Bioengineering Research

Using Multi-inception CNN for Face Emotion Recognition

Ali Salem Altaher[1*], Zahra Salekshahrezaee[1], Azadeh Abdollah Zadeh[1], Hoda Rafieipour[2], Ahmed Altaher[1]

[1] Florida Atlantic University, CEECS Department, Boca Raton, FL, USA

[2] Memorial University of Newfoundland, Computer Science Department, Newfoundland, Canada

Correspondence to: Altaher A.S. (E-mail: aaltaher2018@fau.edu)

Abstract

One integral and necessary part of human behavior is emotion, which affects the way people communicate. Although human beings can recognize and interpret facial expressions, the identification of correct facial expressions continues to be a key and challenging task by computer systems. The main issues stem from the face's non-uniform design and variations in conditions such as light, facial structure, and posture. Several Convolutional Neural Network (CNN) approaches have been introduced for Face Emotion Recognition (FER), but these methods cannot completely reflect temporal variations in facial characteristics. In this study, we use the CMU face data collection of four types of emotions to provide a method for the identification of facial emotions. Four classes of distinguished emotions are happy, sad, angry, and neutral. Pixel values are fed into a Neural Network with different architecture, and the accuracy of those methods has been compared. Restricted Boltzmann machine (RBM), Deep Belief Networks (DBN), Convolutional Neural Networks (CNN), and multi-inception ensemble Convolution Neural Networks are different methods that are used in this research. We note the latter has considerably higher accuracy compared to other ones. The results obtained from the proposed methods Multi-inception CNN is slightly more than 87 percent while for the Restricted Boltzmann Machine (RBM) model it is 26.1 percent and for Deep Belief Networks (DBN) results are almost the same and slightly more than 26 percent finally the results for simple CNN model is 55 percent.

Received: 18 November 2020, **Accepted**: 02 December 2020
DOI: 10.22034/jbr.2021.262544.1037
Keywords: Face Emotion Recognition, FER, Deep Belief, RBM, multi-inception CNN

1. Introduction

Facial expression is one of human beings' easiest, most natural, and most basic signs for expressing their emotions and intentions and intentions. Emotions are expressed in various ways such as verbal/nonverbal interactions, emotional speech, facial expressions, and body gestures. Among the emotions, the classification of facial expressions is beneficial in terms of the capability of social interactions and communicating with machines. Classification of human facial expressions can be done by the Facial Action Coding system (FACS) structure with Action Units (AU) as its building blocks [39] and FACS describes the action units. Action units are categorized into different types such as Main action units, Head movement action units, Eye movement action units, Emotion action units.

There are several studies on the detections of automated facial expression detection as well as the monitoring of drivers and other human-machine interactions (HCI) are commonly used in social studies. The ability to sense the driver's emotion can help to determine the level of attention to configuring the systems so that systems can be tailored. It is because non-verbal signals are important tools for communication and play a key role in interactions.
Neural networks in several areas have been used, including intrusion detection systems [1, 2] and authentication. Face Emotion Recognition domain also is another area where they have been widely used. Some researchers have used different approaches such as Sparse Recovery Classification and Dictionary Learning for Face Emption Recognition [3, 4].

Ekman and Friesen had described as early as the twentieth century six basic emotions, which indicate that certain fundamental emotions, regardless of culture, were understood by individuals [5]. They classified face emotions into six categories, anger, disgust, fear, joy, sadness, and surprise. Another research has recently argued that the six basic emotional models are cultural and not universal [6].

One of the uprisings and demanding areas of research in the field of computer vision is emotion recognition. Labeling an image based on its expression can be a complex machine job. Some emotions can only be differentiated by a slight variation in facial expressions. Emotions such as disgust and anger are usually expressed in very similar ways.

The expressions of the emotions of each person can be very different, with unique features and facial clues. There may be a wide range of different orientations and positions of people's faces in the process of classifying an emotion. By classifying a smaller subset of clearly distinguishable expressions such as anger, happiness, and sadness, better precision can be achieved.

Facial emotion recognition can be a challenging task in dynamic images since people keep facial expressions for short time and the expression may be changed. Besides, lack of facial expression data is another limitation of automated facial expression detection systems. This issue is more challenging for convolutional neural networks in which computational cost is an additional burden [40].

An image or series of frames containing a face may use different representations to view the face [7]. The face can be viewed as being a whole entity known to be holistic representations. If the face is to be a set of characteristics then the solution is analytical representations. The fusion of holistic and analytical methods is a hybrid approach. The analysis process of facial expression can be done in three main steps: the learning of features, the selection of features, and the classifier construction. More detail on the three steps is in the following:

A) Detecting the face within the frame(s) to be viewed as a whole or as a collection of features or a combination of both.

B) Developing processes to gather data on facial expression from facial images

C) Determining the set of classes (e.g. Facial Action Coding System (FACS) is used for the classification of facial expressions)

FER systems can be classified into two main groups: the facial emotion recognition that is based on static image and those, which are based on dynamic sequences. [8] Dynamic FER provides additional information and it is considered to have a higher degree of recognition than static FER systems however, there are some drawbacks to it. As an example, the dynamic features extracted have different phases and different characteristics of facial expression depending on the individual face. Face Emotion Recognition based on deep learning approaches significantly reduces dependency on face-physical models and other pre-processing techniques. They make learning from the input images directly to the "end-to-end."

