ISSN 2278-3091
Volume 11, No.1, January - February 2022
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse031112022.pdf
https://doi.org/10.30534/ijatcse/2022/031112022

Emotion Classification from Facial Images and Videos Using a Convolutional Neural Network

Chevella Anil Kumar[1], Kancharla Anitha Sheela[2]
[1]JNTUH College of Engineering Hyderabad, Kukatpally, Hyderabad-500085, India
[2]JNTUH College of Engineering Hyderabad, Kukatpally, Hyderabad-500085, India

Received Date : December 08, 2021 Accepted Date :January 09, 2022 Published Date : February 06, 2022

ABSTRACT

As a result of its wide range of academic and commercial applications, emotion recognition seems to be a important subject in computer vision & artificial intelligence. The majority of the decisions we make in our life are influenced by emotions. In this technology advancement, researchers found that properly categorising of human emotions may be a major source of development for companies in digital marketing. And that is what we will indeed be focusing on reading emotions of human being from facial image and videos. In the world of artificial intelligence, this concept falls under the category of cognitive systems. Facial expressions are essential to take into account while researching human behaviour including psychological characteristics. In this work, we used deep learning algorithms that recognise basic seven emotions through facial expressions (FER) and videos: happy, surprise, disgust, anger, neutral, sadness, and fear (VER). Deep learning has the potential to improve human-machine communication interaction because of its ability to learn features will allow machines to develop perception. To classify the emotions from facial images using deep learning techniques, we created the Convolution Neural Network Model and trained it on fer2013, a database of pre-recorded images with various emotions. And for emotion recognition from videos, we segment the video into individual frames at 30 frames per second and repeat the process of facial images on each frame, then do sentiment analysis, and finally reframe the emotional analysis output video with all of the available emotional individual frames.

Key words: CNN; cognitive science; Computer vision; FER; fer2013; VER.

1 INTRODUCTION

Emotional analysis is the examination of human feelings, attitudes, and emotions at various levels of the brain. The different states of the human brain reflects the different emotions on human facial expressions. Science involving cognitive processes within computer systems is referred to as cognitive science. Cognitive science's main objective is to learn more about the human brain and indeed the principles that underlie its intelligence [1]. Computer systems related to human intelligence knowledge may be built to analyse face expressions in pictures and videos for emotional content. Face and Facial Component Detection, Extraction Of features, and Expression Classification are the three main stages in the conventional method to Facial Emotion Recognition as shown in figure 1. Facial features such as eyes and nose, other landmarks are first retrieved from a face in the image. Mainly two features are retrieved from the face components: spatial and temporal features. Next, classify the different emotion with pre-trained classifiers such as Adaboost, support vector machine and random forest classifier. It has become clear that deep learning would be the way to go for most machine learning problems because of the existence of large amounts of big data [2] and conventional methods use handmade characteristics in contrast to this.

There is a lot of work being done in the area of computer vision [3]with emotion detection. It's now possible to build intelligent systems that can correctly recognise emotions because to the increasing popularity using Machine Learning and Deep Learning methods. However, advances in disciplines such as psychology as well as neuroscience that are directly linked to emotion detection show that the issue is getting more complicated. Emotion identification may become very complicated when some of these factors are taken into account such as micro expressions, speech tonality, electroencephalography (EEG) data, facial expressions, gestures, and the surrounding environment as well as the limits and problems with existing Computer Vision algorithms. Figure 1 illustrates the stages in a FER system: The process of acquiring images and performing pre-processing steps is known as image acquisition.

Figure 1: Typical Architecture of Facial Emotion Recognition

To get an accurate facial expression categorization, it's critical to provide the classifier with as much relevant evidence as possible under the optimum circumstances. To do this, every input picture is first pre-processed using a conventional FER system.

By enabling "end-to-end" learning associated with the input pictures, deep-learning-based FER methods eliminate the need for pre-processing techniques like face physics models. There are several deep learning networks available, but the CNN, most often used because of its simplicity and easy understanding sown in figure 2. CNN-based methods create a feature map from an input picture by running through a filter

collection inside the convolution layers and each feature map was merged to create fully connected networks. And finally, emotions are classified to a specific class using softmax algorithm.

