INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Real-time video based emotion recognition using convolutional neural network and transfer learning

J Sujanaa[1]*, **S Palanivel**[2]

1 Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalainagar, 608002, Tamil Nadu, India. Tel.: +91-7010181616
2 Professor, Department of Computer Science and Engineering, Annamalai University, Annamalainagar, 608002, Tamil Nadu, India

OPEN ACCESS

Received: 11-07-2020

Accepted: 15-08-2020

Published: 31-08-2020

Editor: Dr. Natarajan Gajendran

Citation: Sujanaa J, Palanivel S (2020) Real-time video based emotion recognition using convolutional neural network and transfer learning. Indian Journal of Science and Technology 13(31): 3222-3229. https://doi.org/10.17485/IJST/v13i31.1118

***Corresponding author**.

Tel: +91-7010181616
sujanaajohn@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: The deep learning approaches have paved their way to construct various artificial intelligence products and the proposed system uses a convolutional neural network for detecting real-time emotions of mankind. The objective of the study is to develop a real-time application for emotion recognition using convolutional neural network and transfer learning methods. **Methods/Statistical analysis**: The proposed system considers happy, normal and surprised categories of emotions. The system consists of four major steps: dataset collection, training, validation, and real-time testing. The dataset is comprised of face images containing emotions such as happy, normal and surprised in the form of video frames. The face and mouth regions are detected using the Haar-Based Cascade classifier at 20 frames per second. **Findings:** The convolutional neural network (CNN) is trained using mouth images and the pre-trained models VGG16 and VGG19 are trained with face images. The trained model is used to detect the emotions in the live webcam video. The experimental results show that the CNN model trained using mouth images gives an accuracy of 85.71% and the pre-trained models trained with face images using transfer learning method achieves an accuracy of 77.78%. The proposed system using CNN outperforms the pre-trained models for recognizing the emotions in real-time video. **Novelty/Applications:** The proposed system is entirely based on the mouth region video frames and the real-time emotion recognition system is developed. This work can detect the three emotions in an unconstrained laboratory environment.

Keywords: Convolutional neural network; mouth detection; pre-trained models; real-time emotion recognition; transfer learning

1 Introduction

The emotion recognition places a very indispensable role in inspecting human feelings and internal thoughts more precisely. The emotions and mood lead in identifying the human mind quickly. According to the psychologist, emotions are mainly for short time and mood is milder than emotion and it is a long-lasting one. Humans may interact with society through emotions in case of the absence of verbal communication [1]. Among the other non-verbal communications, emotions play a very effective way of exchanging internal thoughts with society. Emotions of humans can be detected through distinct ways such as their verbal responses or voice tone, physical responses or through the body-languages, autonomic responses, etc. [1]. The basic types of emotions in a person are happy, normal, surprised, fear, anger, disgust or dislike and sadness. Emotions like happy, normal, surprised, disgust and fear are easy to find out whereas other expressions like disgust, amusement, pride, contempt, and shame are very hard to find in human through facial expressions. There are varieties of applications for facial emotions recognition like student classroom behavior monitoring system, airport/railway suspicious person detection system, autism children expression detection, facial expression-based emotion chat applications, real-time person pain monitoring systems, etc. [2,3].

In this work, the basic expressions like happy, normal and surprised are considered in the mouth and face regions of persons to detect the emotions efficiently. The main issues for constructing a facial emotion recognition system using deep learning needs a larger amount of dataset, emotion recognition in varying lightening conditions and identifying the human expression in real-time under different scenarios. There are numerous architecture models found in the ImageNet competition in the past years which are useful in redefining new models for every problem when their layers are properly fine-tuned and frozen, gives better accuracy and it minimizes the workload to develop a completely new architecture for each problem, thereby reducing the complex task with reusing those pre-trained models instead of training from the scratch.