The Convolutional Neural Network (CNN), among the several deep learning models, is the most common network used for FER [9]. Three types of layers are

used by CNN: a convolution layer, a max-pooling layer, and a fully interconnected layer. Convolutional layers are the first layer to take the image also known as feature maps as inputs, and then transform the inputs into a series of filter sets in a sliding window to output features maps displaying a face image spatial layout.

Subsampling layers then performs averaging or max pooling on the input feature maps given to reduce their dimensions to minimize the effect of variations and slight shifts and distortions reduce the image's spatial resolution. Conventional approaches use hand-crafted features extracted from the input image [9] Last fully connected layer calculates the class.

DBNs have several hidden layers, which could be used for representation or classification [10, 11]. Deep Belief Networks (DBNs) Compared to linear models, has more complex architecture so these networks can recognize more complex features and interpret them. Geoffrey Hinton et al proposed Deep Belief Networks in 2006. Their approach is a combination of Restricted Boltzmann Machine (RBM). Despite deep learning's high learning ability, problems still exist when applied to FER:

I. To avoid over-fitting, deep neural networks need a large amount of training data.

II. Due to the various characteristics such as age, sex, etc., there are significant inter-subject variations. [8]

This paper is arranged as follows. In the following section, we discuss related field work and study. The algorithm proposed in this work will be discussed later and the results are presented. Finally, we discuss the conclusion and future works.

1.1. Related work

This subject was investigated using traditional methods and using new methods of deep learning. We go through some of the previous works and some more new studies in this segment. The previous researches are categorized into two main sections: Recognition of Emotions using unimodal Information and Recognition of Multi-Modal Information.

A) Recognition of emotions by unimodal data Ioannou et al. [12] introduce an expression recognition approach that works accurately for variations in facial expressions among individuals. In this model, facial features are extracted and classified by using a proposed novel neuro-fuzzy system. This method is designed using the Facial Animation Parameter (FAP) to analyze differences in discrete emotional space and 2D continuous activation-evaluation space, which evaluates facial expressions. Mase [13] uses the significant spots of facial muscles to present an approach for emotion recognition. In this technique, the visual flow was used to extract the muscle movements positioned in every face. By using the K-nearest neighbor classification method, they established an 80% precision range of 4 emotions, which are happiness, anger, disgust, and surprise.

Additionally, Yacoob et al. proposed a similar approach in their study [14] compared to Mase [13]. With each emotion they thoroughly studied the motion associated with the edges of the head, lips, and eyebrows, they were able to design an intermediate and high-level description of the various facial expressions based on their method. By using a rule-based system, six basic feelings could be classified with 88 percent accuracy in their study.

Black et al. [15] have derived parametric models for the form, motion, and expression of the head, mouth, and eyebrows. They could also develop an intermediate and high-level design of facial expressions by using the same methodology as Yacoob et al. [14]. Their model could recognize six emotions with an accuracy of 89%. Trujillo et al. [36] presented an unsupervised method of feature extraction for facial expression recognition of thermal images.

In this method, to localize the face features a bi-modal along with the clustering method is used. In the first

step, the Eigen features are extracted to do the task of feature extraction. Eigen features are the covariance matrix of the probability distribution of facial images. Then a Support Vector Machine is applied to perform that task of facial expression classification.

A Fuzzy Inference System (FIS) Facial Expression model is proposed by Ilbeygi and Shah-Hosseini [37]. In this model, Partially Occluded Facial Images are applied for facial expression recognition. For parameter tuning of membership function, a genetic algorithm is used and it provided acceptable performance. Their study results show impressive precision for facial emotion recognition.

B) Recognition of emotions by Multi-Modal Data, in addition to collecting visual information of the face clues, a hybrid system that is also referred to as a multi-modal facial emotion recognition system collects clues from several other modalities.

Busso et al. [16] explored the use of audio-visual evidence for the identification of four characteristics by using attached markers to the faces to collect visual details. They evaluated feature-level and decision-level fusion methods. They found that the bimodal classifier at the feature-level recognized anger and neural state with high accuracy while bimodal classifiers at the decision-level fusion performed well for emotions of happiness and sadness.

Minaee et al. [17] proposed a deep learning model based on a convolutional network. Their approach focuses on critical areas of the face, which leads to significant improvement in older models for multiple datasets such as CK+[1], FERG[2] and FER-2013[3], and JAFFE[4]. In addition, the authors used a mapping technique that is able to find and separate important parts of the face to detect each emotion with higher accuracy. They demonstrated that different emotions of the face are more related to motions on specific parts of the face based on experimental results.