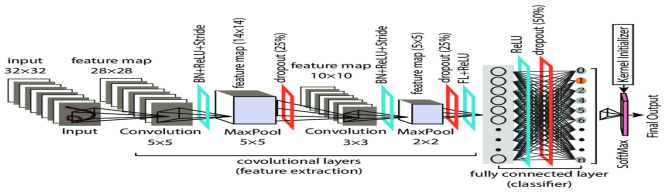


Figure 2: Convolution Neural Network

2 LITERATURE REVIEW

Artificial intelligence and psychological human emotion recognition are the two main areas of research throughout automatic emotion recognition. Verbal and nonverbal information like facial expression changes, voice tone and physiological signs [4] may be used to identify a person's emotional state. During 1967, Mehrabian [5] found that visual cues accounted for 55% of emotional information, vocal cues for 38%, and verbal cues accounted for 7%. That communication channel has piqued the interest of many academics since the earliest indications of an individual's emotional state were being sent via their facial expressions. The tough and delicate process of extracting characteristics from one face and applying them to another must always be accomplished if a better categorization would be to be achieved. In conventional method Facial motions are characterised by Action Units, and the human face was divided into 46 AUs action units, with every AU associated including one or more facial muscles, which have been developed in 1978 by Ekman and Freisen. It's a tough job since each individual communicates his or her emotion during his or their own unique manner. The automatic FER is the most studied by researchers when compared to other modalities to statistics developed by Philipp et al. [6], In just this region, there are many barriers and difficulties to be aware of, such as differences in head position, brightness, age and gender and also the issue of occlusion caused through Goggles, Skill Sickness, scarf, and on and on.

Geometric and texture characteristics like local binary patterns were classic techniques for extracting face features. Facial action components include LBP [7], FAC [8], and Gabor wavelets, as well as local directional patterns [9]. Employing deep learning with emotion recognition has been a huge success throughout recent years due to the results achieved with its designs with its architecture that allow for automatic extraction of features and classification, such as convolutional network [10] and the recurrent neural network (RNN).

3 METHODOLOGY:

There are two kinds of FER: those that utilise images and those that use video images [11]. Static (frame-based) FER is based entirely on static facial characteristics derived from images. Dynamic (video-based) FER, uses spatio-temporal characteristics to capture the dynamics of facial expression sequences from videos. The inclusion of temporal information to dynamic FER improves its recognition rate over static FER, but it also comes with certain drawbacks. That example, the retrieved dynamic features vary depending upon that faces in terms of transition times and facial expression feature properties. Furthermore, the use of temporal normalisation to generate expression sequences with a fixed number of frames may result in the loss of temporal scale information.

3.1 Recognizing Facial Emotions from Images

In past decades, deep-learning algorithms like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have made major advancements in computer vision. Deep learning algorithms were used to extract features, classify, and recognize objects. Using a CNN has the primary benefit of eliminating or significantly reducing the dependence on physics-based models as well as other pre-processing methods since it allows for direct end-to-end learning from input pictures. Due to the combination of these two elements, many areas including that of object, face and emotion recognition have benefited from CNN's cutting-edge findings.

The flow logic of FER: Reading input images and videos that need to be analyzed. Then this input data is given to the multi-cascaded convolutional neural network classifier to train the database and save the best weight features model. After that, the best model for detecting emotions was invoked with the input object sent along. The final result is a collection of emotions, each with a value assigned to it. Furthermore, the 'top emotion' function can isolate the object's highest valued emotion and return it on a scale of 0 to 1.

3.2 Collection of dataset

In deep learning to get the best weight model we have to train the CNN network with the standard available database. To do this task we have several FER databases are available to the researchers. Each database is unique in terms of variations in their standards like size of the database, face pose, resolution and population. The following table shows the few examples of standard databases. And in this paper we used the database of FER2013[12].

