IHMCNN-LSTM: Improved Hybrid Model for Facial Emotions Detection

Arpita Nagpal
GD Goenka University, Gurgaon

Abstract In this manuscript authors have proposed an algorithm for real-time facial emotion detection used to identify human emotions. Facial expressions are the most significant modality to represent a human's emotions Face is a main part of the human body; it is used for non-verbal communication. In this paper, seven human emotions are captured of a single face as well as a group of faces. Emotion detection in a group of people is challenging due to obscuration of the hidden body pose variation, occlusion, variable lighting conditions, indoor-outdoor siting, and image quality. The group of emotions detection is useful in analyses from social media, marking, social event detection, public safety, human computing interaction, and many more. The proposed IHMCNN-LSTM (Improved Hybrid Model convolutional neural network- long short-term memory) Facial Emotions Detection consists of three cycles of detection: Face detection, facial features detection, and emotions detection. The emotions detection structure is based on the Deep Neural Network, It can detect seven different emotions – 'Happy, Anger, Disgust, Scared, Sad, surprised, and Neutral'. In the first step, features are extracted from static images using a convolutional neural network (CNN). The second step, determine the relationship between the transformation of facial features in image sequences and the seven basic emotions using Long Short-Term Memory (LSTM). LSTM is a type of recurrent neural network. The proposed algorithm is a hybrid algorithm named 'IHMCNN-LSTM' which combines the performance of the emotion recognition system by using transfer learning. The significant percentage of accuracy on the seven human emotions is improved using the proposed algorithm as compared with previous algorithms.

Keywords: Facial expression detection, Emotion detection, Convolutional Neural Networks, long short-term memory.

I. INTRODUCTION

Over the past three decades, facial emotion recognition has received increasing attention. Due to the time and requirements of advanced technology, many features have been added to the technology to make the communication system safer and more secure. It is often assumed that a mechanism based on human behavior develops behavioral systems. Face recognition provides information from a digital camera by analyzing images of human faces. It processes all facial feature structures, including the distance between left and right eyes, left and right eyes, mouth, nose, lips, and chin edges [1][2]. Human emotions play an important role in everyday life for nonverbal communication. They also express human emotions, enabling human expression [2,3]. We can readily discern human emotions. Facial expression information is used in an automated emotion recognition system [3]. This study aims to identify seven emotions using Facial expressions such as anger, disgust, fear, happiness, sadness, shock, and neutrality in certain faces and faces.

Emotion recognition contains a wealth of information including facial expressions, body language, tone and tone, and interpretation. Faces are important because they exchange enough information that can be used widely for different purposes in different industries. Furthermore, facial expressions can change similar information across cultures and countries. It is difficult for computers to classify facial features among people of different ages, genders, and ethnic groups under different lighting conditions, settings, and conditions [4]. The contribution of this paper is as follows:

- Proposed an algorithm named 'IHMCNN-LSTM' for facial emotions detection;
- The method of transfer learning is used to improve the model performance;
- A human emotions detection model is developed for an experiment.

Section 2 describes the related work introduced in the field of emotion detection. Section 3 explains the proposed IHMCNN-LSTM algorithm. Section 4 contains the details of evaluation parameters and datasets. In Section 5, the experimental observations and result analysis are presented. The conclusion of the work is there in section 6.

II. RELATED WORK

There are many tasks related to emotion recognition and facial expression detection and recognition, deep neural networks, and transfer learning.

In some studies, it is looks at still pictures and finds your facial features. However, the facial expression is generated by the contraction and relaxation of one's facial muscles. So it is best to view emotions as dynamic and stable factors, some studies may remove geometric and morphological factors basis. A study [5] manually selected the most salient images in the image sequence to experiment with them. This achieved high accuracy; however, it is not a reasonable approach to determine if the method is feasible. Various deep learning techniques have been used to manage complex tasks and improve efficiency, such as CNN and RNN [6], improve deep neural networks [7][8][9], and improve LSTM. CNN has great power to analyze images and Computer vision tasks such as classification, identification and recognition. Long et al. [10] and Chen et al.