Wu, and. Al. [10] proposed a facial expression recognition system consisting of a Local Binary Pattern (LBP) combined with an enhanced Deep Belief (DBN) network. Their research focuses on robustness, by using LBP to extract the feature, and an enhanced deep-confidence network to identify and distinguish LBP features. They reported a significant improvement in the rate of recognition for JAFFE datasets.

Liu et al. [18] suggested a novel Boosted Deep Belief Network (BDBN) to execute the three FER stages in one loop in an alliterative way. With proposed models, a set of features can be acquired and chosen to create an improved, efficient classifier, which is useful in characterizing expression-related facial appearance/form changes.

Pan et al. [28] presented a multi-modal hybrid method for speech emotion recognition, names as a multi-modal attention network (MMAN). Their results show that to have a greater speech emotion recognition, visual and textual cues are significantly valuable on the IEMOCAP dataset. One of the benefits of the model is its scalability to beyond three modalities. In this work, the attention technique empowers the model to grater data association. Moreover, it requires a lower number of parameters.
The Multi-modal Conditional Attention Fusion model, which is presented by Chen et al. [29], considers various modalities at each stage. In this model, a Long-short term memory recurrent neural networks (LSTM-RNN) is in charge of capturing dependencies, which those dependencies are a long time. Present input features and information of recent history change the modalities weights automatically. The model is tested on the AVEC2015 dataset and it surpasses other fusion strategies.

As the end-to-end memory networks showed remarkable success on multi-attention emotion recognition, Beard et al. [30] suggested a method for

[1] http://www.consortium.ri.cmu.edu/ckagree/

[2] https://grail.cs.washington.edu/projects/deepexpr/ferg-db.html

[3] https://datarepository.wolframcloud.com/resources/FER-2013

[4] Japanese Female Facial Expression available at http://www.kasrl.org/jaffe_download.html

multi-modal emotion recognition that includes an external shared memory. The memory is renewed with different gated analysis iterations. In this method, visual, audio, and textual data are applied to do the task of emotion recognition. To evaluate the model, they examined it on three large datasets such as CREMA-D, RAVDESS, and CMU-MOSEI. The achievement of this study confirms that joining a global contextualized memory update with a gated memory update facilitates the emotion recognition task.

Formulating the emotion recognition problem like a probability distribution task is done by Zhao et al. [31]. They proposed a framework consists of emotional space and visual space. The framework is called a Weighted Multi-Modal Conditional Probability Neural Network (WMMCPNN). For training and testing, both spaces are applied. Based on emotion space and visual space, a distribution prediction is built. Two series of experiments are designed in this study. The first experiment is conducted to assess uni-modal feature-based approaches in terms of their performance and the second experiment series are designed to compare feature fusion techniques. The proposed framework is compared with some baseline algorithms such as Convolutional Neural Network Regression (CNNR), Conditional Probability Neural Network (CPNN), shared sparse learning (SSL), and Weighted Multi-Modal SSL (WMMSSL). The results of this framework show outstanding performance compared to benchmark emotion distribution prediction methods.

A semi-supervised emotion recognition algorithm by the means of an adversarial network is presented by Liang et al. [32]. The proposed model is a multi-modal semi-supervised method and it includes an enhanced generative adversarial network.
Visual and acoustic modalities are applied to build a semi-supervised multi-modal model. To boost the performance of classification, a combination of unlabeled data fed into GAN and multi-modalities are applied. The improved GAN is consists of a Generator (G) network and a Discriminator (D) network. A noise of size 100-dimensional is fed and changed into a feature map. In the second stage, it becomes the input of the discriminator. Various convolutional layers along with a dropout layer for each convolution layer are embedded in the discriminator. In the end, it contains a fully connected layer added to a softmax layer. Two Convolutional Generative Adversarial Networks are applied to provide visual and acoustic multi-modality.

The fusion of multi-modality is done by concatenation representation of visual and acoustic features. Another softmax and fully connected layer are designed to support the concatenation representation and multi-modality. To evaluate the performance of the proposed model, complementary experiments with different settings are done. The proposed model is compared to some adjusted versions of it such as uni-modal Fully Supervised Baseline (FSBase), two uni-modal Multi-modality Fully-supervised Baseline (MFSBase), Semi-Supervised Baseline (SSBase) with extra unlabeled data, uni-modal Semi-supervised with GANs (SSGAN), Multi-modality Semi-supervised with GANs(MSSGAN).

Experimental results confirm that mixing the multi-modality with unlabeled data fed into the GAN enhances the performance of classification. Among feature representation methods, Self-Supervised Learning (SSL) is a notable method of feature representation. Siriwardhana et al. [33] suggested a method of multi-modalities feature representation based on Self-Supervised Learning called Self Supervised Embedding Fusion Transformer (SSE-FT). Vision, audio, and text are the modalities used in the proposed method. Although the SSL learning method provides a strong feature representation, it suffers from high dimensional feature space and extensive feature size. The SSE-FT applied transformers and graph convolution nets for fusion techniques. CMU-MOSI, CMU-MOSEI, IEMOCAP, which is talking conversation of female and male, and MELD datasets(12,000 Conversations of Friends series) are used to evaluate the Accuracy of the SSE-FT model. The results of this study

demonstrate that SSE-FT outperforms of state of the art similar methods.

