Facial Expression Recognition by Using Convolutional Neural Network

[1]Sajja Radharani[*], [2]G Kalyan Kumar, [3]B Ravi Teja,
[4]Ch Varun Kumar
[1,2,3,4] Department of CSE, VFSTR Deemed to be University, Vadlamudi, Guntur, India.
radharani.sajja6@gmail.com

Abstract

Many fields, especially image processing, medical science, and machine learning, place a significant emphasis on emotion recognition systems. The use of emotion recognition systems is widespread, and they are essential to fault detection and gaming applications. Therefore, automatic facial expression detection is presently a study subject that attracts a lot of awards, necessitating a deep understanding of the trends. In this study, we were given the opportunity a Conventional Neural Network for facial expression identification. By default, this solution is using a straightforward 4-layer CNN Model to recognize emotions on all faces in the webcam broadcast.

Keywords: *Convolutional Neural Networks (CNN), Open CV, Deep Neural Networks (DNN), Pre-trained models.*

1. Introduction

The process of face detection involves locating the presence of a face in a still image or a video. The face in the scene can be tracked using primary facial attributes like eyes, lips, forehead, and chin aligned in the order of a familiar face. The fundamental idea behind face detection is that a human eye can accurately detect objects that a machine cannot. From the perspective of a machine, it seems exactly like a person using his senses to look for something. The factors, which make the face detection task difficult to solve are variations in the image plane and pose, incomplete face in the frame and other structural components, lighting conditions, and background.

Expression Recognition

Face expression has been a significant study area for the past two decades, in the fields of image processing and computer vision. In these sectors, significant advancements have been made, and numerous strategies have been put forth. However, Figure 1 below illustrates a fully automated hand gesture recognition system in its typical state. Each emotion has its facial muscle movements that trigger the specific emotion which can be used to differentiate one expression from another.

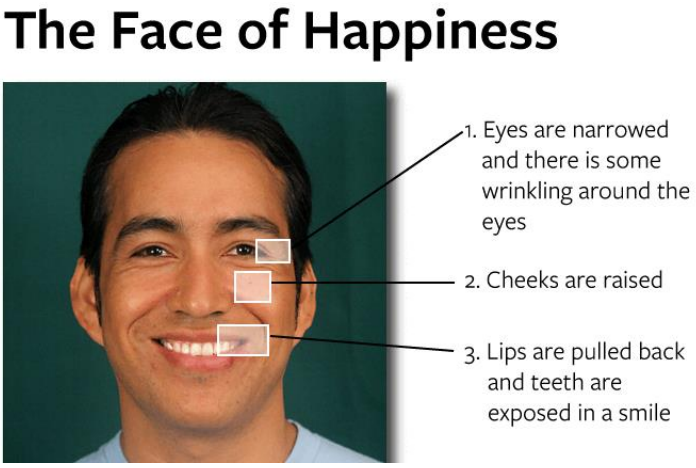


Fig 1 Image of Expression Recognition

Experimentation:

They used a customized CNN model with six convolutional layers, "Rectified Linear Activation Function" (ReLU) three max-poolings are used as the activation function, with the first two using pool size (3,3) and stride (2,2) and the third using pool size (2,2) and stride (2,2). Each max pooling is preceded by every two convolutional layers, and two drop out with a value of 0.2. two dense layers, one layer that has been flattened one of which has the activating function "RELU," and the other of which has the activating function "Softmax" and 1.2 million parameters that can be trained. With no pre-processing or feature extraction methods required, the six CNNs layer architecture used in this work performed well, achieving an accuracy of 61.7 percent on the FER2013 test. The modern test's ensemble of CNNs achieved a 75.2% accuracy rate for recognizing the seven different emotion categories. We carried out numerous trials with different batch sizes and epochs. However, after many experiments with after experimenting with various batch sizes and epochs, the optimum test accuracy was discovered with a batch size of 512 and 10 epochs.

Fine-tuned Visual Geometry Group (VGG-16) to recognize emotions on FDR 2013 as images:

The researchers used the FER-2013 dataset as the raw data to describe their experiments in this publication. The researchers have also compiled a large number of papers that used FER-2013 and a variety of approaches, including ensemble-based neural network (EBNN) and standalone-based neural network (SBNN) techniques.The suggested network belongs to the SBNN class, with VGG-16 serving as the basic model. GAP was one of thirteen convolutional layers in the final pooling layer. The network is then subjected to an experiment in which variables like as data distribution, batch normalisation usage or nonuse, GAP, optimizer selection, and freezing a specific number of my layers are changed. According to the research team's testing of 23 different models, the use of an imbalanced dataset, GAP, non-frozen layer, and SGD optimizer leads in the highest overall accuracy, or 69.40 percent. With this result, the network significantly outperformed the majority of reported networks, which coincides with the present results.Since it allows end-to-end training and is significantly simpler than the other three top SBNN models, the model takes less time and memory.

