2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)

Human Abnormality Classification Using Combined CNN-RNN Approach

Md. Mohsin Kabir
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
m97kabir@gmail.com

Farisa Benta Safir
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
farisabentasafir@gmail.com

Saifullah Shahen
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
shahriar350@gmail.com

Jannatul Maua
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
jannatulmaua035@gmail.com

Iffat Ara Binte Awlad
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
iffatara7878@gmail.com

M. F. Mridha
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh
firoz@bubt.edu.bd

***Abstract*—With the advent era of big data, Facial Expression Recognition (FER) becomes a promising area for Deep Learning. The facial expression reflects our mental activities and provides useful information on human behaviors. With the incrementing progress of the deep learning-based classification method special demands for human stability measurement using facial expression have emerged. Recognizing human abnormalities such as drug addicts, autistic, criminals is very challenging due to the limitation of existed FER systems. Besides, there are no datasets that consist of helpful images that describe the true expressions of the human face that can detect human abnormality. To evaluate the best performance on human abnormality recognition we have created a Normal and Abnormal Humans Facial Expression (NAHFE) dataset. In this paper, we proposed a new model by stacking the Convolutional Neural Network and Recurrent Neural Network (RNN) together. The proposed combined method consists of convolution layers followed by the recurrent network. The associated model extracts the features within facial portions of the images and the recurrent network considered the temporal dependencies which exist in the images. The proposed combined architecture is evaluated based on the mentioned NAHFE dataset and got the superior performance to detect human abnormalities.**

Keywords—Computer vision, Deep learning, Convolutional neural networks, Recurrent neural networks, Facial expression recognition.

I. INTRODUCTION

The human face is exceptionally imaginative, able to notify innumerable emotions without accents a word. Facial Expression Recognition (FER) is a fundamental problem in the computer vision, and image processing field [1, 2]. Facial expressions differ from man to man, and also influenced by gender, age, ethnicity, and so on. As a whole, the facial expression constitutes the real feelings of the human being. So FER becomes more pledging because of its human behavior analytical power. Hence, using this analytic capability of FER, we have developed a combined method of CNN and RNN to classify human abnormalities. This approach analyzed the human face and finds the abnormalities, such as Drug addiction, Autism, Criminalism.

The analysis of facial expressions is a laborious task for Deep Learning (DL) approaches, because the human being can deflect the way they express their expressions significantly [3, 4]. The convolutional neural network that is a part of deep learning, mostly used for image analysis, and image processing tasks [5, 6]. Deep CNNs have gained momentous success and proven specifically well suited for image recognition tasks from massive datasets[7, 8]. Recent CNN architectures employ several ways to shorten the training time and enhance generalization over input data, including data augmentation [9], dropout regularization [10], ReLU activation functions, and GPU acceleration [7]. RNN has been getting popular exponentially because RNN not only assesses its input(s) momentarily but also its evaluation lies on the past input(s) [11, 12]. Thus, the result is generated from a composition of information coming from the past and present. Hence, in this paper, we have developed an architecture that uses both CNN and RNN to classify human abnormalities. The overall contributions of this work are given below:

- Identifying present difficulty to analyze human abnormalities in Facial Expression Recognition (FER) problems.
- A new dataset named Normal and Abnormal Humans Facial Expression (NAHFE) dataset is created that consists of 1936 images of 4 different classes.
- A novel CNN-RNN combined approach to classifying human abnormalities is proposed and got a convincing result. This is a unique approach that is believed to have enormous potentials. Also, the result of the proposed CNN-RNN combined architecture is compared to the result of basic CNN architecture.

The rest of this paper is constructed as follows: The related research is described in section 2. In section 3, the overall architecture of deep CNN-RNN combined

architecture is described. Section 4 contains the model's evaluation and compares the results of the architectures. At last, Section 5 concludes the paper.

II. Related work

Facial expression recognition or analysis of facial effect has inducted significant attention in the computer vision researchers during the past few years [13, 14]. Popular approaches classify 7 basic expressions namely happy, sad, surprised, disgusted, fear, angry, and neutral but no prior work on classifying human abnormalities has been done yet.