Table 1: List of some emotional databases

Databases	Descriptions	Emotions
MultiPie	More than 7,50,000 captured by 15 view and 19 illumination videos	Anger, Disgust, Neutral, Happy, Surprise
MMI	2900 videos	Neutral and Basic six emotions
GEMEPFERA	289 images Sequences	Anger, Fear, Sadness, Relief and Happy
SFEW	700 Images with different ages, illumination and head pose	Neutral and Basic six emotions
CK+	593 videos for posed and non-posed expressions	Neutral and Basic six emotions
FER2013	35,877 gray scale images with different emotions	Neutral and Basic six emotions
JAFFE	213 gray scale images posed by Japanese females	Neutral and Basic six emotions

fer2013 standard dataset comprising faces in grayscale, each 48*48 pixels in size. That dataset has 35,877 emotional pictures with a training set of 28,709 images as well as a test set comprising 7168 images, and that it contains seven distinct emotions with associated labels (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

Figure 3: FER2013 database

3.3 Data Pre-processing

Machine learning and deep learning image processing applications need pre-processing methods such as resizing, reshaping, converting to grayscale, and normalising. Utilizing vectorization and OpenCV, create Numpy arrays from the pictures and use pandas to categorise the results.

3.4 Building the model using Convolution Neural Network

Deep learning algorithms includes Convolution Neural Network which encode input data like multidimensional arrays and perform well with huge quantities of labelled data and it is a feed-forward neural network is often used in artificial intelligence. Every portion of the image is known as receptive field, which is extracted by CNN. And it weights every neuron according to the significance of the receptive field. As a result, it's able to tell which neurons are more important than others. Here, we built a CNN model using the hyper parameters listed below.

- 16 convolution layers
- Zero Padding and
- maxpooling with 2x2 strides and
- ReLu activation function in convolution layers and
- 3x3 filters kernel size
- Softmax activation at fully connected layer
- Dropout=0.5
- Batch size=32
- Epochs=30
- Stochastic gradient descent optimizer
- Categorical crossentropy for calculating loss function

Convolutional Layer – this layer convolving the input image neurons with the different filters and generate the feature maps. This output may be calculated using the formula below:

$$Conv_{x,y} = \sum w_i * v_i \quad (1)$$

w_i =kernel weights, v_i =values of the input map's spatially corresponding elements

$$Z_{x,y} = \sum w_i * v_i + b_i \quad (2)$$

$z_{x,y}$ is the output of the layer with adding bias value b_i

Rectified Linear Unit Layer − it applies an activation function to the output from the preceding by setting thresholds at zero; it introduces non-linearity into the network.

$$R_{x,y} = \max(0, Z_{x,y}) \quad (3)$$

Pooling Layer − with each new node in the network, we can reduce the amount of data we need to keep by pooling. This pooling layer spatially resizes every input depth slice separately. The kinds of pooling it covers are many and including average, maximum and L2 pools. Filter (F) and stride (S) are the two hyper parameter in this study.

In general, if we have input dimension W1 x H1 x D1, then

W2 = (W1−F)/S+1

H2 = (H1−F)/S+1

D2 = D1

Where W2, H2, and D2 are the width, height, and depth of output.

Flattening- Flattening is converting the data into a 1-dimensional array for inputting it to the next layer. We flatten the output of the convolution layers to create a single long feature vector. And it is connected to the final classification model, which is called a fully-connected layer

Fully Connected layer- Each neuron inside the previous layer was completely connected to each neuron inside the following layer, similar to what they've been connected in conventional neural networks. W is indeed the number of unrestricted parameters inside the network and X is indeed a k-dimensional input.

$$F(X,W) = \emptyset(W * X), \text{ where } \phi \text{ is Non-Linear Activation function preferably Bi-Polar sigmoid}$$

Output layer − the output scores from the previous layer are computed in this layer. The resulting output is of the size 1×1× L, where L is the number of training dataset classes.

Softmax Layer- The softmax layer propagates network errors backwards, and indeed the following equation may be used to compute the grade for each class.

$$S(x) = \frac{e^{x_i}}{\sum_{j=0}^{k} e^{x_j}} \quad i=0,1,2.....k \qquad (6)$$

3.5 Training and testing the model

For training, we used the above built CNN model to train fer2013 emotional database for building the best weight feature model and saved this model in HDF5 data file. Its best weight feature model was stored after 30 epochs if the validation accuracy remains constant or earlier stopping point.

3.6 Recognizing Facial Emotions from Videos

A video is a collection of sequentially changing images, or frames. As a consequence, we'll apply the same techniques to videos like we did to images or individual frames for the purpose of emotion detection[13] As a result, any algorithm that applies to both videos and pictures is the same. The only extra step for video processing would be to split the video into every one of its constituent frames and afterwards apply image processing techniques to recognize the emotions from images.