[11], which has been rationalized classification using deep CNN; Yu and Zhang [12] used deep CNNs for facial feature recognition based on static images. A study investigated the effect of CNN depth on large image recognition [13]. So many studies have successfully used CNN in facial expression recognition for filtering features. To enhance HRI and robot–robot interaction capabilities, the study [14] proposed a CNN algorithm for robots to recognize emotions. The network consists of three convolution layers and one fully connected layer as are outcomes and the input is information from speech, gesture, and facial recognition. Although there are only four layers in the CNN system, it performs well in experiments and has an acceptable sense of recognition. The only drawback of the CNN framework is that the method can only be used on still images. This method cannot be used for images that are real-time and dynamic in nature. It is proposed that RNN can cope with dynamic data in a time sequence. RNNs are widely used in contexts of speech data as it has internal memory for processing the time-input sequence. A study [15] showed that RNN is a powerful model for sequential data and LSTM works well for sound recognition. Sunder Meyer et al. [16] showed that the LSTM network offered better performance than the standard RNN for one great English language and French language modeling task. Xu et al. proposed face ant spoofing using LSTM and CNN algorithms [17]. These LSTM and CNN algorithms can be trained on temporal features with a face ant spoofing database with diverse attacks [18]. This construction is useful and the experimental results suggest that it is working well for face ant spoofing. However, the face ant spoofing method can be trained and applied to only one data set. Paper [19], proposed human-robot interaction with the help of multimodal emotion recognition and evolutionary computation. There are several algorithms using human-robot interaction with a maximum detection rate was 97%. Emotion-based communication on a small humanoid robot is practical and its' applications are adopted in many human robot interaction applications. The methods proposed in this study are specific to the data set. Prior knowledge can also be transferred and reused for the next training step by the transfer learning technique.

III. PROPOSED MODEL COMBINING CNN AND LSTM

The methodology proposed in this paper aims to find out the relation between the image sequences of human facial expressions and their corresponding labels. The facial expressions possess both appearance and temporal features. As they are generated by the amalgamation of the contraction and relaxation of facial muscles. In the fig. 1 three phase cycles are illustrated in which three different stages of operation are used for evaluation and validation purpose. The step wise approach is discussed below in algorithm. The convolution layers in the front of the fully connected layers of the CNN are cut and connect to two LSTM layers with 256 cells. Therefore, the 128 × 128 pixel input image is scaled down to a 2048 × 1 feature vector after the convolutional part. The vector is fed to the LSTM layers as the input. Then, the number of parameters in the first LSTM layer is 2,360,320; this is dramatically lower than in the case of the LSTM network without a feature extractor.

IV. TRANSFERRING PARAMETERS CNN TO LSTM

To improve the model performance, use to transfer the knowledge of parameters from the source domain data. The source data cannot be directly used in the target model. In the first step to extract the information, CNN is trained with a large number of labelled static images. The second step is to flattened the image into a 1-D feature map by removing the last layer of the CNN. This 1-D feature map which is output of model trained by source domain data is important for conducting transfer learning. This 1-D feature map is then imported into LSTM as time sequence information. The final part of the CNN is extracted and transferred to LSTM for target field data training. In this way, the pre-trained CNN model with source field information can transfer the knowledge to the new LSTM model, improving the performance of the LSTM model In this study, the source field data are static data sequences, and target field data are dynamic data sequences. Through this transfer learning, we can transfer the knowledge of human facial emotion recognition of the static images to the new LSTM.

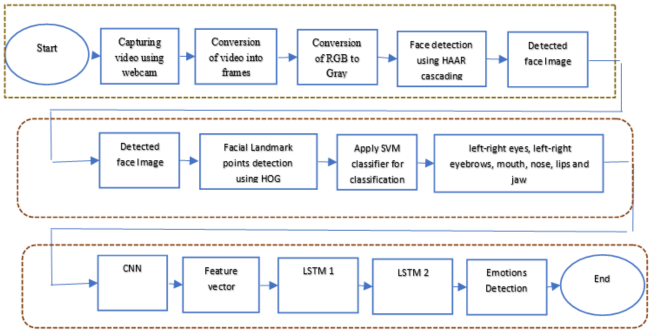
Start
Capturing video using webcam
Conversion of video into frames
Conversion of RGB to Gray
Face detection using HAAR cascading
Detected face Image
Detected face Image
Facial Landmark points detection using HOG
Apply SVM classifier for classification
left-right eyes, left-right eyebrows, mouth, nose, lips and jaw
CNN
Feature vector
LSTM 1
LSTM 2
Emotions Detection
End