By using Machine Learning methods the Facial Expression Recognition of FER2013:

Multiple obstacles prevented the FER2013 dataset from learning an effective model. To address the problems of FER2013, we want to create a data augmentation method in this paper. In addition, we contrast SVM, AlexNet, VGG, and ResNet, four alternative deep learning techniques. The trials on FER2013 demonstrate that the VGG16 is the ideal model for recognizing of the face emotions, and FER2013 performance is significantly enhanced by our data augmentation. The majority of real-world applications can benefit from our method's in-depth classification study, which is offered. We find that facial expression recognition is more accurate with the VGG16 model. The accuracy rates are forSVG+HOG system as 46.0 percent, AlexNet+SVM system as 52.7 percent, VGG16 system 67.0 percent, and for ResNet18, 58.6 percent.

By Using a Gabor Convolutional Network Efficient and Fast Facial Expression Recognition:

Multiple obstacles prevented the FER2013 dataset from learning an effective model. To solve the challenges of FER2013, we want to create a data augmentation approach in this study. In addition, we contrast SVM, AlexNet, VGG, and ResNet, four alternative deep learning techniques. The trials on FER2013 demonstrate that the VGG16 is the right model for detecting the expressions of face, and our data augment significantly and also improves it in FER2013. Our approach offers in-depth classification research that is broadly applicable to the majority of real-world applications. We discover that the VGG16 model works better for identifying facial expressions.These resulted as SVM + HOG 46.0%, AlexNet+SVM 52.7%, VGG16 67.0%, and ResNet18 58.6% which are the final accuracy rates.

1. **Using alexnet transfer learning the Facial Emotion Recognition:**

They used both the FER2013 and CK+ datasets in their paper. To determine whether or not picture augmentation helps to increase accuracy, they applied it. Because of the image enhancement, CK+'s accuracy increased from 96.35 to 99.44. From 70.52 to 66.20, it decreased FER2013. For the proposed model's evaluation, Adam and Adam, two different optimizers, were employed. On both datasets, the model trained with Adam was less efficient than the fine-tuned Alexnet network trained with the Adam optimizer. The adaptive learning rate can be turned on or off in Adam depending on the gradient variance, which allows it to perform better than Adam. The warm-up learning rate no longer needs to be manually adjusted thanks to Adam. In this paper, we suggest using FER to achieve Alexis's transferrable learning. The proposed model employs the complete framework. In order to learn features particular to emotion classes, the final layer of the pre-trained Alex net is substituted with a fully connected layer. Without freezing any layers, the entire model is refined using emotion datasets. Image scaling and grayscale conversion are the only basic pre-processing steps needed for this technique. Additionally, Fine-tuned Alexnet doesn't need any extra manual feature extractions. FER2013 and CK+ datasets are used to assess the suggested model. By obtaining accuracy of 99.44 percent and 70.52 percent for the CK+ dataset and the FER dataset, respectively, Fine-tuned Alexnet surpassed other current techniques. Due to its fine-tuning method, the suggested model can even perform well with a tiny dataset like CK+.

PROPOSED WORK:

The input and output are so instantaneous that they both occur in a matter of milliseconds. Additionally, the input and output are so human-computer interactive that no significant button pushes or output instructions are needed to display the name of the output expression or to provide the input image. The headings below outline the entire work, methodologies, models, etc.

Working Principle:

This research tries to classify using deep convolutional neural networks, it is possible to classify a person's facial expression into one of seven groups. The FER-2013 dataset, which was released at the International Conference on Machine Learning, is the platform after which only one is known about it is as used to train the model (ICML). The faces in this collection represent the seven emotions of anger, contempt, fear, happiness, neutrality, sadness, and surprise. There are 35887 grayscale, 48x48 pixel photos in this collection.This project works more in a human-computer interactive way, which means whenever we hit the display command then automatically the webcam turns on and starts detecting faces and opens a live video which automatically boxes the faces that are identified and display the expression name above the box and this is as simple as that. The model doesn't bother the moving face or a slightly incomplete face and even detects the emotion as long as the face stays in the frame. The input and output are both done simultaneously and in real-time with no latency.

