extraction part comprises four convolutional layers, every two followed by the max-pooling layer and the activation function, rectified linear unit (ReLU). They are then trailed by a dropout layer and two fully connected layers. The spatial transformer comprises two convolution layers each followed by max-pooling and ReLU and two

fully connected layers. The spatial transformer tries to focus on important regions of the face.

4.2.2 MicroExpNet

To train the student network, Inception_v3 was used as the teacher model for training using the KD method. Four (M, S, XS, XXS) student networks were created out of which MicroExpNet is the smallest student network which is 1MB in size. The activation function employed was ReLU (Rectified Linear Unit). The network has two convolutional layers (conv1, conv2) each followed by a max-pooling layer. The network has two fully connected layers (fc1, fc2). The difference between the student networks is in the number of neurons in the first fc layer and the number of parameters (size).

4.3 Hyperparameters

4.3.1 Deep-Emotion

The model is trained for 500 epochs with a learning rate of 0.005 and the optimizer used was the Adam optimizer.

4.3.2 MicroExpNet

The student model is trained for 3000 epochs with a learning rate of 0.0001 and the optimizer used was the Adam optimizer with a dropout rate of 0.5. The batch size employed was 64 and the weight of distillation taken is 0.5.

4.4 Inspecting Datasets

4.4.1 FER2013

FER 2013 dataset has a total of 35,887 48*48 images with seven classes or emotions including the neutral. It is one of the challenging datasets as not all images are frontal. The images have occlusions, some are not full images, some have low contrast and some have eyeglasses in the images.

4.4.2 FERG

FERG dataset is a collection of 55767 animated facial expression images. The images are in seven different facial expressions including the neutral.

4.4.3 CK+

The CK+ dataset contains a mixture of posed and spontaneous facial expressions of 123 people. The dataset comprises 1574 images in eight different facial expressions including the neutral. Contempt, which is considered to be a difficult emotion to detect, is the extra emotion considered in this dataset.

4.4.4 Oulu CASIA

Oulu CASIA dataset contains images that were taken in three different illumination conditions. There are a total of 480 images which was posed by 80 subjects with six classes.

4.4.5 JAFFE

JAFFE is a Japanese dataset that contains 10 Japanese female posing seven different frontal facial emotions. There are a total of 213 images in the dataset.

4.5 Results

Table 1 Result Analysis

Dataset /Method	FER 2013	JAFFE	CK+	FERG	Oulu-CASIA
Deep-Emotion	70.02 %	70.02 %	98 %	99.3 %	-
MicroExpNet	-	-	96.9 %	-	95 %

The models were trained and tested on different datasets such as CK+, FER2013, JAFFE, FERG, and Oulu-CASIA. The table above summarizes the results expressed as a percentage. For the FERG dataset, the Deep-Emotion model produces the highest accuracy making it the state-of the-art (SOTA). Similarly, for the Oulu-CASIA dataset, the MicroExpNet model produces the highest accuracy making it the state-of-the-art (SOTA).

The Deep-Emotion yields a 70.02% accuracy on the FER2013 dataset, which is a good figure given the challenging nature of the dataset. The model can recognize emotions from faces that are not full, frontal, or both.

MicroExpNet is the smallest student model of Inception_v3. The model provides significant insight for designing a microarchitecture for the FER system despite being less accurate than the state of the art (Cugu et al., 2019).

4.6 Factors

The factors that contribute to the efficiency of the FER models are the hyperparameters taken for training and the datasets used. Hyperparameter tuning plays a significant role and determines how the trained model will perform in the testing phase. The type of dataset is used to train the model also affects accuracy. The accuracy for the FER2013 dataset by Deep-Emotion is comparatively low compared to those for other datasets. The diversity of the images and the accuracy of a model are correlated. A model trained on diverse input would perform better.

5. CONCLUSION AND FUTURE SCOPE

This paper compares two different Facial Emotion Recognition (FER) models. The first model is Deep-Emotion based on Attentional Convolutional Network (ACN) and MicroExpNet which works on Knowledge Distillation (KD) method. Analysis and comparison of the model's network architecture, hyperparameters, and results on different datasets were executed. This paper aimed at a comparative review on two machine learning methods of emotion recognition through facial recognition and so was achieved.

FER has many applications in psychology, security, robotics, and gaming. Much research has been carried out in FER and there is still room for improvement in this field. The FER systems have been progressively improving. The ultimate aim of FER systems is to increase efficiency and accuracy. This achievement will furthermore have positive implications in robotics, psychology, and many more.

The comparative study carried out can benefit an implementer of FER to choose an appropriate method based on their domain of application. The paper will help a researcher in the future gain insight into the problem statement for a new FER system. It provides reviews on two models that were state-of-the-art for two separate datasets; this can provide recommendations on what can be borrowed from these models and methods along with the improvements that can be made.

The datasets used for the training and testing of the two models are inconsistent in their complexity and variety. The characteristics in each of the dataset point to a requirement for the new dataset. For instance, from the CK+ dataset, it is understood that emotion contempt must be present in the dataset for the model to be able to detect difficult emotions. From the FER 2013 dataset, the requirement of occlusions in the dataset is recognized for practical facial emotion detection.

The analysis carried out in this paper lays the foundation for an implementation based comparative review of two machine learning models. From this study, the strength and weaknesses of each model are understood along with their applications. Further, recognizing the diversity of the datasets used during their training and testing prepares a requirement for the dataset for training and testing when carrying out an implementation-based comparison.

The comparative analysis presented in this report is based on a literature review and does not involve any practical implementations. The recommendations and possible future work in this field or topic can be to do a comparative study of the models through implementation. The models can be retested on different existing open-source datasets. Additionally, to draw a clear conclusion on the comparison the models can be tested on a common dataset for fair accuracy comparison and the dataset could be a newly collected one. The models such as MicroExpNet can be modified to detect emotions in images that are not frontal. Student models of large models based on ACN can be trained.

6. ACKNOWLEDGEMENT

We would like to extend our gratitude to Mr. Yeshi Jamtsho for guiding and helping us in our research work. The research panel members have additionally been of great nourishment to the group through the question and concerns put forward during the proposal presentation. It allowed us to refine our objectives and make them implementable in our current scenario.

REFERENCES

Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). *Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution.*

Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., & Liwicki, M. (2016). *DeXpression : Deep Convolutional Neural Network for Expression Recognition*. 1–8.

Cugu, I., Sener, E., & Akbas, E. (2019). *MicroExpNet :*

An Extremely Small and Fast Model For Expression Recognition From Face Images. *Table II*.

Gavde, M., & Pednekar, P. S. (2018). *Comparative Study on Mood Detection Techniques*. *6*(Iv), 1456–1457.

Khan, N. U. (2012). *A Comparative Analysis of Facial Expression Recognition Techniques*. 1262–1268.

Khandait, S. P., Thool, R. C., & Khandait, P. D. (2011). *Automatic Facial Feature Extraction and Expression Recognition based on Neural Network*. *2*(1), 113–118.

Minaee, S., & Abdolrashidi, A. (2019). *Deep-Emotion : Facial Expression Recognition Using Attentional Convolutional Network*.

M, P. (2015). *A study on human facial expressions and mood analysis based on digital image processing techniques.* Retrieved from http://hdl.handle.net/10603/203243