Real Time Facial Expression Recognition Based On Deep Neural Network

T. Ambikadevi Amma[1], M. R. Sruthy[2], S. Divya[3], P. Renuka[4*]

[1]*Professor & Principal, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady, India*

[2,3]*Assistant Professor, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady, India*

[4]*PG Scholar, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady, India*

*Corresponding author: renuka15297@gmail.com

***Abstract*: Now-a-days with the continued development of artificial intelligence facial emotion recognition has become more popular. The emotion recognition plays a major role in interaction technology. In interaction technology the verbal components only play a one third of communication and the non-verbal components plays a two third of communication. A facial emotion recognition (FER) method is used for detecting facial expressions. Facial expression plays a major role in expressing what a person feels and it expresses inner feeling and his or her mental situation or human perspective. This paper aims to identify basic human emotions with the combination of gender classification and age estimation. The facial emotions such as happy, sad, angry, fear, surprised, neutral emotions are considered as basic emotions. Here proposes a real time facial emotion recognition system based on You Look Only Once (YOLO) version 2 architecture and a squeezenet architecture. The yolo architecture is a real time object detection system. Here it used for identify and detect faces in real time. These images are captured by using anchor boxes for accuracy. The second architecture is squeezenet and is used for gender classification and age estimation. It provides significant, accurate object detection and extracts high-level features that help to achieve tremendous performance to classify the image and detecting objects. Both the architectures provide accurate result than other methods with the large no of hidden layers and cross validation in the neural network.**

***Keywords*: Artificial Intelligence (AI), Convolutional Neural Network (CNN), Emotion recognition, Facial expression recognition (FER), YOLOv2.**

1. Introduction

Facial expression recognition technology uses biometric markers to identify and predict human emotions in the human face. It is a sentiment analysis tool that can be able to detect the six basic emotions such as happy, sadness, surprise, fear, disgust and anger. Facial expression recognition or computer-based facial expression recognition system is important because of its ability to mimic human coding skills. Facial expressions and other gestures convey nonverbal communication cues that play an important role in interpersonal relations. These cues help the listener to understand the intended meaning of the sentences. Therefore, facial expression recognition, because it extracts and analyzes information from an image or video feed, it is able to deliver unfiltered, unbiased emotional responses as data. Similarly, artificial intelligence voice recognition technology using the sense of hearing and AI speakers has been commercialized because of improvements in artificial intelligence (AI) technology. Through the use of such technologies that recognize voice and language, there are artificial intelligence robots that can interact closely with real life. By using these information’s the robots or the other systems can manage the daily schedules of people and playing their favorite music. Research is a process of arriving at an appropriate solution to a problem through a systematic approach. Technologies for communication have traditionally been developed based on the senses that play a major role in human interaction. In particular, artificial intelligence voice recognition technology using the sense of hearing and AI speakers has been commercialized because of improvements in artificial intelligence (AI) technology. Through the use of such technologies that recognize voice and language, there are artificial intelligence robots that can interact closely with real life, in such ways as managing the daily schedules of people and playing their favorite music. However, sensory acceptance is required for interactions more precisely mirroring those of humans. Therefore, the most necessary technology is a vision sensor, as vision is a large part of human perception in most interactions. In artificial intelligence robots using interactions between a human and a machine, human faces provide important information as a clue to under-stand the current state of the user. Therefore, the field of facial expression recognition has been studied extensively over the last ten years

2. Comparative Study

C. A. Corneanu [1] proposed a system based on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition. It is an automatic facial expression analysis and

RGB, 3D, thermal and multimodal methods are used. It is an automatic facial expression analysis method. Here RGB, 3D, termal and multimodel methods are used for face detection. Describes the facial muscles with the jaw or tongue derived from analysis of facial anatomy. The second literature study proposed by P. viola[2]. It is a rapid object detection using a boosted cascade of simple features. This is a machine learning approach fir visual object detection and uses integral images for feature selection. This method capable for processing images extremely rapidly and achieving high detection rates. Provides fast computation than the traditional recognition methods. Select small number of visual features from the input image and corresponding output image with recognized emotion retrieve as output.