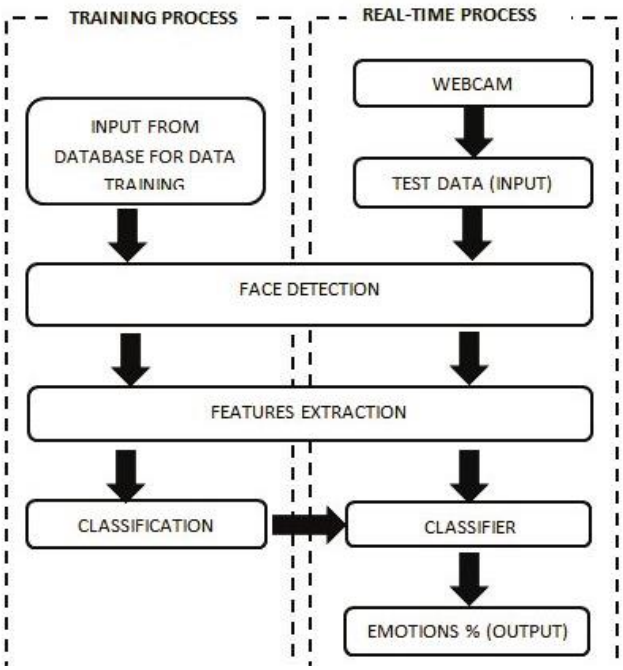


Fig2: Proposed Model

About Images Dataset:

Data is a crucial component of any AI model and, essentially, the only cause of the current rise in machine learning's popularity. Scalable ML algorithms are now feasible as standalone solutions that can add value to a business rather than being a by-product of its core operations because of the availability of data.

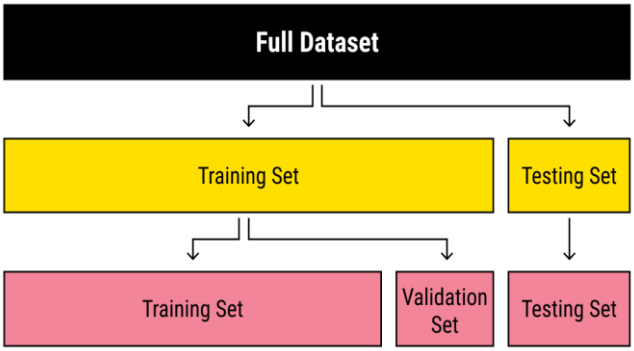


Fig3: Dataset

Training:

A machine learning model's parameter (weights) must be fitted before it can be trained using examples, training data is a collection of samples (such as a collection of images or videos, texts, audio files, etc.) with comprehensive and relevant labels (classes or tags). This means that any ML model must include training datasets. According to the objectives of an AI project, they are required to instruct the algorithm on how to create precise predictions. Machines need them to start recognizing patterns in the data, much like humans learn better from examples. Computers, on the other hand, require a lot more examples because they do not think similarly to humans.

They are not as good at seeing objects in pictures or identifying persons as humans are at doing so. They communicate using unique, differently structured programming languages. For training a machine learning model to identify emotions from videos, they ask for a lot of labor and data.

One image will suffice to explain what a cat is to a young child. To accurately distinguish a cat from, for example, a dog, a computer must be taught to recognize thousands of photos of various cats in all sizes, colors, and forms. On the other hand, a sufficiently advanced ML model can produce results that are more accurate than those produced by a human. This may seem paradoxical, but it also has to do with the ways that humans and machines interpret information differently.

Validation:

The collection of data is used to assess a model's fit to modify model hyper-parameters while using the training dataset. The evaluation becomes more skewed when skill from the validation dataset is added to the model setup.

Testing:

The dataset collection is a portion of the training dataset used to objectively assess a model's performance. There are additional techniques for calculating an impartial or biased assessment of model competence on unknown data depending on the validation dataset.

FER2013 Dataset:

A 48x48 pixel grayscale image of a face makes up the FER2013 data. As a result of the faces being automatically registered, each image now has a face that is roughly centered and takes up about the same amount of space. Each face must be assigned to one of seven categories,

with denotion of anger, disgust, fear, happiness, sadness, surprise, and neutrality. "Emotion" and "Pixels" are two columns in the train CSV file. The expression of the emotion that is depicted as the image is by a numeric code in the "emotion" column that ranges from 0 to 6, inclusive. For each image, a string enclosed in quotes is present in the "pixels" column. The values in this string are separated by spaces and are arranged in row-major order. You must forecast the emotion column-based exclusively on the "pixels" column in the test CSV.
28,709 instances make up the training set. 3,589 different samples make up the 3,589 sample public test set utilized for the leader board. Additional 3,589 cases are included in the final test set, and they were utilized to choose the contest winner.
This dataset of the experiment in the present study was as Pierre-Luc Carrier and Aaron Courville(2013) which was a part of an ongoing research project. The workshop organizers have graciously received a draught of their dataset, which will be utilized in this competition.