The leading applications in the industries have used deep learning methods for identifying facial emotions in photos and videos. The dataset used in the automated facial expression recognition systems viz. Extended Cohn-Kanade Dataset (CK+), Facial Expression Recognition 2013 (FER-2013), Denver Intensity of Spontaneous Facial Action Database (DIFSA), Japanese Female Facial Expression Dataset (JAFFE), MMI Facial Expression Database, Emotion Recognition in the Wild Challenge (EmotiW) dataset, Radboud Faces Database (RaFD) datasets [4–7]. In facial image-based emotion recognition techniques, a hybrid method was proposed combining CNN and Recurrent Neural Network (RNN) classifiers using Rectified Linear (ReLU) activation units and it gives better accuracy of 94.46% [6]. In a combined approach for both facial expression recognition and gender classification [8], the feature extraction is done using the viola-jones algorithm. The eye and nose region are detected using Haar-cascade classifiers and lip corners are detected using the Sobel edge detection method where 19 patches are extracted from the segments with the landmark. These landmarks are trained with Quadratic Discriminant Analyzer (QDA) and Support Vector Machine (SVM) classifier gives better accuracy than the state-of-art methods. In a video frame-based emotion recognition system [7], a method based on LBP along with the Adaboost algorithm to read the Linear Binary Pattern (LBP) features and then fed to Gaussian Mixture Models (GMM) for emotion classification and this model gives maximum accuracy and minimum time consumption. The Electroencephalography (EEG) signals are significant features for classifying six emotional states as they are better suitable for clinical diagnosis. In an unsupervised learning method called hypergraph-based emotion recognition method [9], acoustic features like MFCC (Mel Frequency Cepstral Coefficients) combined with epochs (glottal closure) features from speech signals are used and it gives good accuracy than their individual accuracies. The collective feature group comprising of MFCC, spectral centroids and MFCC derivatives along with the bagged ensemble methods consisting of 20 support vectors gives an accuracy of 75.69% [10]. The feature fusion vector is formed with DBN (Deep Belief Network) features and statistical features like Electro-Dermal Activity (EDA), Photoplethysmogram (PPG) and Zygomaticus Electromyography (zMEG) and is used for classifying the emotions using Fine Gaussian Support Vector Machine (FGSVM) gives 89.53% overall accuracy [11]. A method was proposed for combining MFCC with Residual features to extract useful information from each emotion and models are created to detect the music emotion recognition system using Auto Associative Neural Network (AANN), SVM and Radial Basis Functional Neural Network (RBFNN) where SVM shows highest accuracy of 99.0% for combined features than the other classifiers [12]. In the transfer learning method, pre-trained models like (AlexNet [13], VGG-S [14], VGG-M [14], and VGG-VD16 [15]) are used to extract the low-level features using "MatConvNet" toolkit to predict the human activity in the surveillance video camera. The author, Mehmet Akif OZDEMIR [2] identified seven emotions by training the LeNet architecture and obtained a validation accuracy of 91.81%. The CNN model is adopted by the authors by Denis Sokolov and Mikhail Patkin [3] for detecting emotions in real-time using iPhone SE or higher versions smartphones where the CNN model is trained for 1 hour using WeSee dataset and obtained an accuracy of 63.01% on test data. The eye-region is taken for emotion recognition instead of considering the entire face images to detect emotions in real-time [16]. The authors generated RGB images from the Spectrogram arrays of the speech signals with 8 to 4 kHz frequency. The AlexNet is used as the pre-trained CNN model to detect the emotions with 79.7% average weighted accuracy [17].

2 Proposed Methodology

The objective of this work is to detect the emotion of the person in real-time using CNN architecture and transfer learning of pre-trained models. The system consists of these main steps: mouth and face detection, training, validation, and real-time testing. The mouth emotion images are trained using CNN architecture and the face emotion images are Keras library[(18)] with a backend TensorFlow library[(19)] surprised as these emotions are the most prevailing emotions expressed through mouth region.

2.1 Mouth and Face Detection

The "OPENCV- 4.1.0" is a Python open-source library for image and video processing, packaged with pre-trained Haar-feature based cascade classifiers which consist of XML files for detecting face, eye and mouth region in the image. This library is specially developed to potentially teach a machine about the objects existing in real life. This cascade classifier is an algorithm created by Paul Viola and Michael Jones[(20)] using machine learning techniques. These cascade classifiers are trained with samples containing face images and non-face images. In this work, the Haar-based cascade classifier is used for detecting the mouth and face in every video frame.