P. I. Wilson [3] proposed a Facial feature detection using haar classifiers system. It is a method to accurately and rapidly detect faces within an image. This technique can be adapted to accurately detect facial features. The haar cascade method detects and analyzes various elements of human face. By regionalizing the detected area of the image, false positives are eliminated. Hence the speed of detection is increased due to the reduction of the area examined. The disadvantage with this paper is that it only detect the facial features like eyebrow and the mouth. Hence the result is not accurate than the proposed method.

Q. Zhu and M.C. Yeh proposed fast human detection using a cascade of histograms of oriented gradients which is mainly uses cascade of rejectors approach. It mainly consists of two methods. They are cascadeof rejectors approach with the Histograms of Oriented Gradients (HoG) features [4]. These methods help to achieve a fast and accurate human detection system. The main features used in the system are HoGs of variable size blocks. These blocks are capture from human faces. Use a AdaBoost technique for feature selection and identify the appropriate set of blocks, from a large set of possible blocks. Comparatively this is not a fast and accurate human detection system than the other algorithms.

R. Rivera [5] proposed a system local binary patterns and its application to facial image analysis. It uses a LDTP method for facial image analysis and prediction. It efficiently encodes the information's of emotion related features. Here it divides the face image in to several regions and sample the codes. That code is used for compare the input images with the training set data. A two level grid is used to construct the face descriptor. As a typical application of the LBP approach, the LBP based facial image analysis is successful than previously used approaches.

B. Yang [6] proposed facial expression recognition using weighted mixture deep neural network based on double-channel facial images. In this method it automatically extracts the facial features from the input image. When compared to other methods it is more effective with the facial emotion recognition tasks. Uses two channels for image recognition. They are gray scale images and corresponding local binary pattern facial images. Features of LBP facial images are extracted by a shallow convolutional neural network (CNN). S. Xie and H. Hu [7] proposed Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. Here propose a novel method, named Deep Comprehensive Multi-patches Aggregation Convolutional Neural Networks (DCMA-CNNs). This is used to solve facial emotion recognition problem. The proposed method consists of two branches of convolutional neural network. The one branch that extracts local features from image patches and the other branch extract holistic features from the whole expressional image. Here aggregate the both local and holistic features before making the classification and result.

K. Zhang [8] Facial expression recognition based on deep evolutional spatial-temporal networks. Here propose a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN). It analyzes the facial expression information of temporal sequences which uses both recognition and verification signals as supervision. It also calculates the loss functions and it helpful to increase the variations of different expressions. PHRNN model is effective to extract temporal features based on facial landmarks from consecutive frames.

Y. Liu and X. Ma [9] proposed a method Facial expression recognition with pca and lbp features extracting from active facial patches. The paper proposes an algorithm based on the combination of two features. They are gray pixel value and Local Binary Patterns (LBP) features. Principal component analysis is used to reduce dimensions of the features. They are combined by the gray pixel value and Local Binary Patterns (LBP) features. All the features are extracted from the active facial patches of input image. A softmax regression classifier is used to classify the six basic facial expressions.

H. Jung [10] proposed a Joint fine-tuning in deep neural networks for facial expression recognition. Here the deep network is based on two different models. It extracts temporal appearance features and temporal geometry features. The temporal features extracts from the image sequences and the temporal features from the landmark points. Finally, these two models are combined using a new integration method and the corresponding result is considered as output.

3. Methodology

Emotion recognition is the process of identifying human emotion, and can be from facial expressions as well as from verbal expressions. The facial emotion recognition is both something that humans do automatically. Here the system that identify the six basic emotions with the combination of gender classification and age estimation. Mainly two algorithms are used for detection and classification. They are YOLO version 2 architecture and squeezenet architecture.

A. Data set

Kaggle dataset has been used for the experiment. It consists pre-cropped gray scale images with size of 224x224. The images are labeled in 7 emotions (0 = Anger, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral).

B. *Preprocessing*

In general scenario, human vision system, first detects the faces, and then subsequently it recognizes the emotion associated with that face. In the same way, in this work, face detection is the preprocessing or prior work of the emotion recognition task.