2. Method

2.1. Proposed Method

The CMU face image dataset[5] was used here. Four different emotional gestures are represented in this dataset; neutralist, happiness, angriness, and sadness. The matrix representation of the images is used with assigning the related emotional label to them. Figure 1 shows samples of this dataset.

Figure 1. Samples of CMU face images dataset

The dataset consists of 20 objects that include 32 instances for each of them, containing various emotional gestures. The subjects wear sunglasses in many instances as well. The total number of face images is 640, which are white and black colors. The images sizes vary between (128 × 120), (64 × 60) and (32 × 30), we chose the first dimensions as a reference for our work, while with the multi-inception CNN proposed model, whole the dataset were exploited due to unifying the sizes of whole images.
The performances of the following classification models are examined for comparison demands.

2.2. Restricted Boltzmann Machine

One of its features is the simplicity of its structure, two layers, known as a visible layer (denoted by m) and hidden layer (denoted by n). The units denoted by m_0, m_1, m_2,…,m_N and n_0, n_1, n_2, … n_T forms these layers, where N and T indicate the number of units forming the layers, respectively. The Restricted Boltzmann Machine (RBM) structure is shown in below Figure 2.

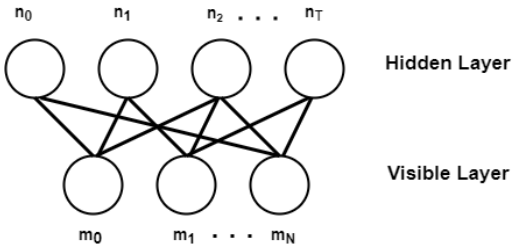

Figure 2. Restricted Boltzmann Machine General structure.

The structure Energy is illustrated via Eq. 1 [19] below.

$$E(m,n) = -\sum_{i=1}^{N} a_i m_i - \sum_{j=1}^{T} b_j n_j - \sum_{i=1}^{N} \sum_{j=1}^{T} m_i W_{i,j} n_j \quad (1)$$

Weighted connections are used to connect the units in different layers, $W_{i,j}$ is the denotation for these connections, where i and j refer to the i^{th} unit of the hidden layer and the j^{th} unit of the visible layer, respectively. The bias weights (offsets) for the visible and the hidden layers are represented by a_i and b_j, respectively.

The RBM probability distribution can be calculated via the energy function $P(m,n)$ given in Eq. 2 [19] below.

$$P(m,n) = \frac{1}{\tau} e^{-E(m,n)} \quad (2)$$

Where the normalizing constant is represented by the variableτ, which is the summation of the $e^{-E(m,n)}$for

[5] https://archive.ics.uci.edu/ml/datasets/CMU+Face+Images

every applied configuration. The Marginal distribution probability shown in Eq. 3 [19] below is calculated for the visible layer – input layer as the sum over all hidden layer configurations.

$$P(m,n) = \frac{1}{\tau}\sum_{n=1}^{T} e^{-E(m,n)} \qquad (3)$$

Thus, the conditional probability of any configuration for the visible unit m can be calculated for a specific configuration of the hidden unit n by Eq. 4 [19].

$$P(m\backslash n) = \prod_{i=1}^{N} P(m_i/n) \qquad (4)$$

Contrariwise, the conditional probability of any configuration for the hidden unit n can be calculated for a specific configuration of the visible unit m, by Eq. 5 [19].

$$P(n\backslash m) = \prod_{j=1}^{T} P(n_j/m) \qquad (5)$$

Finally, the individual activation probabilities are shown by Eq. 6 and Eq. 7 [19]. The Logistic sigmoid function is represented by σ.

$$P(n_j = 1\ /m) = \sigma\ (\ b_j + \sum_{i=1}^{N} W_{i,j} m_i\ ) \qquad (6)$$

$$P(m_i = 1\ /n) = \sigma\ (\ a_i + \sum_{j=1}^{T} W_{i,j} n_j\ ) \qquad (7)$$

The updating of the weights in this article is based on [20]. The constructive divergence (CD) algorithm was used to train the RBM and optimize the weights vector.

Considering recent works, Alphonse et al. [34], proposed a Multi-Scale and Rotation-Invariant Phase Pattern (MRIPP) method for facial expression recognition. In this method, robust, rotation-invariant, and blur-intensive features are brought out. The MRIPP benefits from a series of Restricted Boltzmann Machines (RBM). A stack of RBM is applied given the fact that the convergence in DBM is slow. The batch of RBM provides a notable accuracy of classification.

2.3. Deep Belief Networks

This network is a type of unsupervised learning network; it combines the interactions that occur between the variables that comprise the hidden layers or the visible layer (input). Figure 3 illustrates the Deep Belief Networks (DBN) structure.