The flow of Logic: we will follow few more steps for video analysis even if the basic technique should be the same for both images and videos.

- Video analyze (): this step is used to extract the individual frames of videos with 30fps and analyzing them individually.
- Each frame analysed through this method is saved as a distinct image in the code's working directory's root folder. Additionally, this function duplicates the original video simply drawing a box from around face and displaying genuine emotions in the video clip.
- From these processed values, we generate a Pandas Data Frame, which we plot using matplotlib. We can see every emotion plotted against time in this plot.
- Using the model, we can go further into this data frame to recognize the particular emotion values of each frame and determine which sentiment would have been most prominent during the video.
- By separating and analysing individual frames from videos, we may perform FER algorithms on them to analyze the emotional analysis from video. The following figure 4 shows the Block Diagram of Emotion Recognition from Videos.

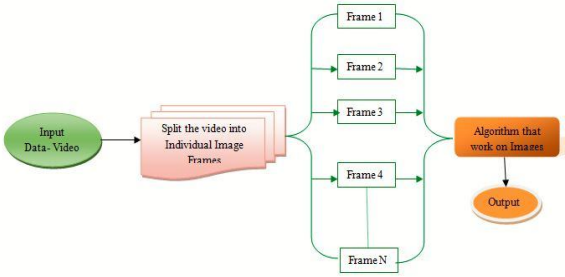
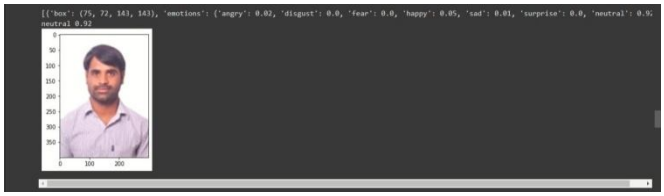


Figure 4: Block Diagram of Emotion Recognition from Videos

4 EXPERIMENTAL RESULTS

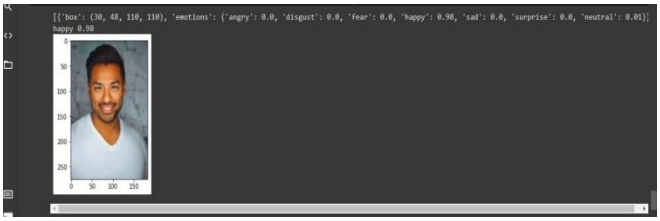
A Convolutional Neural Network model has been used to classify the seven distinct emotions from images and videos of people's facial expressions. For FER, we used a CNN model to detect emotions in a local image and showed an emotion value ranging from 0 to 1 next to it as shown in figure 5. Ultimately, the input image's emotion is assigned a class value based on the image's highest class. For Video-Based Emotion Recognition, we used a video as input; transform it into individual images/frames with 30 frames per second, then using the CNN model to assess the emotion status of each image and emotion values are shown next to each face in a image encircled by a rectangular box. Finally, the method publish a new video in .mp4 format that will have a box around the face of the human with live emotion value and also generate the zip file of individual frames of a live video as shown in figure 6. And also we plot the results emotion versus its prominent value in the video shown in figure 7 and Emotions against the time in the video as shown in figure 8.



(a)

(b)

(c)



(d)

Figure 5: (a), (b), (c) and (d) shows the result of different emotions from different images

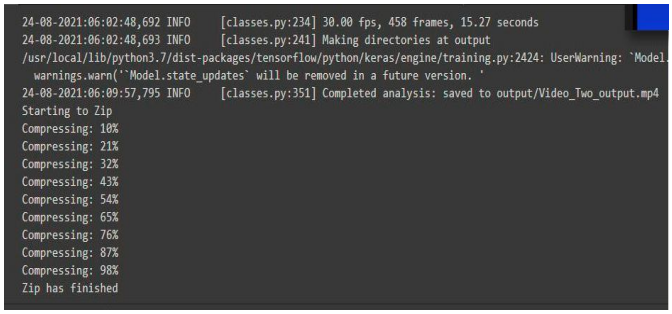


Figure 6: converting lives video into frames, analyzes the emotions of each frame and generates the output .mp4 file and zipped file of individual frames

```
13  score_comparisons = pd.DataFrame(emotions, columns = ['Human Emotions'])
14  score_comparisons['Emotion Value from the Video'] = emotions_values
15  print(score_comparisons)
```