Bounding box is cropped and reshaped into 224X224 pixels. All the images are of frontal face. Non frontal faces (image of side face) and non-relevant images (images that were some random image or those with hands covering face, etc.) were removed. The preprocessing method that normalizes the data hence it boost the performance of CNN and achieve better accuracy.

C. *Training phase*

Fig. 1. Training of Data

The figure 1 shows the training of datasets. The deep neural network uses supervised learning approach hence the system trained with large no of datasets. The data base contains seven standard categories of emotions two sets of gender classification and age estimation categories.

D. *Training phase*

Fig. 2. Testing of data

Figure 2 shows the testing of data. The system tested with new images. Like training phase, in the testing phase, feature extraction is completed using proposed convolutional neural network.

E. *CNN-based FER approaches*

Facial expressions are the natural way to communicate emotional states. In recent years with the development of hardware technology, many algorithms based on deep learning have been researched. In this paper propose a method for facial expression recognition based on features extracted with convolutional neural networks (CNN). Automatic facial expression recognition (FER) has been studied due to its practical importance in many human behavior analysis tasks such as interviews, health care industry for patient monitoring, autonomous driving, and medical treatment, among others. Figure 3 shows the procedure of convolutional neural network facial emotion recognition approach. Facial emotion recognition (FER) is an important topic in the fields of artificial intelligence and computer vision technology. The CNN consist of three layers input layer, multiple of hidden layers and a output layer. Additionally, in the hidden layer there are convolutions, pooling and fully connected layers used in the architecture of CNN. Initially the input images passed through the convolution layer are convolved using filters in the convolution layers. From the convolution results, feature maps are constructed and maxpooling (subsampling) layers lower the spatial resolution of the given feature maps. CNNs apply fully connected neural-network layers behind the convolutional layers.

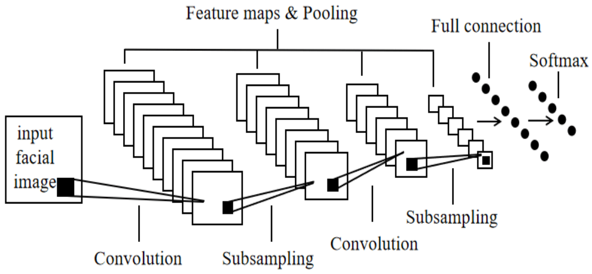


Fig. 3. CNN based FER approach

F. *Proposed approach*

Face detection and emotion recognition is one of the important tasks of object detection. Here introduce a new method with the combination of two architecture such as yolo v2 and a squeezenet architecture. There are many models are under development such as RCNN, RetinaNet, and YOLO. Here we use You look only once algorithm for real time face detection. It also helps to detect multiple faces at real time. This recognized image the passes through both the CNN architectures for classification.

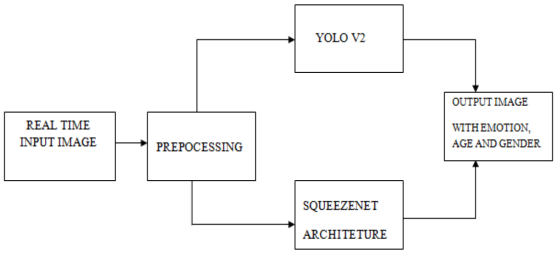


Fig. 4. Proposed system architecture

Here the figure 4 shows the proposed system architecture. The web cam captures the real time facial included videos. From the video stream system detect the faces and taken as the input. Similarly, all the faces that is present in the web cam taken as input and corresponding values given as output. The detected images undergo preprocessing states and here it transforms in to 224x224 size gray scale images for comparing and testing with the trained dataset. Similarly, the same input image passes through the squeezenet architecture and age and gender are classified. Finally, the results of two sections are combined together. Hence in the screen we can see each faces in the anchor boxes with recognized facial emotion, estimated age and gender classification.