Fig4: Samples from dataset

Open CV:
OpenCV is a frequently employed technology in computer vision. It is a Windows, Linux, and Mac operating system compatible real-time computer vision library that was written in C and C++. It is available as open-source software for free download at http://sourceforge.net/projects/opencvlibrary/. The traditional conception of a digital image is that it is composed of discrete groups of light intensities that have been organized into a two-dimensional matrix of picture elements, or pixels, by a device like a camera. These pixels can all be represented numerically and stored using a particular file format (such as jpg or gif) [8].

- A picture can be represented by Open CV in more ways than just as a pixel array. It does so by presenting an image as a data structure known as a pill image, which makes useful image information or fields like:
- A width is a number that indicates the width of a picture in pixels. Height is a number that indicates the height of the image in pixels.
- The amount of colors per pixel is indicated by an integer in the data, which is an array of pixel values pointed to a pointer channel.
- Depth is a number that represents as bits per pixel.
- With step, an integer representing the number of bytes per image row is used. An integer represents the size of an image in bytes.
- ROI is a pointer as a structure that designates the focus area inside the picture.

Functions used in Gesture Recognition:

CV2.Video Capture

It is used to obtain a video capture object for the camera, and the read () method is used to read the frames while using the newly formed video capture object in an infinite loop. Either the device index or the device name used to capture the frames may be used as its parameter. It provides a picture with 640*480 pixels by default.

CV2.Rectangle

CV2. Rectangle (image, start point, end point, color, thickness) is the basic syntax.
Any image that was formed can have a rectangle drawn on it. When the ROI is found, a rectangle is drawn across the hand movements that were found.

CV2.COLOR

A picture is transformed from one color space to another using it. In OpenCV, there are more than 150 techniques for converting color spaces. CV2.CVT color syntax is used (src, code, test, design). Src stands for the source image, code for the color space conversion code, DST for the output picture's size and depth, and design for the number of channels.

CV2.Gaussian Blur

It is used to smooth the image using Gaussian Blur. Using this method, the sharp edges in an image are smoothed. The syntax for applying the Gaussian Blur function is cv2.GaussianBlur(Input_array, Output_array, size, border type).

CV2.Threshold

The thresholding is applied using this function. The first argument is the source image, which must be in grayscale. The second argument is the threshold used to classify pixel values, and the third argument specifies the maximum value assigned to pixel values that surpass the threshold. We have implemented both Binary Inversion and OTSU thresholding techniques in our model. There are many types of thresholding techniques.

Deep Learning Model

In this Section First, we discuss Convolutional Neural Network which performs both Feature Extraction and Classification, and our own CNN Model that we use for Feature extraction.

Convolutional Neural Network

Typically used to evaluate visual pictures by processing data with a grid-like architecture, convolutional neural networks are feed-forward neural networks. A CNN is another name for it. An image's objects can be found and classified using a convolutional neural network. In CNN, every image is displayed as an array of pixel values.

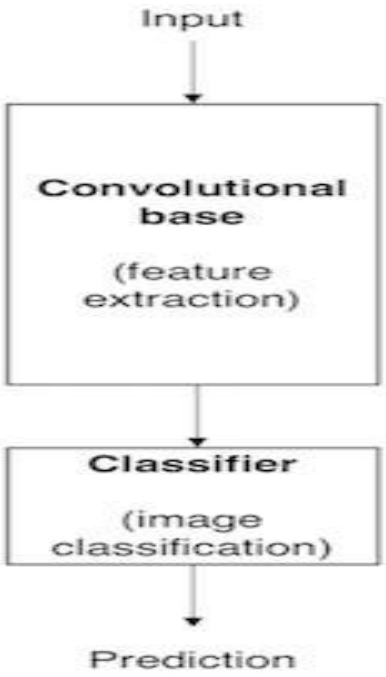


Fig5: Architecture of a model based on CNN.

Feature Extraction

During this phase, we extracted the most significant features of the face using our model. Were shaped the faces in the video according to the fixed input dimensions, accepted by the model which is trained using our own CNN model. After passing images through our model, the output of the first and second fully connected layers (fc1 & fc2) as features. For each input image, we got a feature vector of 1024 dimensions from those two layers. The features extracted from our model are proved to be faster than the hand-crafted features as these models can represent the images efficiently.