Due to different facial appearances, micro-expression classification is a challenging task. Takalkar et al. [17] analyzed the use of deep learning for micro-expression classification. Using data augmentation, authors generated massive datasets of synthetic images named CASME and CASME 2 databases. Finally, A CNN-based micro-expression recognizer is developed by combining and tuning both of these datasets which gave a maximum of 78.2% accuracy. Jung et al. [16] proposed two deep network models using CNN and DNN for the FER problem. By using FER 2013 database they achieved 72.78% accuracy for DNN architectures and 86.45 for CNN. The authors first detected face from input images by Haar-Like features and then applied the Deep Learning model. Xiujie Qu et al. [15] proposed a real-time and fast face recognition model using CNN. The authors divide the process into two parts. First, the network is trained on the PC and then the network is implemented on the Field Programmable Gate Array (FPGA) and got 99.25% which is state of the art. Neha Jain et al. [19] proposed the face emotion recognition model which is a combination of deep CNN and RNN model. In this model author used two datasets one is MMI Facial Expression Database (TFD) and another is the Japanese Female Facial Expression (JAFFE). In this model, 80% of datasets are used for training and 20% of datasets are used for validation. Achieve 94.91% accuracy when used JAFFE datasets and achieve 92.07% accuracy when used MMI datasets. Fathallah et al. [20] experimented that CNN is very effective while recognizing facial expressions. Authors fine-tuned CNN architecture with the VGG model and trained famous datasets like CK+, MUG, and got nearly 99% accuracy which is state of the art. Ali et al. [18] proposed deep neural network architecture that has been presented for automated facial expression which has two convolutional layers one is max pooling another is four inception layers and firstly applies the inception layer. The proposed approach takes a facial image as input and classifies that image into 6-expression or neutral. The author used different databases such as MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013 and achieve 94.7% which is the best accuracy using the CMU MultiPIE database.

However, all these FER techniques only analyzed the six or seven basic expressions but do not give any solution to human stability identification. In this paper, we have discussed how the CNN-RNN combined approach solves this issue.

III. Methodology

The combined CNN and RNN architecture is implemented and benchmarked to evaluate human abnormalities classification from image datasets. In the following sub-sections, the network structure of the proposed models is briefed.

A. Data Pre-processing

The data pre-processing has occurred in two stages, data normalization and data augmentation. These two techniques are described below.

Data Normalization: Normalizing image is a significant pre-processing methodology. It reduces the inner-class feature discrepancy and is viewed as intensity offsets. The intensity offsets are fixed in the local region, So standard deviation and gaussian normalization are useful while normalizing. The resulted image after normalization is computed by (1) [19].

$$\Psi(\pi,\theta) = \frac{\xi(\pi,\theta) - \mu(\pi,\theta)}{6\sigma(\pi,\theta)} \quad (1)$$

Where μ is a local mean and σ is a local standard deviation [12].

$$\mu(\pi,\theta) = \frac{1}{M^2}\sum_{k=-\alpha}^{\alpha}\sum_{n=-\alpha}^{\alpha}\xi(K+\mu, n+\theta)$$

$$\mu(\pi,\theta) = \sqrt{\frac{1}{M^2}\sum_{k=-\alpha}^{\alpha}\sum_{n=-\alpha}^{\alpha}[\xi(K+\mu, n+\theta) - \mu(\pi,\theta)]^2}$$

Data Augmentation: We carry out various transformations to the training and testing samples during evaluation to increase the network resistance to mutation in input samples. This image mutation technique is performed on the CPU simultaneously with network evaluation and training on GPU. Deep learning architectures constantly demands a vast number of input samples to achieve better accuracy. Even though our NAHFE dataset has 1936 images for 4 classes, it is still inadequate for evaluating a deep learning architecture. So before evaluating the architecture, we augmented the dataset with several transformation techniques for propagating diverse tiny variations in appearances and poses. We engaged five image appearance filters (Gaussian, disk, unsharp, average, and motion), and six affines transform matrices by joining short geometric transformations to the identity matrix. By generating this augmentation, for every actual image in the dataset, we have created (5x6) = 30 different images, therefore the number of samples increased to (1936x30) = 58,080. Then we normalized all the images using the above-mentioned approaches. Finally, the dataset is divided into two portions train, and validation to experiment and evaluate the model. 80% of the dataset is used to train the model and 20% is used for validation.