1) YOLO v2

The deep learning based algorithms in object detection have grown rapidly. These algorithms are mainly classified in to two. They are two-stage detector like Faster R-CNN and one-stage detector like YOLO. It is a real time object detection algorithm with high speed i.e., better than real time. In this method the network understands generalized object representation. This is the second version of yolo, faster version architecture and can be freely available. YOLO performs well when facing normal size objects. The accuracy decreases notably when dealing with objects that have large-scale changing like faces. It mainly focused on improving recall and localization and also in maintaining classification accuracy. To achieve better performance, they use some ideas. They are Batch Normalization, High resolution classifier and a convolutional with anchor boxes. In the batch normalization method, it normalizes the image into corresponding gray scale image size. In the second stage they train the classifier network at 224x224 pixels. Increase the resolution to 448 for detection. The present approach includes using anchor boxes more appropriate for face detection and a more precise regression loss function with a multi object prediction per grid cell. The improved detector significantly increased accuracy while remaining fast detection speed. These are the post-processing steps needed to get final bounding boxes after the image is passed through Yolo network. Each Convolution block has the Batch Norm normalization and then Leaky Relu activation except for the last Convolution block.

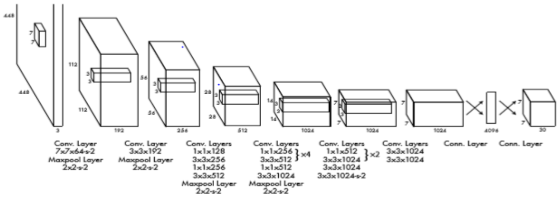


Fig. 5. YOLOv2 architecture

Figure 5 shows the architecture for the YOLOv2. The network architecture of YOLO consists of a new classification model named Darknet-19. It is used as a backbone for YOLOv2. These networks provide more accuracy and complexity. The darknet-19 consists of two layer. 19 convolutional layers and a set of 5 maxpooling layers. The advantage with this network is that, it provides 91.2% top-5 accuracy on imageNet, 90% better than VGG and 88% better than YOLO network. The reorganization layer takes every alternate pixel and puts that into a different channel.

In the training phase maily three stages take place. The classification stage train the darknet-19 network imagenet 1000 class classification dataset with input size 224x224. It gives the top-1 accuracy of 76.5% and a top-5 accuracy of 93.3%. In the detection phase removed the last Removed the last convolutional layer from Darknet-19 and add three 3 × 3 convolutional layers and a set of1x1 convolutional layer the last stage is Multi-Scale Training. Here the model trained the model for different input sizes.

2) Squeezenet

The squeezenet is an 18-layer network consist of 1x1 and 3x3 convolutions. Additionally, a 3x3 maxpooling layer is included. The major component in this architecture is fire module. It consists of two layers they are Squeeze layer and an expand layer. It reduces the total no of weight. Figure 6 shows the squeezenet architecture. This architecture stars with a single convolutional layer followed by fire modules with alternate MaxPooling 2D. The network fuses the convolution and maxpool layers and written back the output. Decrease the stride with later convolution layers and thus creating a larger activation/feature map later in the network.

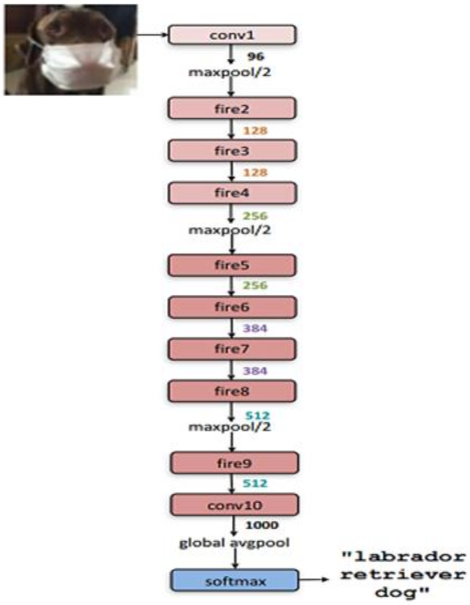
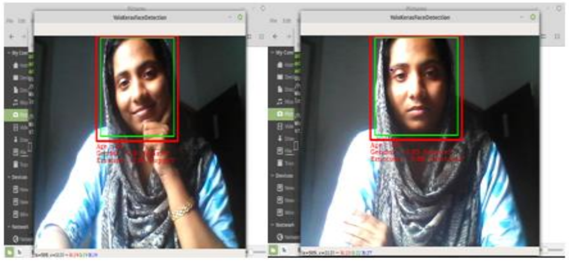