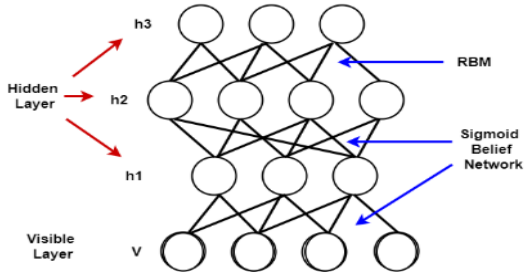

Figure 3. Deep Belief Network structure [21].

The indirect connections between the hidden layers $h^{(1)}$ - $h^{(2)}$, $h^{(2)}$ - $h^{(3)}$ and layers v - $h^{(1)}$ are illustrated in Figure 3 above. The top layers (i.e. $h^{(2)}$ and $h^{(3)}$) comprise indirect connected RBM, while others comprise a direct connected Bayesian network.

By adding a layer to the last layer, the priors can be improved, training of the DBN implement this idea.

The training of the DBN criteria is [22]:

a) First, the weights of the connections between one layer of both visible X and hidden $h^{(1)}$ layers are learned.
b) After, a new hidden layer ($h^{(2)}$) is added, and the new weights between $h^{(1)}$ and $h^{(2)}$ are learned. The previous step weights initialize between the layers $h^{(1)}$ and X in the new step.
c) Next, another hidden layer is attached and the model is trained based on the previous step parameters and weights which initialize weights between $h^{(2)}$, $h^{(1)}$ and X to obtain the new weights between $h^{(2)}$ and $h^{(3)}$.

d) The Up-Down algorithm is applied by the end for fine-tuning [21].

Li et al. [35] presented a parallel sparse Deep Belief Network method based on multi-objective optimization. This technique accelerates the speed of Sparse DBN networks and reduces the inference time. To have a multi-objective optimization, the Sparse DBN applied the Self-adaptive Quantum Multi-objective Evolutionary Algorithm based on Decomposition (SA-QMOEA/D). One of the advantages of this method is that it is not sensitive to the dimension increasing of the dataset compared to other networks. The experimental tests show that Sparse DBN inference time is less than conventional DBN. The remarkable speedup is in performance along with superior accuracy compared to traditional facial expression recognition.

2.4. Convolutional Neural Network

The convolutional, pooling, and fully connected layers are the basic components of a Convolutional Neural Networks (CNN) structure. Most model parameters are handled in the fully connected and can be reduced by using the operation of max pooling. This operation reduces the feature map via a window filter, which takes the highest value that the window covers to represent the whole values there. This improves the speed performance of the hardware systems used.

The Convolution stage and the sampling stage are the main two stages of CNN. These tasks are performed by the convolutional layers and the max-pooling layers, respectively. Rectangular inputs - known as the local receptive field - with dimensions of m x n are fed to each node of the convolutional layer from the previous layer through the convolution process.

The network parameters could be considered as a trainable filter or kernel function *F*, this is due to the local receptive fields that have the same biases and weights. The feature maps is the name used to describe these trainable filters. Figure 4 describes this process.

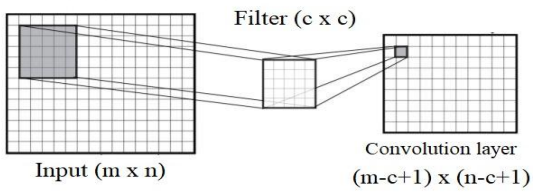

Figure 4. Process of Convolution [23].

The step that occurs between the convolution and pooling layers is called sampling. The pooling layer performs the subsampling by fetching small regions of rectangular shape from the convolution layer. The subsampling process is illustrated in Figure 5 below.

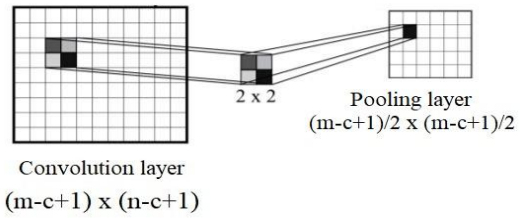

Figure 5. Subsampling process [23].

The CNN nodes are driven by their activation functions that may vary between a node and another in the same model. Some of those activation functions are tan h, sigmoid, and ReLU. The Adam function may be used as one of the optimizing for model training.

Some of the neurons can be dropped to not take part in the feedforward and backpropagation passes, this is known as the dropout regulation, support the effectiveness of the training process, and decrease hugely the overfitting probability of the model. The dropout neurons are selected randomly and from our model, it is chosen to be 0.5. The input data will be the only influencer on the remaining neuron's parameters update.

In a novel work, Priyasad et al. [38] presented a model to automatically recognize emotion by applying Deep Convolution Neural Networks (DCNN) and transfer learning. This study shows prominent accuracy of around 97% with cross-validation on two RML and eNTERFACE05 datasets. Comparing to the- state-of-the-art works the model reveals impressive accuracy differences mostly on the RML dataset.

2.5. Multi-inception CNN

It is a new technique used to classify images. Figure 6 below shows an example in which multilayers of inceptions construct each CNN.