B. Convolutional Neural Network

Facial expression images appear in several shapes and qualities, so we define the data pre-processing technique that can work with any type of input shape and quality. In this architecture, CNN constructs with six convolutional layers

and two dense layers, each with a ReLu activation function, and dropout for training. Equations (2) and (3) Explain the convolution and fully connected layers operations. Moreover, we applied regularization for every weight matrix that shortens the volume of the weights at the separate layer to several fixed hyperparameters. Equation (3) Explains the regularization process.

$$Y_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{m_i^{(l-1)}} K_{i,j}^{(l)}\ Y_j^{(l-1)} \tag{2}$$

Where the output Y_i^l of layer l consists of the m_3^l feature of size $m_1^l \times m_2^l$. The ith feature map denoted Y_i^l and $B_i^{(l)}$ is a bias matrix and $K_{i,j}^{(l)}$ is the filter of size.

$$d(x) = Activation(w^T x + b) \tag{3}$$

Here, $w = [w_1, w_2, \dots, w_n]^T$ represents the weight vector of the dense layer, and b represents the bias value of the dense layer.

$$ReLU = \max(0, x) \tag{4}$$

If the ReLu function gets any non-positive value, it returns zero, but for any positive input of x, it returns that input value.

$$Dropout(x, p) = \begin{cases} x, & with\ prob.\ p \\ x, & with\ prob.\ 1-p \end{cases} \tag{5}$$

Where x be the output of a particular neuron in the network and p the dropout possibility. Fig. 1. Demonstrate the CNN approaches for learning and extracting features from input images.

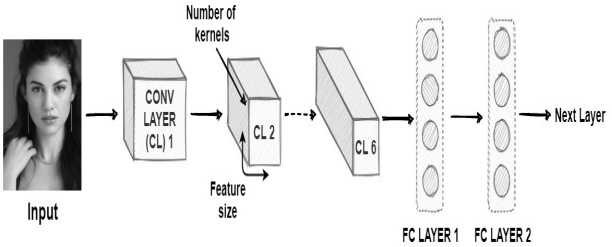

Fig. 1. The convolutional neural network architecture is the feature learning and extracting portions of the overall CNN-RNN combined architecture. Each of the cubes represents an output of the convolution. The height and width are the gained information, and each cube's depth is equal to the number of kernels. Each convolution is followed by batch normalization and an activation layer. After the final convolution, it is converted into a linear set of nodes.

C. Recurrent Neural Network

Recurrent neural networks (RNNs) are a branch of neural networks that are typically used to processing time-series and other sequential data. In RNN the tensors transit both forward and backward by circulating loops in the network. It generates an output at each time step and has recursive connections between hidden units. The mathematical model of RNNs can be expressed as follows:

$$h_t = \sigma_h(W_h x_t + U_h y_{t-1} + b_h),$$
$$y_t = \sigma_y(W_y h_t + b_y) \tag{6}$$

Where x_t is an input vector, h_t is a hidden layer vector, y_t is the output vector, W_h, U_h and b_h, b_y are weighting matrices and vectors, and σ_h and σ_y are activating vector functions.

D. Combined CNN-RNN Architecture

The proposed architecture combines continuous data using RNN to expand and learn the information. To regulate all the parameters, the CNN feature extraction method is used. The RNN classifies the images by adding the extracted features from the successive CNN network of each image and finally for the prediction uses Softmax. While experimenting, when the image is served to the CNN network, from the dense layer, 200-dimensional vectors will be uprooted. For the learned time t, the network takes P frames from the past ([t − P, t]). Thereafter every frame runs from time t − P to t to the CNN and extracts P vectors for every input. After that, each vector passes by a node of RNN and each node of that model gives some outputs of the valence label. The experiment and evaluation of the architecture are done by various layers of CNN as input features and the proposed one has acquired the maximum score on test data. To calculate the cost function, the mean squared error is used while optimizing. The overall architecture is illustrated in Fig. 2.

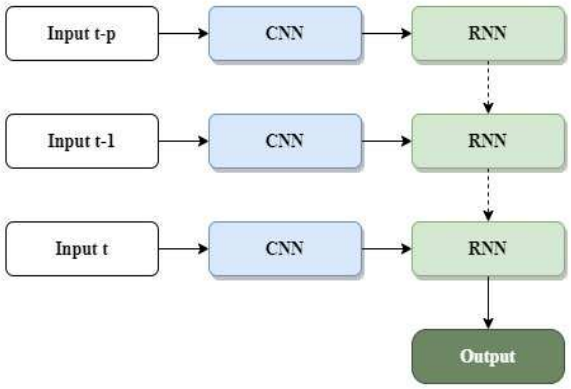

Fig. 2. The combined CNN-RNN architecture to classify human abnormalities using the NAHFE dataset. CNN extracts the information and the RNN classifies the images. The input images first pass to CNN and then the output from the dense layer of CNN goes through the RNN and classifies the exact class of the image.