Model Training and Evaluation

During this phase, we train the model with the FER2013 training dataset for comparative studies. We apply the trained model to test data and evaluate its performance. We use a Deep Neural Network of 7 hidden layers each with 32,64,128, and 128 units respectively, and ReLU activation. Categorical Crossentropy loss with adam optimizer is preferred for these experiments. To prevent the model from over-fitting, we introduced a 0.5 dropout after the input layer. This made the model more reliable. For training and validation, we used the FER2013 training folder and test folder, respectively. The architecture of the suggested method for model training and evaluation is shown

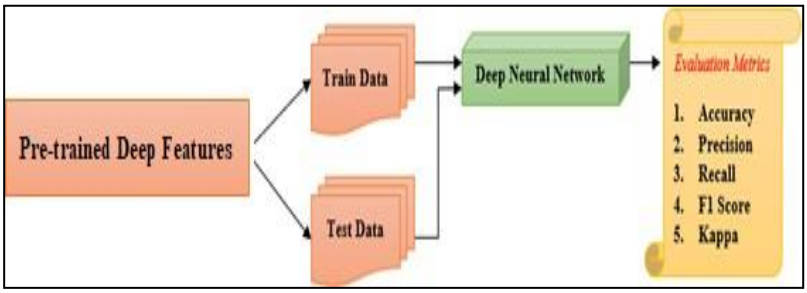


Fig6: Model Training and evaluation

RESULTS:

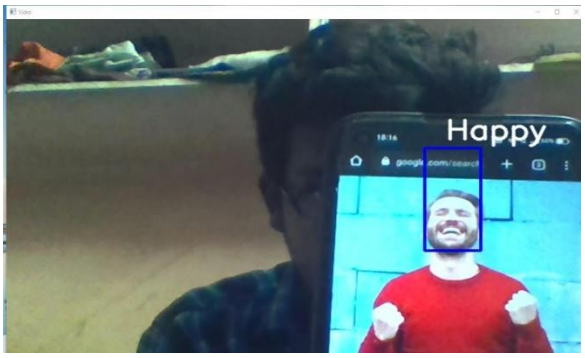
To get the fast and best outcome of the CNN features, we have used the benchmark dataset available on Kaggle.com. We collected the data from FER2013 with CNN Challenge on the same website. The dataset is a large set of black and white images of distinct faces under a variety of imaging conditions. The dataset consists of nearly 30000 samples.

Table1: Models Comparison

Model	Conv Layers	Training		Testing		Epochs
		Accuracy	Loss	Accuracy	Loss	
CNN	8	79.50	0.0232	77.08	0.1149	50
CNN	5	78.35	0.0941	76.88	0.1066	50

Outputs:

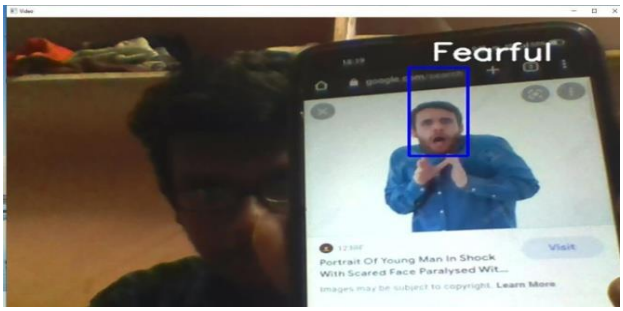
Happy:



Sad:

Angry:

Fearful:



REFERENCES:

1. Breuer, P., Eckes, C. and S Muller, (2007). Hand gesture recognition with a novel IR time-of-flight range camera - a pilot study, Proceedings, 28-30 March, Springer-Verlag pp 247-60.

2. Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, Mo Nours Arab, (2008). "Human-Computer Interaction: Overview on State of the Art", International Journal on Smart Sensing and Intelligent Systems, Vol. 1(1).

3. Hervé Lahamy and Derek Litchi, "REAL-TIME HAND GESTURE RECOGNITION USING RANGE CAMERAS".

4. Mokhtar M. Hasan, Pramoud K. Misra, (2011). "Brightness Factor Matching for Gesture Recognition System Using Scaled Normalization", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3(2).

5. Nguyen T.-N. , Huynh H.-H. , and J. Meunier, "Static hand gesture recognition using artificial neural network," Journal of Image and Graphics, vol. 1, no. 1, pp. 34–38, 2013.

6. Pei Xu, "A Real-time Hand Gesture Recognition and Human-Computer Interaction System".

7. Stergiopoulou E and N. Papamarkos, (2009)."Hand gesture recognition using a neural network shape fitting technique," Engineering Applications of Artificial Intelligence, vol. 22, no. 8, pp. 1141–1158.