2.2 CNN

The CNN is a deep neural network (DNN) performing numerous operations in various layers like convolution, sub-sampling, flattening and action similar to backpropagation type of learning in the dense layers. In the convolution layer, the convolutional filter will be applied over the input matrix or another dot product operation may be performed. In the sub-sampling layer, a max-pooling operation is performed which involves reducing the size of the matrix by preserving important information in the feature maps. Then those values are flattened, followed by fully connected layers and the last layer is the output layer with Softmax activation to classify the outputs to their respective classes thereby following the strategy of supervised learning technique. The CNN has several nodes/neurons which has the capacity to learn the weights and biases by itself through continuous training as shown in Figure 1.

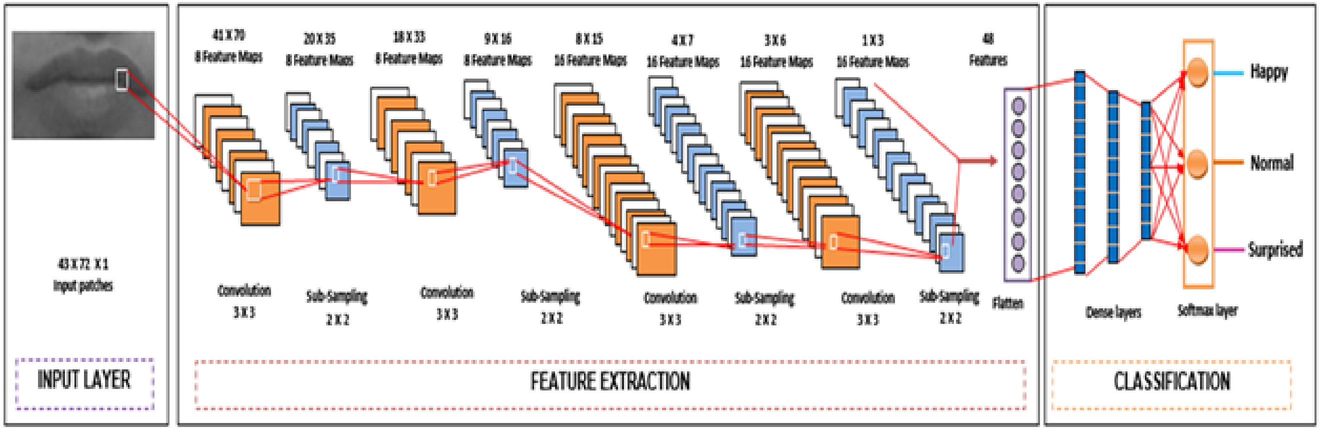


Fig 1. Proposed CNN architecture.

The images first undergo a simple pre-processing step where the RGB images are converted to grayscale images. We have used the Scikit-library[(21)]for pre-processing the input frames. The frames are converted into a form of NumPy array to feed to the CNN. The training samples 6720 are converted to the form, (6720, 43, 72, 1) and validation 1680 are converted to the form (1680, 43, 72, 1). In the CNN model, the dataset is passed as input to the CNN models to classify and detect mouth emotion expressions, which includes four convolution layers and four max-pooling layers, and then they are flattened. The flattened output is given as input to the varying number of fully connected dense layers. The initial layer consists of the raw pixels of the mouth image of the size 43x72. The first convolution layer with 8 filters performs a dot product of their weights and input image pixel values and reduces the dimension to 41x70 (8 images). The first max-pooling layer is applied along the spatial dimension (height x width) to perform a downsampling operation and this layer reduces the dimension to 20x35. In the second layer, the CNN filter of size (3x3), 8 filters are applied and it gives output dimension as 18x33 (8 images) followed by the second max-pooling layer of size (2x2) is applied and it gives output 9x16. In the third layer, the CNN filter of size (3x3), 16 filters, gives the output as 8x15 (16 images). The third max-pooling layer of size (2x2) gives output 4x7. In layer 4, CNN filter of size (3x3), 16 filters, gives output 3x6. The max-pooling layer (2x2) for the fourth time is applied and it gives output 1x3 (16 images). Then the images are flattened in layer 5 to form 48 features. These 48 features are passed to the dense layers which are comprised of a varying number of hidden neurons. Different models have been developed with varying numbers of dense layers and their hidden units. The last layer will be the output layer with Softmax activation function comprising three output neurons which will classify the output into normal, happy and surprised categories for the proposed mouth-based emotion recognition system.