Fig. 1. Illustration of Methodology

V. EXPERIMENTAL RESULTS

The extended Cohn-Canada (CK+) data contains 123 subjects and 593 image classes. The images are taken sequentially using the front-facing camera issue. All images are 640 × 480 grayscale images. Most of the subjects are Euro-American women. Of the 593 series of images, 327 are labeled with seven emotions: happy, sad, shocked, scared, disgusted, angry and contemptuous. All emotions vary in length, and must contain at least 10 images. We made some choices from the 232 series with more than ten images and six basic emotions. The last ten images of 232 sequences are taken in the experiments.

The input for face detection and emotion detection is taken from captured video. The performance of the proposed IHMCNN-LSTM algorithm is optimized by testing the performance parameters in each sampling cycle. The performance parameters for quantitative analysis in the proposed method are recall, precision, sensitivity, specificity, and error rate measurements.

TABLE I. RESULT FROM ANALYSIS FOR EMOTIONS

Emotions / Detection probability %	Happy	Neutral	Anger	Sad	Surprised	Disgust	Scared
Happy	97.8	3.24	1.06	1.32	6.23	0.09	9.12
Neutral	12.5	98.8	5.23	10.2	2.31	0.57	14.3
Anger	0.75	13	96.9	21.3	0.38	1.3	10.3
Sad	0.34	20.6	15.1	93.3	0.3	0.08	21
Surprised	1.04	3.23	3.37	0.09	94.4	5.76	13.5
Disgust	0.05	34.5	21.1	9.09	17.9	91.4	14.4
Scared	24.9	20.2	13.1	1.76	5.23	3.8	94.7

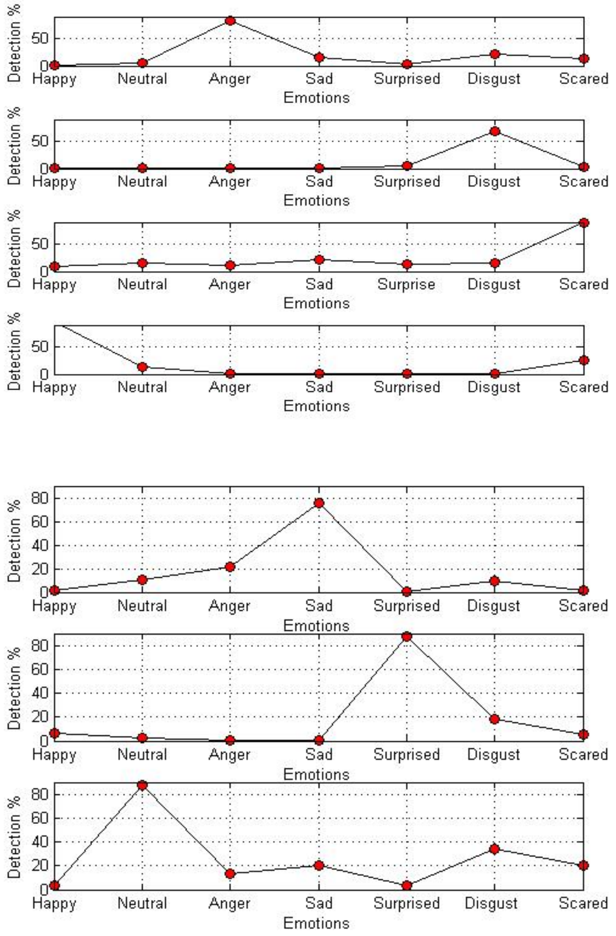

Fig. 2. Graphical evaluation of all seven emotions

A. CROSS-VALIDATION

To verify the accuracy of (IHMCNN-LSTM) our proposed model, we use cross-validation to evaluate (IHMCNN-LSTM) our proposed model and compare with another model, like CNN, - LSTM, and - CNN-LSTM, that have the same structure as our proposed model but without knowledge transfer. Tables II show the Comparisons between emotion detection rate with different models

TABLE II. COMPARISONS BETWEEN EMOTION DETECTION RATES WITH DIFFERENT MODELS.