	Human Emotions	Emotion Value from the Video
0	Angry	20.80
1	Disgust	0.00
2	Fear	32.58
3	Happy	137.37
4	Sad	76.43
5	Surprise	13.04
6	Neutral	177.10

Figure 7: Emotion vs it's prominent in the live video

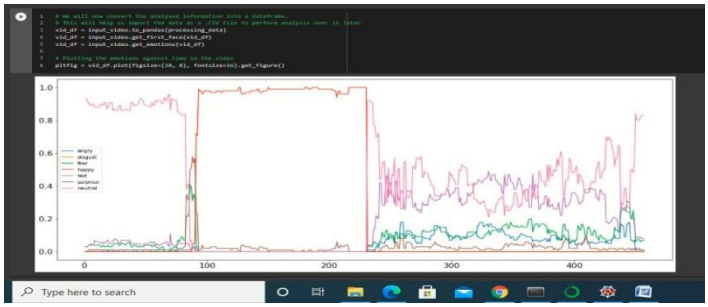


Figure 8: Emotions against the time in the video

5 CONCLUSION

The Proposed work is designed to develop a CNN network models to predict the emotions from facial image and videos. The classified emotions are represented in seven states Happy, sad, angry, surprised, neutral, disgust and Fear with corresponding scores in between 0 and 1 as shown in above results. We trained the network with fer2013 dataset to generate the best model to recognize and classifying the emotions from facial image and videos. In our model trained dataset contain total 35,877 emotional images with training set consists of 28,709 images and the test set consists of 7168 images. For this model we achieved the accuracy of 97%. To further increase the accuracy of the model, we can expand the training dataset.

REFERENCES

[1] S. Mukherjee, A. K. Gupta, S. Aziz, T. Aziz, P. Jaiswal, and S. Chatterjee, **"Cognitive science: A review,"** in *2017 4th International Conference on Opto-Electronics and Applied Optics (Optronix)*, Nov. 2017, pp. 1–7, doi: 10.1109/OPTRONIX.2017.8349979.

[2] S. E. Kahou *et al.*, **"EmoNets: Multimodal deep learning approaches for emotion recognition in video,"** Mar. 2015, [Online]. Available: http://arxiv.org/abs/1503.01800.

[3] A. Pangestu Lim, G. Putra Kusuma, and A. Zahra,

"Facial Emotion Recognition Using Computer Vision."

[4] L. Shu *et al.*, "**A review of emotion recognition using physiological signals,**" *Sensors (Switzerland)*, vol. 18, no. 7. MDPI AG, Jul. 01, 2018, doi: 10.3390/s18072074.

[5] C. Marechal *et al.*, "**Survey on AI-based multimodal methods for emotion detection,**" in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11400, Springer Verlag, 2019, pp. 307–324.

[6] P. Werner, F. Saxen, and A. Al-Hamadi, "**Facial Action Unit Recognition in the Wild with Multi-Task CNN Self-Training for the EmotioNet Challenge.**"

[7] C. Shan, S. Gong, and P. W. McOwan, "**Facial expression recognition based on Local Binary Patterns: A comprehensive study,**" *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009, doi: 10.1016/j.imavis.2008.08.005.

[8] M. H. Alkawaz, D. Mohamad, A. H. Basori, and T. Saba, "**Blend Shape Interpolation and FACS for Realistic Avatar,**" *3D Res.*, vol. 6, no. 1, Mar. 2015, doi: 10.1007/s13319-015-0038-7.

[9] T. Jabid, M. H. Kabir, and O. Chae, "**Robust facial expression recognition based on local directional pattern,**" *ETRI J.*, vol. 32, no. 5, pp. 784–794, Oct. 2010, doi: 10.4218/etrij.10.1510.0132.

[10] C. Anil kumar and K. A. Sheela, "**Emotion Recognition from Facial Biometric System Using Deep Convolution Neural Network (D-CNN),**" 2021, pp. 381–391.

[11] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "**Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition,**" *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, Apr. 2019, doi: 10.1109/TAFFC.2017.2695999.

[12] "**https://www.kaggle.com/msambare/fer2013.**" .

[13] Institute of Electrical and Electronics Engineers and PPG Institute of Technology, *Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES 2020) : 10-12, June 2020.*