2.3 Transfer learning

The pre-trained CNN architectures are the models that are already trained with a subset of ImageNet dataset during ImageNet Competitions and have learned the weights on that larger dataset. So, the weights of these models can be used to learn the proposed emotion recognition problem. The knowledge obtained from one domain can be applied to another domain is said to be 'transfer learning'. Transfer learning can be done using two methods: fine-tuning and freezing. In the fine-tuning method, the pre-trained model layers are tuned with varying filters, layers and hidden units to optimize their learning in current problems thereby increasing the accuracy in learning the newly defined problem. In the case of freezing, the pre-trained model layer weights are frozen (locked), thereby not allowing those weights from being changed during the current training. The pre-trained models reduce the burden of training the CNN architecture models from scratch. The pre-trained convolution base consists of various layers like convolution block, max-pooling layer, rectified linear activation unit (ReLU), batch normalization layer, separable convolution layer, inception layers, etc. These blocks that are already trained for the ImageNet dataset are frozen to preserve/lock their weights from being trained for the proposed emotion recognition task. Figure 2 shows the illustration of transfer learning used in the proposed facial emotion recognition.

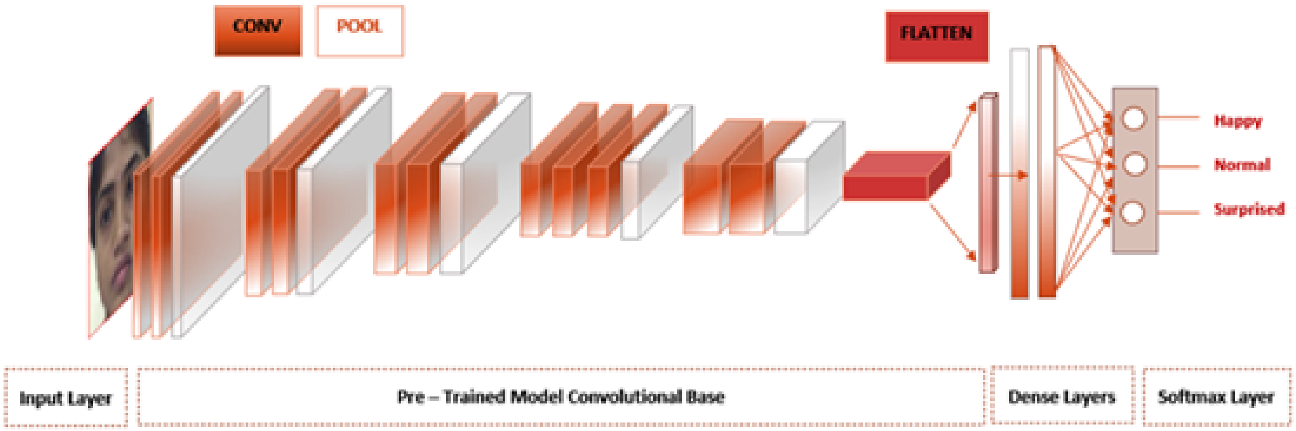


Fig 2. Illustration oftransfer learning for emotion recognition using a pre-trained model.

The pre-trained models like Visual Geometry Group (VGG) are made the publicly available ConvNet models for all computer vision problems used in image and video recognition tasks, where VGG16 and VGG19 are used in this work to find out the emotions. VGG was developed by K. Simonyan and A. Zisserman [15], the runner-up model in the ImageNet competitions conducted in the year 2014 called ImageNet Large-Scale Visual Recognition Challenge(ILSVRC) [22] for classifying the subset of ImageNet database objects into 1000 classes and trained on NVIDIA Titan Black GPU's. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. The architecture of VGG16 and VGG19 are described [13].

3 Experimental Results

3.1 Data sets

The dataset is collected using a web camera with a resolution of 1280 x 720 in the laboratory environment. A total of 8400 mouth images and 5400 face images are used from 20 subjects (10 Male, 10 Female). The dataset is divided into training and validation sets. For training the CNN architecture, 6720 mouth images are used for training and 1680 mouth images are used for validation. For testing, 2100 mouth images are used for each expression in real-time. For training using transfer learning methods, 4320 face images are used for training and 1080 face images are used for validation and 2100 face images are used during real-time testing.