IV. Evaluation

We conduct experimental research on evaluating the proposed human abnormalities classification problems solution on our NAFHE dataset. This study carried out the impact of the combined CNN-RNN approach to classify human abnormalities. Firstly, we demonstrate the dataset used for the experiments. Secondly, the empirical setup is explained. Thirdly, the metrics used to measure the system accuracy are described. Fourthly, the results of the combined CNN-RNN approach are presented. Finally, the results using basic CNN architecture is shown and compared with the proposed model.

A. Dataset

We observe that most of the Facial Expression Recognition (FER) image datasets on the Web are built for classifying six or seven basic expressions of the human face. But to analyze human stability, we need a dataset that is divided into four classes named Drag addiction, Autism, Criminalism, and Normal. Therefore we have used web gathering approaches to get Normal and Abnormal human

images from the Web and create our Normal and Abnormal Humans Facial Expression (NAHFE). As we have narrated, we add four classes to our dataset: Drag addiction, Autism, Criminalism, and Normal. Using respecting keywords to each of the four classes in addition to the name of the class (e.g., sinful, convicted, sinner for Criminal), we have gathered a good number of images that belong to the same class. Finally, we placed 1936 images for the four classes. In evaluations, we have used 80% (1548) of the images for training and the rest 20% (388) for testing. Also, images of four classes are distributed evenly and the number of samples in each class is 484. A couple of sample images from the NAHFE dataset are shown in Fig. 3.

Fig. 3. The sample images of Normal, Autistic, Drug addict, and Criminal human faces from the dataset we used in this research from top to the right. These images are gathered from the web using the web gathering technique.

B. Evaluation Matrices

To evaluate the performance of the architecture, we used accuracy, precision, recall based on the confusion matrix. Accuracy, precision, and recall can be calculated using the following formulas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

Where TP means True Positive, TN means True Negative, FP means False Positive, and FN means False Negative.

C. Experimental Setup

Python programming language is used for pre-processing data, experimenting, and evaluations of the model. The proposed architecture is implemented using TensorFlow and Keras. Besides, NumPy is used for mathematical operations on the architecture.

D. Experiments and Comparisons

We use the NAHFE dataset to implement our combined CNN-RNN architecture. Table I shows the prediction accuracy, precision, and recall of the combined CNN-RNN approach for classifying human abnormalities. This table demonstrates the performance of the proposed architecture and compared the predicted result of the proposed architecture with basic CNN architecture. Our investigation found that using only CNN architecture does not classify the human abnormalities properly. Basic CNN architecture gives only 0.732 accuracies while CNN-RNN combined approach gives 0.895. It is noted that the performance measurement by precision and recall also higher for CNN-RNN combined approach comparing to the basic CNN architecture. So, to get better results combining CNN with RNN was found more effective. Every result in this research is given as a mean of four runs. The proposed model is trained with a limit of 100 epochs.

TABLE I. THE TABLE SHOWS THE VALIDATION ACCURACY, PRECISION, AND RECALL OF THE PROPOSED CNN-RNN COMBINED APPROACH AND BASIC CNN ARCHITECTURE.

Model	Accuracy	Precision	Recall
CNN	73.20	72.97	70.38
Proposed Model	**89.50**	**87.98**	**88.13**

Besides, the proposed CNN-RNN combined architecture evaluated under different hyperparameters to significantly tuning the model and investigated the improvement of the model in different circumstances. The final result of the proposed CNN-RNN combined approach comes out from investigations of using the different number of hidden units and hidden layers.

TABLE II. THE TABLE SHOWS THE VALIDATION ACCURACY, PRECISION, AND RECALL OF THE PROPOSED CNN-RNN COMBINED APPROACH BASED ON DIFFERENT HIDDEN UNITS.

Number of Hidden Units	Accuracy	Precision	Recall
50	85.24	84.97	87.08
100	86.54	86.67	87.20
150	**88.75**	**87.03**	**87.98**
200	87.70	86.60	87.10

Table II presented the result of the proposed model using a different number of hidden units. The investigation says that the model performs the best result while using 150 hidden units. The accuracy, precision, and recall scores increase with the increasing number of hidden units. But when the number

of hidden units crosses 150, the model start behaving negative result.

TABLE III. THE TABLE SHOWS THE VALIDATION ACCURACY, PRECISION, AND RECALL OF THE PROPOSED CNN-RNN COMBINED APPROACH BASED ON DIFFERENT HIDDEN LAYERS.