Emotions	CNN	HAAR-CNN[24]	CNN-LSTM[25]	IHMCNN-LSTM
Happy	86.21	90.06	95.89	97.8
Neutral	83.13	93	89.24	98.8
Anger	75.23	94.7	84.92	96.98
Sad	85.23	87.01	91.94	93.3
Surprised	74.72	77.4	88.34	94.4
Disgust	70.38	72.03	87.97	91.4
Scared	79.38	88	86.12	94.7

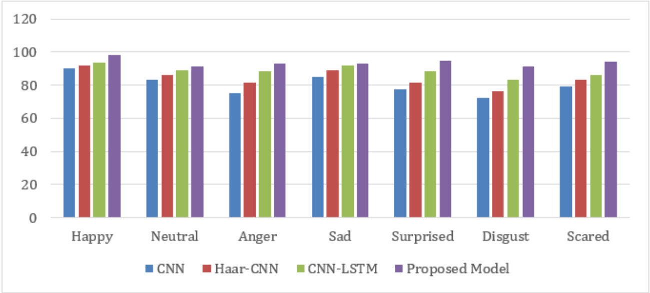

Fig. 3. Graphical evaluation of emotions comparison analysis with existing models

Table III shows the results obtained from leave-one-out cross-validation. The second column shows the number of successes in 1000 validations. The model performance is improved to 90.51%, from 58.62% (CNN-LSTM) as shown in table IV. which depicts the success of transfer learning. The results show that IHMNCNN-LSTM has better performance than CNN (68.1% accuracy) and LSTM (63.79% accuracy). However, if transfer learning and pre-training are not used in the first half of the same architecture, the accuracy is only 58.62%. After CNN architecture pretraining of transfer learning, recognition capabilities Accuracy improved by 90.51%. Although deep neural network architecture is important, recognition performance improvement through transfer learning is even more vibrant.

TABLE III. ACCURACY OF THE PROPOSED MODEL WITH DIFFERENT MODELS

Methods	Number of successes (1000)	Accuracy (%)
CNN	865	86.5
LSTM	675	67.5
CNN-LSTM	819	81.9
IHMCNN-LSTM	953	95.3

TABLE IV. COMPARISON ACCURACY WITH EXISTING MODELS

Methods	**CNN**	**LSTM**	**CNN-LSTM**	**IHMCNN-LSTM**
Accuracy (%)	68.1	63.79	90.51	95.32

VI. CONCLUSION

Humans can show thousands of facial expressions during their communication that depend upon their mood, person, behavior, and many other factors. In this paper, we have discussed the human emotion detection process using a deep neural network approach. The proposed model (IHMCNN-LSTM) combines CNN and LSTM and exploits the advantages of CNN and RNN. Leave-one-out cross-validation shows a significant improvement in model performance. The feasibility and practicality of this model have been validated. The face was displayed with Haar Cascading. Our experimental results show that our (IHMCNN-LSTM) proposed model can recognize facial expressions from video with an average accuracy of 95.32% much better compared to existing methods. In the future, we can apply this model to such social media platforms across multiple domains to analyze emotions in video calls. We can use this method in various social media applications like Twitter, WhatsApp, Messenger, etc. The mental health of any human can be detected while they are on chats or video calls. This can take out persons from trauma situations.

REFERENCES

[1] N. Sandhu, A. Malhotra and M. Kaur.,” Human Emotions Detection Using Hybrid CNN Approach”, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 9, Issue. 10, October 2020, pg.01 – 09.

[2] A. Jindal and R. Priya.,” Landmark Points Detection in Case of Human Facial Tracking and Detection”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019.

[3] A. Jindal and R. Priya,” Human Feelings Identification using Facial Gesture”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020.

[4] A. Jindal and R. Priya,” Human Face Tracking and Detection”, International Conference ICSCC-2020, G. D. Goenka University Gurgaon, April 2020.