3.2 Training and Validation

The four-layer CNN models are trained with varying numbers of dense layers such as two, three, four, five and six. The models are trained with various numbers of parameters and their weights are adjusted for 50 epochs with patience=5. The model with two dense layers is trained with a total of 4,731 training parameters gives good classification results. The training time of this model is 8 min 7 sec on a system with 2.29 GHz CPU. In the simple four-layer CNN with two dense layer model, the training loss and accuracy saturated at 19 epochs and yields an accuracy of 99.05% for the training data and 96.96% for the validation data and it is shown in Table 1.

Table 1. Loss and accuracy for training and validation data using CNN models

CNN Model	No. of Dense layers	Training Time (mm:ss)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Model 1	2	8:07	0.0261	0.9905	0.1229	0.9696
Model 2	3	6:49	0.0242	0.9930	0.1595	0.9672
Model 3	4	7:30	0.0305	0.9905	0.1759	0.9625
Model 4	5	5:00	0.0622	0.9804	0.1825	0.9613
Model 5	6	5:08	0.0602	0.9804	0.1784	0.9589

For the pre-trained models, varying number of layers are frozen and trained on free cloud-based service, Google Colab network GPU's [23]. All the models are trained for 25 epochs and the model is saved with patience=5. The loss/error and accuracy for the training and validation data for different models are shown in Table 2, where model 6 and model 7 are obtained from VGG16, and model 8 and model 9 are obtained from VGG19 using transfer learning.

Table 2. Loss and accuracy for training and validation data using transfer learning models

Transfer Learning Model	No. of Layers Frozen	Training Time (mm: ss)	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Model 6	15	19:21	0.0272	0.9919	0.0325	0.9796
Model 7	14	18:01	0.0150	0.9970	0.0948	0.9898
Model 8	18	23:70	0.0215	0.9947	0.0562	0.9815
Model 9	17	14:06	0.0201	0.9944	0.0474	0.9833

In model 6, 15 layers are frozen and trained with the output layer 3 neurons and it has 12,291 trainable parameters. In model 7, 14 layers are frozen and dense layer hidden neurons are changed from 4096 to 512 neurons in the first dense layer and 128 neurons in the second dense layer and given to 3 output classes with a dropout of 50% and it has 66,051 trainable parameters. In model 8, 18 layers are frozen contains 12,291 trainable parameters. In model 9, 17 layers are frozen and dense layer hidden neurons are changed from 4096 to 512 neurons in the first dense layer and 128 neurons in the second dense layer and given to 3 output classes with a dropout of 50% and it has 66,051 trainable parameters.

3.3 Real-time testing

The CNN models are tested with real-time video for varying numbers of dense layers. The CNN model that gets trained with 2 dense layers performs well compared to 3, 4, 5 and 6 dense layer models. The two-layer dense model, when tested for every frame gives 71.88% accuracy. For each emotion, 700 mouth images are extracted from the live webcam and it is used to evaluate the performance of the system for consecutive 'n' images (n= 1, 3, 7 and 10).

Similarly, when the model is tested for every 3 frames and 7 frames gives 75.71% and 80% respectively. The proposed CNN model with three dense layers achieves an accuracy of 85.71% for every 10 frames. Among the four models obtained using transfer learning methods, model 7 and model 9 gives the highest validation accuracies and these models are tested in the real-time video to evaluate the performance of the system. For each emotion, 700 face images are extracted from the live webcam and it is used to evaluate the performance of the system for consecutive n images (n=1, 3, 7 and 10). The snapshot of real-time emotion recognition.

The performance of the classifier can be evaluated using measures like precision, recall, F1-score, and accuracy. The precision, recall, F1-score, and accuracy for real-time testing images are shown in Table 3 for every 10 consecutive testing images. The performance of emotion recognition for real-time video is carried out where the Model 7 gives maximum accuracy of 77.78% using the transfer learning method.