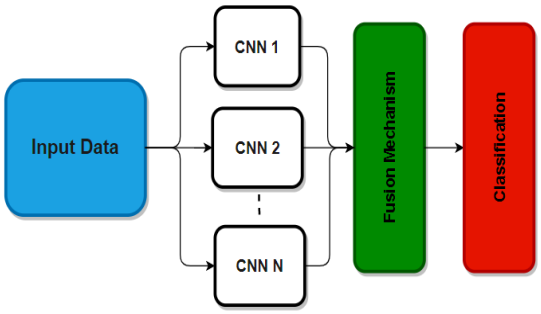

Figure 6. Multi-inception CNN

These nodes and inception blocks are generated randomly for each CNN. In this paper, an ensemble of a multi-inception model comprised of five individual Inceptions was generated. Each structure consists of an image input layer, randomized hidden inception blocks, a fully connected layer, a dropout layer, and a softmax layer. Each randomly generated inception block consists of four convolution layers of size 5x5, and 3x3 with a random number of filters, a ReLU layer, and a max-pooling layer, as shown in Figure 7.

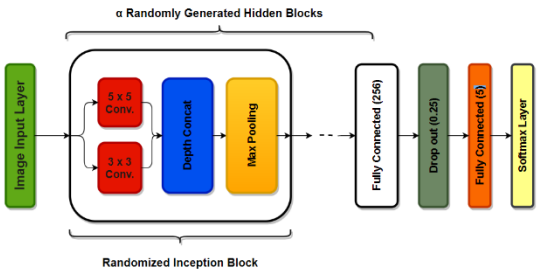

Figure7. Proposed model.

A parallel training process takes place once the CNNs are generated. Each input data is fed into the CNN's and returned to all neurons of the fully connected layer the labels of classification and the activation vector. Finally, these results are fused by the Pattern-Net [24-26].

3. Experimental Results

For evaluating the classification performance of the emotions, the previously discussed models are implemented. Models are trained with the CMU face dataset [27]. For the RBM model, a structure of a single visible layer and a hidden layer is used. The DBN structure is quite the same as the RBM with considering to double the number of hidden. To prepare the dataset and make it reliable for the CNN and Multi-inception CNN models, images were resized (dimensions unified), randomly rotated, and translated.

Input layer to CNN model is fixed to be 120 x 128 x 1, three convolution layers are applied with (3 x 3) filter, and ReLU layers, and the quantity of filter in every convolution layer is 64, 128 and 128. Three max-pooling layers (one after each convolution layer) with a size of (2x2) are also used. A fully connected layer with 256 nodes exhibited after the last max-pooling layer. A dropout layer is then added with a probability of 0.25. Then an additional fully connected layer is attached with four nodes (to classify the four emotions). Ultimately, connect to the softmax and classification layer.

The Multi-inception CNN model was tested using an augmented dataset with image resolution the same as the one used for CNN above. The results were compiled using the 5-fold cross-validation procedure. The overall accuracy for the four-class classification task of facial emotions is 78%, which was achieved by using a prediction model consisting of 5 randomly generated sub CNNs.Figures 8 to 11 illustrate the accuracy of the classification for the emotions individually and summarize the accuracy of those models.
The individual emotion classification is illustrated through diagonal percentages, while the most right down percentage represents the model classification accuracy.

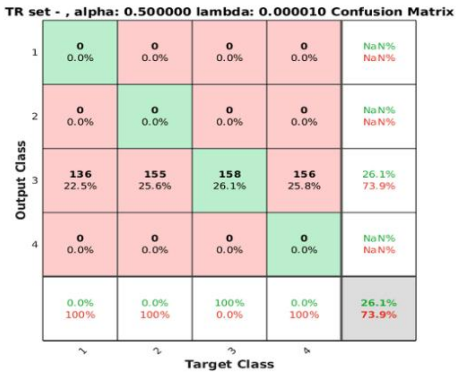

Figure 8. Classification accuracy of RBM.

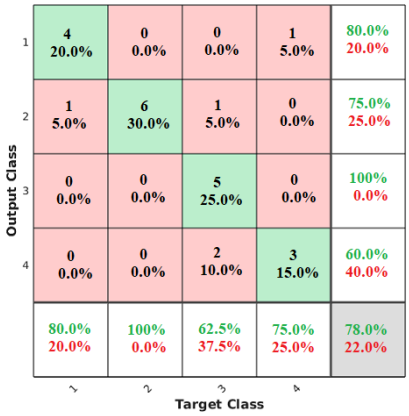

Figure 11. Classification accuracy of Multi-inception CNN.

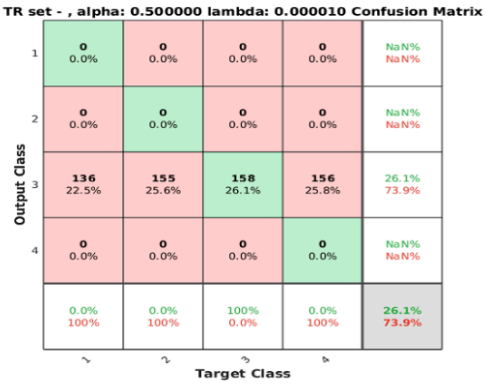

Figure 9. Classification accuracy of DBN.