[5] A. Lopes, E. de Aguiar, and A. F. Souza, and T. Oliveira-Santos, ‘‘Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order,’’ Pattern Recognition., vol. 61, pp. 610–628, Jan. 2017.

[6] J. Hopfield., ‘‘Neural networks and physical systems with emergent collective computational abilities,’’ Proc. Nat. Acad. Sci. USA, vol. 79, no. 8, pp. 2554–2558, 1982.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, and I. Sutskever, and R. Salakhutdinov, ‘‘Dropout: A simple way to prevent neural networks from overfitting,’’ J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[8] S. LIoffe and C. Szeged, ‘‘Batch normalization: Accelerating deep network training by reducing internal covariate shift,’’ 2015, arXiv:1502.03167. [Online]. Available: https://arxiv.org/abs/1502.03167

[9] X. Gloria, A. Bordes and Y. Bengio., ‘‘Deep sparse rectifier neural networks,’’ in Proc. 14th Int. Conf. Artif. Intell. Statist., Jun. 2011, pp. 315–323.

[10] E. Shelhamer, and T. Darrell, ‘‘Fully Convolutional Networks for Semantic segmentation,’’ in Proc. IEEE Conf. Computer. Vis. Pattern Recognition., Jun. 2015, pp. - 3431–3440.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yulee, ‘‘Deep Lab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,’’ IEEE Trans. Pattern Anal. Mach. Intel., vol. 40, no. 4, pp.- 834–848, Apr. 2017.

[12] Z. Yu and C. Zhang., ‘‘Image based static facial expression recognition with multiple deep network learning,’’ in Proc. ACM Int. Conf. Multimodal Interact., November. 2015, pp. 435–442.

[13] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in Proc. Int. Conf. Learn. Representations, September 2014, pp.- 1–4.

[14] M. Ghayoumi and A. Bansal, ‘‘Multimodal architecture for emotion in robots using deep learning,’’ in Proc. Future Technol. Conf. (FTC), December. 2016, pp. - 901–907.

[15] A. Graves, A. Mohamed, and G. Hinton, ‘‘Speech Recognition with Deep Recurrent Neural Networks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Vancouver, BC, Canada, May 2013, pp. -6645–6649.

[16] M. Sundermeyer, R. Schlüter, and H. Ney, ‘‘ LSTM neural networks for language modeling, ’’ in Proc. Inter-speech, September 2012, pp. 194–197.

[17] Z. Xu, S. Li, and W. Deng, ‘‘Learning Temporal Features using LSTM-CNN Architecture for Face Anti-spoofing,’’ in Proc. 3rd IAPR Asian Conf. Pattern Recognition., November. 2015, pp. 141–145.

[18] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, ‘‘Face Anti-spoofing database with diverse attacks,’’ in Proc. 5th IAPR Int. Conf. Biometrics (ICB), April. 2012, pp. -26–31.

[19] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, ‘‘Multimodal emotion recognition with evolutionary computation for human-robot interaction,’’ Expert Syst. Appl., vol. 66, pp.- 42–61, December. 2016.

[20] S. J. Pan and Q. Yang, ‘‘A survey on transfer learning,’’ IEEE Trans. Knowles. Data Eng., vol. 22, no. 10, pp. -1345–1359, Oct. 2010.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Region-based convolutional networks for accurate object detection and segmentation,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp.- 142–158, January . 2016.

[22] A. Mollahosseini, B. Hasani, and M. H. Mahoor, ‘‘AffectNet: A database for facial expression, valence, and arousal computing in the wild,’’ IEEE Trans. Affective Comput., vol. 10, no. 1, pp.- 18–31, Jan./March. 2019.

[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, ‘‘The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,’’ in Proc. IEEE Computer. Soc. Conference. Computer. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2010, pp. 94–101.

[24] N. Sandhu, A. Malhotra, M. Kaur,” Human Emotions Detection Using Hybrid CNN Approach”, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 9, Issue. 10, October 2020, pg-.01 – 09

[25] Tzuu-hseng s. Li, ping-huan kuo, ting-nan tsai, and po-chien luan,” CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot”, Received June 19, 2019, accepted July 4, 2019, date of publication July 11, 2019, date of current version July 30, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2928364. VOLUME 7, 2019. IEEE Access.