Table 3. Performance of emotion recognition for real-time video

Model	n=1 (in %)	n=3 (in %)	n=7 (in %)	n=10 (in %)
Model 1	71.88	75.71	80.00	85.71
Model 7	69.52	71.43	73.33	77.78
Model 9	45.71	46.19	49.27	53.49

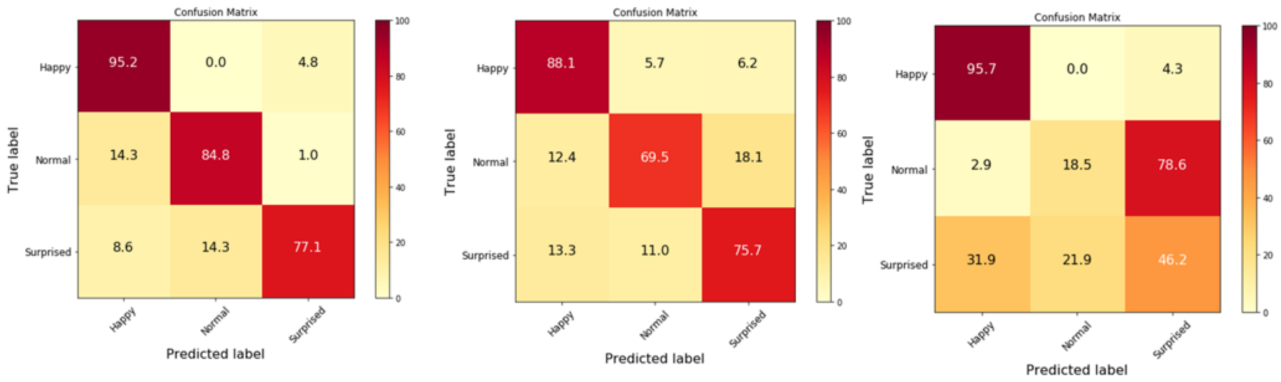
3.4 Performance analysis

The performance of the classifier can be evaluated using measures like precision, recall, F1-score, and accuracy. The precision, recall, F1-score, and accuracy for real-time testing images are shown in Table 4 for every 10 consecutive testing images.

Table 4. Performance of emotion recognition

Model	Emotions	Precision (in %)	Recall (in %)	F1-score (in %)	Accuracy (in %)
Model 1	Happy	95.23	80.64	87.32	90.79
	Normal	84.76	85.58	84.17	90.16
	Surprise	77.14	93.10	83.39	90.48
Model 7	Happy	88.09	77.05	81.22	87.46
	Normal	69.52	80.66	73.69	84.29
	Surprise	75.71	75.71	74.72	83.81
Model 9	Happy	95.71	73.36	83.06	86.98
	Normal	18.57	45.88	26.44	65.56
	Surprise	46.19	35.79	40.33	54.44

The confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for classification problems where the output can be of two or more types of classes. It is used to assess the performance of the classifier. In confusion matrix [(24)], true positive is the total number of emotion frames which are correctly identified as the respective classes, true negatives are the number of emotion frames which are correctly identified as other classes. The confusion matrix (in %) for real-time recognition of emotion using 10 consecutive face images is shown in Figure 3 . In model 1, 95.20% testing is correctly classified as happy classes, 85.80% testing as a normal class and 77.10% testing as a surprised class respectively. In total, 85.71% testing is correctly classified and 14.29% testing is misclassified as shown in Figure 3(a). Similarly, for model 7, 77.78% testing is correctly classified and 22.22% testing is misclassified to other classes is shown in Figure 3(b) and for model 9, 53.49% testing is correctly classified and 46.51% testing is misclassified to other classes is shown in Figure 3(c).



(a) *(b)* *(c)*

Fig 3. Performance of emotion recognition for real-time testing images using (a) Model 1 (b) Model 7(c) Model 9

4 Conclusion

This work has proposed a real-time emotion recognition system for recognizing the emotions normal, happy and surprised using mouth images with CNN architecture and face images with pre-trained models. The face images are extracted in an unconstrained laboratory environment using a web camera. The experimental results show that the proposed system with CNN recognizes the three emotions using mouth images with an accuracy of 85.71% and the transfer learning of pre-trained models using face images gives an accuracy of 77.78%. The proposed system using CNN gives better performance than the pre-trained modes for recognizing the emotions in real-time video.

Acknowledgement

The authors are thankful to Annamalai University, Annamalainagar, for providing a good environment and well-equipped Vision and speech laboratory in the Department of Computer Science and Engineering to carry out research in an efficient manner.