Fig. 6. Squeezenet architecture

4. Result and Discussion

In this experiment, the proposed architectures provide the accurate output than the existing facial emotion recognition methods. It finds out the faces in real time and predicts the facial expression with age and gender classification. Here the yolo architecture that provides more accurate result with improved accuracy and precision rate. The sueezenet architecture predicts the corresponding age limit and gender based on the training datasets.



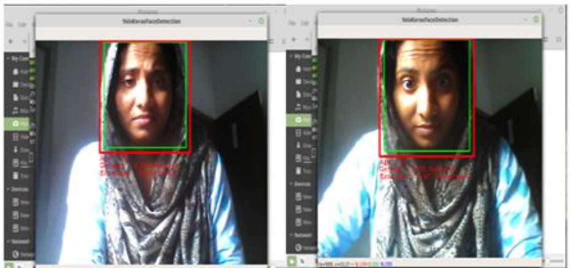


Fig. 7. Recognized emotions of female

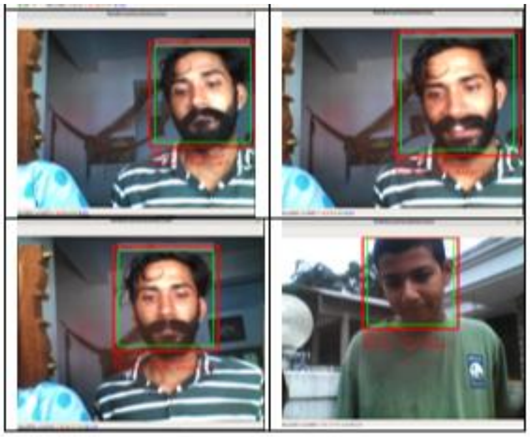


Fig. 8. Recognized emotions of male

5. Conclusion and Future Scope

The use of machines in society has increased widely in the last decades. Nowadays, machines are used in many different industries. As their exposure with human's increase, the interaction also has to become smoother and more natural. In order to achieve this, machines have to be provided with a capability that let them understand the surrounding environment. Specially, the intentions of a human being. Emotion recognition is still a difficult and a complex problem in computer science because every expression is a mix of emotions. Here proposed an efficient real time facial expression recognition system with the combination of two algorithms such as yolo version 2 and squeezenet architecture based on deep neural networks for more accurate and efficient facial expression recognition. The future scope can be an action that is done upon recognition of the emotions. If get a sad emotion, can have the systems plays a song or tells a joke or send his/her best friend a message. This can be the next step of AI where the system can understand, comprehend the user's feelings and emotions and react accordingly. This bridges the gap between machines and humans. We can also have an interactive keyboard where the users can just use the app and the app will then identify the emotion and convert that emotion to the emoticon of choice.

References

[1] C. A. Corneanu, M. O. Simon,´ J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 8, pp. 1548–1568, 2018.

[2] P. I. Wilson and J. Fernandez, "Facial feature detection using haar classifiers," Journal of Computing Sciences in Colleges, vol. 21, no. 4, pp. 127–133, 2017.

[3] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2. IEEE, 2016.

[4] Ryu, A. R. Rivera, J. Kim, and O. Chae, "Local directional ternary pattern for facial expression recognition," IEEE Transactions on Image Processing, vol. 26, no. 12, pp. 6006–6018, 2017.

[5] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mix-ture deep neural network based on double-channel facial images," IEEE Access, vol. 6, pp. 4630–4640, 2017.

[6] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 211–220, 2016.

[7] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Transactions on Image Processing, vol. 26, no. 9.

[8] Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma, "Facial expression recognition with pca and lbp features extracting from active facial patches," in 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR). IEEE, 2016, pp. 368–373.

[9] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–299.