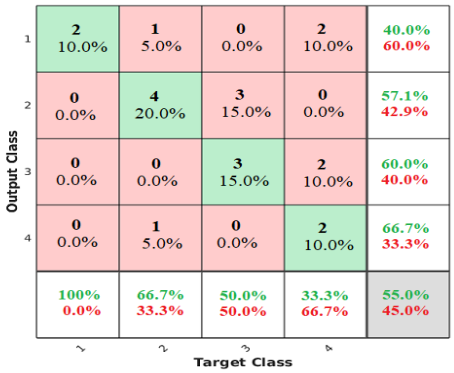

Figure 10. Classification accuracy of CNN.

Table 1 summarizes the classification accuracies of the model. As we can see, the RBM and DBM present the lowest accuracy while the Multi-inception CNN model has as high as 78% accuracy.

Those results show the superiority of our proposed multi-inception CNN model in handling such a small database task with accomplishing promising classification accuracy results compared to the same training and testing conditions considered for the regular CNN and other mentioned classifiers. This stimulates the expectations of achieving such improvements with other tasks using this model.

Table 1: Classification accuracy of the examined models

Model	Classification Accuracy %
RBM	26.1
DBN	26.1
CNN	55
Multi-inception CNN	78

4. Conclusion and future work

In this work, we applied multiple deep learning algorithms to classify facial emotions. The CMU facial dataset is used to evaluate the performance of models. The accuracy of models such as Restricted Boltzmann machine (RBM), Deep Belief Networks

(DBN), Convolutional Neural Networks (CNN), and multi-inception ensemble Convolution Neural Networks are evaluated. One of the challenges imposed on our study was the dataset size. Our experiments show that multi-inception CNN overcome the issue of the small dataset and it shows an accuracy of 87%. The RBM and DBM show extremely low as 26% accuracy. Among the models, CNN was the moderate one and it shows an accuracy of 55. The high ranked accuracy of multi inception CNN confirms that parallel training process solution is the key factor to maintain the precision of a model. According to this result, we have a plan for future work to examine the combination of some classification algorithms and present ensemble models.

Conflict of interest

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgments

No applicable.

References

[1] Z. Salek, F. M. Madani and R. Azmi (2013), Intrusion detection using neural networks trained by differential evaluation algorithm,10th International ISC Conference on Information Security and Cryptology (ISCISC), Yazd, pp. 1-6. DOI: 10.1109/ISCISC.2013.6767341

[2] Altaher, A. S., & Taha, S. M. R. (2017). Personal authentication based on finger knuckle print using quantum computing. International Journal of Biometrics, 9(2), 129-142.

[3] Ali, A. M., Zhuang, H., & Ibrahim, A. K. (2017). An approach for facial expression classification. International Journal of Biometrics, 9(2), 96-112.

[4] Ali, A. M., Zhuang, H., & Ibrahim, A. K. (2020). Multi-pose facial expression recognition using rectangular HOG feature extractor and label-consistent KSVD classifier. International Journal of Biometrics, 12(2), 147-162.

[5] Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. Journal Of Personality And Social Psychology, 17(2), 124-129. DOI: 10.1037/h0030377

[6] Jack, R., Garrod, O., Yu, H., Caldara, R., & Schyns, P. (2012). Facial expressions of emotion are not culturally universal. Proceedings Of The National Academy Of Sciences, 109(19), 7241-7244. DOI: 10.1073/pnas.1200155109

[7] Das, D., & Chakrabarty, A. (2016). Emotion recognition from face dataset using deep neural nets. 2016 International Symposium On Innovations In Intelligent Systems And Applications (INISTA). DOI: 10.1109/inista.2016.7571861

[8] Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE Transactions On Affective Computing, 1-1. DOI: 10.1109/taffc.2020.2981446

[9] Ko, B. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors, 18(2), 401. DOI: 10.3390/s18020401

[10] Wu, Y., & Qiu, W. (2017). Facial expression recognition based on improved deep belief networks. AIP Conference Proceedings 1864, 020130 (2017). Retrieved from https://doi.org/10.1063/1.4992947

[11] Al Ani, L. A., & Al Tahir, H. S. (2020). Classification Performance of TM Satellite Images. Al-Nahrain Journal of Science, 23(1), 62-68.

[12] Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis, T., Karpouzis, K., & Kollias, S. (2005). Emotion recognition through facial expression analysis based on a neuro-fuzzy network. Neural Networks, 18(4), 423-435. DOI: 10.1016/j.neunet.2005.03.004

[13] Kenji, M. (1991). Recognition of facial expression from optical flow. IEICE TRANSACTIONS On Information And Systems, E74-D(10), 3474-3483.

[14] Y. Yacoob and L. Davis(1994), Computing Spatio-temporal representations of human faces .Computer Vision and Pattern Recognition. Proceedings CVPR'94., 1994 IEEE Computer Society Conference On. IEEE, pp. 70–75.