Number of Hidden Layers	Accuracy	Precision	Recall
3	87.03	86.89	86.08
4	87.64	87.06	86.97
5	88.15	87.21	87.10
6	**89.50**	**87.98**	**88.13**
7	89.37	86.14	88.01

Similarly, Table III shows the result of using different hidden layers in the model and it is found that using 6 hidden layers gives the best result of classification. Increasing the hidden layers to 7 decreases the result of the architecture. Finally, the experiment finds that the proposed classification model performs best while using 150 hidden units and 6 hidden layers.

This paper broadly investigated the significance of the CNN-RNN combined approach to classify human abnormalities and founds satisfactory performance with 150 hidden units and 6 hidden layers.

V. CONCLUSION & FUTURE SCOPE

This paper experiments and evaluates a human abnormalities classification method using the NAHFE dataset, created by us. We practiced a combined method of CNN and RNN to train and test our method precisely. We observe that the combined CNN-RNN approach gives better performance for human abnormalities classification. The architecture proposed in this paper, to the best of our knowledge, is the first architecture that classifies human abnormalities using a novel CNN-RNN combined approach. The performance of the proposed architecture is also compared with different CNN baseline architectures. We strongly believe that this research work will pave the way for significant research on human abnormalities classification and will enhance the intelligence and the practicability in future work.

ACKNOWLEDGMENT

The authors would like to thank the Advanced Machine Learning (AML) lab for their resource sharing and precious supports.

REFERENCES

[1] Lv, Y., Feng, Z. and Xu, C., 2014, November. Facial expression recognition via deep learning. In 2014 International Conference on Smart Computing (pp. 303-308). IEEE.

[2] Song, I., Kim, H.J. and Jeon, P.B., 2014, January. Deep learning for real-time robust facial expression recognition on a smartphone. In 2014 IEEE International Conference on Consumer Electronics (ICCE) (pp. 564-567). IEEE.

[3] Jung, H., Lee, S., Yim, J., Park, S. and Kim, J., 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE international conference on computer vision (pp. 2983-2991).

[4] Mollahosseini, A., Chan, D. and Mahoor, M.H., 2016, March. Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.

[5] Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y., 2003. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 16(5-6), pp.555-559.

[6] Mollahosseini, A., Chan, D. and Mahoor, M.H., 2016, March. Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.

[7] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), pp.84-90.

[8] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[9] Simard, P.Y., Steinkraus, D. and Platt, J.C., 2003, August. Best practices for convolutional neural networks applied to visual document analysis. In Icdar (Vol. 3, No. 2003).

[10] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), pp.1929-1958.

[11] Kobayashi, H. and Hara, F., 1993, October. Dynamic recognition of basic facial expressions by discrete-time recurrent neural network. In Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan) (Vol. 1, pp. 155-158). IEEE.

[12] Zhang, T., Zheng, W., Cui, Z., Zong, Y. and Li, Y., 2018. Spatial–temporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics, 49(3), pp.839-847.

[13] Fathallah, A., Abdi, L. and Douik, A., 2017, October. Facial expression recognition via deep learning. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) (pp. 745-750). IEEE.

[14] Majumder, A., Behera, L. and Subramanian, V.K., 2016. Automatic facial expression recognition system using deep network-based data fusion. IEEE transactions on cybernetics, 48(1), pp.103-114.

[15] X. Qu, T. Wei, C. Peng and P. Du, "A Fast Face Recognition System Based on Deep Learning," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 289-292, doi: 10.1109/ISCID.2018.00072.

[16] H. Jung et al., "Development of deep learning-based facial expression recognition system," 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), Mokpo, 2015, pp. 1-4, doi: 10.1109/FCV.2015.7103729.

[17] M. A. Takalkar and M. Xu, "Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets," 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, NSW, 2017, pp. 1-7, doi: 10.1109/DICTA.2017.8227443.

[18] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1-10, doi: 10.1109/WACV.2016.7477450.

[19] Neha Jain,Shishir Kumar,Amit Kumar,Pourya Shamsolmoali and Masoumeh Zareapoor, " Hybrid Deep Neural Networks for Face Emotion Recognition," Pattern Recognition Letters, doi: 10.1016/j.patrec.2018.04.010.

[20] Abir Fathallah, L. Abdi and A. Douik, "Facial Expression Recognition via Deep Learning," 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, 2017, pp. 745-750, doi: 10.1109/AICCSA.2017.124.