References

1) De A, Saha A. A comparative study on different approaches of real time human emotion recognition based on facial expression detection. In: and others, editor. Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA. 2015;p. 483–487.
2) Ozdemir MA, Elagoz B, Alaybeyoglu A, Sadighzadeh R, Akan A. Real time emotion recognition from facial expressions using CNN architecture. *TIPTEKNO 2019 - Tip Teknol Kongresi*. 2019;p. 1–4.
3) Sokolov D, Patkin M. Real-time emotion recognition on mobile devices. In: and others, editor. Proc - 13th IEEE Int Conf Autom Face Gesture Recognition;vol. 787. 2018.
4) Majumdar A, Ward RK. Discriminative sift features for face recognition. *Canadian Conference on Electrical and Computer Engineering*. 2009;p. 27–30.
5) Happy SL, Routray A. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*. 2015;6(1):1–12. Available from: https://dx.doi.org/10.1109/taffc.2014.2386334.
6) Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*. 2018;115:101–106. Available from: https://dx.doi.org/10.1016/j.patrec.2018.04.010.
7) Kaya H, Gürpınar F, Salah AA. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*. 2017;65:66–75. Available from: https://dx.doi.org/10.1016/j.imavis.2017.01.012.
8) Anusha AV, Jayasree JK, Bhaskar A, Aneesh RP. Facial expression recognition and gender classification using facial patches. In: and others, editor. 2016 Int Conf Commun Syst Networks, ComNet. 2016;p. 200–204.
9) Liang Z, Oba S, Ishii S. An unsupervised EEG decoding system for human emotion recognition. *Neural Networks*. 2019;116:257–268. Available from: https://dx.doi.org/10.1016/j.neunet.2019.04.003.
10) Bhavan A, Chauhan P, Hitkul, Shah RR. Bagged support vector machines for emotion recognition from speech. . *Knowledge-Based Syst*. 2019. Available from: https://doi.org/10.1016/j.knosys.2019.104886.
11) Hassan MM, Alam MGR, Uddin MZ, Huda S, Almogren A, Fortino G. Human emotion recognition using deep belief network architecture. *Information Fusion*. 2019;51:10–18. Available from: https://dx.doi.org/10.1016/j.inffus.2018.10.009.
12) Nalini NJ, Palanivel S. Music emotion recognition: The combined evidence of MFCC and residual phase. *Egyptian Informatics Journal*. 2016;17(1):1–10. Available from: https://dx.doi.org/10.1016/j.eij.2015.05.004.
13) Karpathy A, Leung T. Large-scale Video Classification with Convolution Neural Networks. In: and others, editor. Proceedings of International Computer Vision and Pattern Recognition. 2018;p. 10–20.

14) Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*. 2016;116:1–20. Available from: https://dx.doi.org/10.1007/s11263-015-0823-z.

15) Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*. 2016;116:1–20. Available from: https://dx.doi.org/10.1007/s11263-015-0823-z.

16) Wu H, Feng J, Tian X, Sun E, Liu Y, Dong B. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In: MobiSys 2020 - Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services. Association for Computing Machinery, Inc. ;p. 448–61.

17) Lech M, Stolar M, Best C, Bolia R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Frontiers in Computer Science*. 2020;2. Available from: https://dx.doi.org/10.3389/fcomp.2020.00014.

18) Keras Library. . Available from: https://keras.io/.

19) Abadi M. Large-Scale Learning on Heterogeneous Distributed Systems. *Prelim White Pap*. 2016;1(212):1–19. Available from: http://download.tensorflow.org/paper/whitepaper2015.pdf.

20) Viola P, J M. Rapid Object Detection using a Boosted Cascade of Simple Features. . *Comput Vis Patter Recognit*. 2001;394(1-3):1–9. Available from: https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf.

21) Pedregosa FF, Michel V, Grisel OO, Blondel M, Prettenhofer P, Weiss R. Scikit-learn: Machine Learning in Python. *J Mach Learn Res [Internet]*. 2011;12:2825–2855. Available from: http://scikit-learn.sourceforge.net.

22) ImageNet Overview . 2019. Available from: https://en.wikipedia.org/wiki/ImageNet#ImageNet_Challenge.

23) Bisong E, Bisong E. Google Colaboratory. In: and others, editor. Building Machine Learning and Deep Learning Models on Google Cloud Platform. 2019.

24) Ganesan K. Text Mining, Analytics & More: Computing Precision and Recall for Multi-Class Classification Problems. 2015. Available from: http://text-analytics101.rxnlp.com/2014/10/computing-precision-and-recall-for.html.