[15] M. J. Black and Y. Yacoob, (1995), Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. Proceedings of IEEE International Conference on Computer Vision, Cambridge, MA, USA, pp. 374-381. DOI: 10.1109/ICCV.1995.466915

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan (2004), Analysis of emotion recognition using facial expressions, speech and multimodal information . Proceedings of the 6th international conference on Multimodal interfaces. ACM, pp. 205–211.

[17] Minaee, S., & Abdolrashid, A. (2019). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. Retrieved from http://arXiv:1902.01019v1 [cs.CV] 4 Feb 2019

[18] P. Liu, S. Han, Z. Meng, and Y. Tong,(2014), Facial Expression Recognition via a Boosted Deep Belief Network. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1805-1812. DOI: 10.1109/CVPR.2014.233

[19] Restricted Boltzmann machine. (2020). Retrieved 25 June 2020, from https://en.wikipedia.org/wiki/Restricted Boltzmann machine

[20] Y. Bengio. (2009). Learning deep architectures for AI. Foundations and trends® in Machine Learning, vol. 2, no. 1, pp. 1–127

[21] G. E. Hinton, S. Osindero, and Y.-W. (2006). A fast learning algorithm for deep belief nets. Journal of Neural computation, vol. 18, no. 7, pp. 1527–1554

[22] H. Larochelle, "Neural networks [7.7]: Deep learning - deep belief network." [Online]. Available: https://www.youtube.com/watch?v=vkb6AWYXZ5I

[23] Wang, C., & Xi, Y. (). Convolutional Neural Network for Image Classification. Johns Hopkins University Baltimore, MD, 21218.

[24] Li, H., Ellis, J. G., Zhang, L., & Chang, S. F. (2018, June). Pattern net: Visual pattern mining with a deep neural network. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (pp. 291-299).

[25] Khalid, M., Wu, J., Ali, T. M., Ameen, T., Altaher, A. S., Moustafa, A. A., ... & Xiong, R. (2020). Cortico-Hippocampal Computational Modeling Using Quantum-Inspired Neural Network. Frontiers in Computational Neuroscience, 14, 80.

[26] Abidalkareem, A. J., Abd, M. A., Ibrahim, A. K., Zhuang, H., Altaher, A. S., & Ali, A. M. (2020, July). Diabetic Retinopathy (DR) Severity Level Classification Using Multimodel Convolutional Neural Networks. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1404-1407). IEEE.

[27] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza (2012) .Disentangling factors of variation for facial expression recognition. Springer ECCV, pages 808–822.

[28] Pan, Z., Luo, Z., Yang, J., and Li, H., 2020. Multi-modal Attention for Speech Emotion Recognition. arXiv preprint arXiv:2009.04107,.

[29] Chen, S., & Jin, Q. (2016, October). Multi-modal conditional attention fusion for dimensional emotion prediction. In Proceedings of the 24th ACM international conference on Multimedia (pp. 571-575).

[30] [30] Beard, R., Das, R., Ng, R. W., Gopalakrishnan, P. K., Eerens, L., Swietojanski, P., & Miksik, O. (2018, October). Multi-modal sequence fusion via recursive attention for emotion recognition. In Proceedings of the 22nd Conference on Computational Natural Language Learning (pp. 251-259).

[31] Zhao, S., Ding, G., Gao, Y., & Han, J. (2017, October). Learning visual emotion distributions via multi-modal features fusion. In Proceedings of the 25th ACM international conference on Multimedia (pp. 369-377).

[32] Liang, J., Chen, S., & Jin, Q. (2019, November). Semi-supervised Multimodal Emotion Recognition with Improved Wasserstein GANs. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 695-703). IEEE.

[33] Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S. (2020). Multimodal Emotion Recognition with Transformer-Based Self Supervised Feature Fusion. IEEE Access.

[34] Alphonse, A. S., Shankar, K., Rakkini, M. J., Ananthakrishnan, S., Athisayamani, S., Singh, A. R., & Gobi, R. (2020). A multi-scale and rotation-invariant phase pattern (MRIPP) and a stack of restricted Boltzmann machine (RBM) with preprocessing for facial expression classification. Journal of Ambient Intelligence and Humanized Computing, 1-17.

[35] Li, Y., Fang, S., Bai, X., Jiao, L., & Marturi, N. (2020). Parallel Design of Sparse Deep Belief Network with Multi-objective Optimization. Information Sciences.

[36] Trujillo, L., Olague, G., Hammoud, R., & Hernandez, B. (2005, September). Automatic feature localization in thermal images for facial expression recognition. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops (pp. 14-14). IEEE.

[37] Ilbeygi, M., & Shah-Hosseini, H. (2012). A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. Engineering Applications of Artificial Intelligence, 25(1), 130-146.

[38] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2019). Learning salient features for multimodal emotion recognition with recurrent neural networks and attention-based fusion. In 15th International Conference on Auditory-Visual Speech Processing (AVSP).

[39] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, 115, 101-106.

[40] Ekundayo, O., & Viriri, S. (2019, March). Facial expression recognition: a review of methods, performances, and limitations. In 2019 Conference on Information Communications Technology and Society (ICTAS) (pp. 1-6). IEEE.