Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 18 (2023) pp. 058-066
© MIR Labs, www.mirlabs.net/jias/index.html

Comprehensive Study on Emotion Detection with Facial Expression Images Using YOLO Models

Awais Shaikh[1], Keshav Mishra[2], Pradnya Kharade[3] and Mahendra Kanojia[4][0000-0002-7628-8683]

[1] Sheth L.U.J and Sir M.V College, Mumbai University,
Mumbai, Maharashtra, India
awaiscsfy2031@gmail.com

[2] Sheth L.U.J and Sir M.V College, Mumbai University,
Mumbai, Maharashtra, India
keshavcsfy2025@gmail.com

[3] Sheth L.U.J and Sir M.V College, Mumbai University,
Mumbai, Maharashtra, India
pradnyabhabal@gmail.com

[4] Sheth L.U.J and Sir M.V College, Mumbai University,
Mumbai, Maharashtra, India
kgkmahendra@gmail.com

***Abstract*: Facial expressions, a kind of nonverbal communication, can be used to interpret human emotions. A technology known as facial emotion recognition studies facial expressions in photos and movies. Due to the numerous applications, it has, emotion identification is a crucial subject. One of the most difficult pattern recognition challenges is emotion detection using facial expressions. Emotion detection using facial expressions includes a number of face-related applications, such as face verification, facial recognition, face clustering, and many more. In this article, we have given a comprehensive study of face and emotion detection using YOLO models. We have described the architecture of the YOLO model and its versions used for the review objectives. We have implemented various YOLO models and presented the experimental results of YOLOv5. We infer from our comparative study that YOLO models are explored for face detection but very little work has been found for expression detection. The initial hypothesis of this review was that there is an increase in accuracy with every new version of the YOLO model for face and emotion detection which turns out to be false.**

Keywords: Emotion Detection, Face Detection, Deep Learning, Convolution Neural Network, Pattern Recognition, YOLO.

I. Introduction

Recent developments in pattern recognition, machine learning, and biometrics analysis, along with the increased usage of cameras, are mainly responsible for the expansion of the FER (Facial emotion recognition) technology. For applications like surveillance, self-driving cars, and gaming that need quick and precise object recognition, YOLO is a popular option [1]. Emotion recognition has been added to the YOLO architecture recently, enabling it to categorize the emotional state of people in a picture or video. In order to do this, the YOLO model must first be trained on a sizable dataset of faces and emotions, such as the Affect- Net dataset, before it can be used to forecast the emotions of people in fresh photos or videos. Happiness, sorrow, wrath, surprise, and contempt are among the feelings that YOLO-based emotion detection systems frequently identify [2]. In several applications, including security, human-computer interaction, and other fields, YOLO-based emotion detection has been found to be relatively quick and accurate [3].

To research facial expressions, cameras are utilized to identify faces and capture real-time human responses to circumstances. The way that the facial muscles flex and contract differently in response to each facial expression makes it easier for deep learning algorithms to recognize emotion. It has been discovered that YOLO-based emotion recognition is reasonably quick and accurate, and it is employed in many applications including market research, security, and human-computer interface [4]. The quality of the training data, the size of the model, and the facial expressions of the people in the image are a few examples of variables that can have an impact on how well YOLO-based emotion detection systems perform. There are seven basic human emotions: surprise, contempt, rage, fear, happiness, and sadness. These emotions can be recognized by a range of facial expressions, such as the position of the mouth and the positioning of the eyes and brows [5].

This technology can be employed, and a variety of applications can be made, using the YOLO algorithm. Depending on their response, it may be clear if they are eager to talk to us or not. Face recognition can be used for a number of purposes, such as personal identification in surveillance and user authentication for security systems. Pattern recognition and computer vision have seen an increase in the use of automated face detection. Face identification systems formerly had limited application outside of simple circumstances, but deep learning methods have improved their versatility. Compared to YOLO, the first successful one-shot CNN architecture, the R-CNN structure was significantly different [6]. In order to train the YOLO network for emotion recognition, a dataset of images or videos that have been labeled with various emotions is used. The network gains the capacity to identify faces and ascertain the emotion behind each one. The final outputs from the network consist of a set of bounding boxes encircling the faces in the image together with the expected emotion for each face. The main advantage of YOLO for emotion identification is its real-time performance, which makes it suitable for use in applications that require quick and efficient emotion detection [7]. How well YOLO can identify emotions will depend on both the quality of the training data and the intricacy of the network. In the conclusion section, we tabulated the comparative analysis of the YOLO models in the proposed work. In this article, we have detailed the architecture of all variants of YOLO.

II. Literature Review

In this section we have described the well-recognized work done in facial and emotion detection using YOLO models. The section is organized in order of YOLO models and its implementations.

D Garg, P Goel et al in 2018 [8]. The application of YOLO, a real-time object detection system, to face detection was shown in this study. In order to conduct face identification on fresh photos, the authors trained YOLO using a library of images labeled with face-bounding boxes. They observed that YOLO performed well in terms of accuracy and speed when compared to two other well-known face identification systems, Multi-task Cascaded Convolutional Networks (MTCNN) and Single Shot Multi Box Detector (SSD). This research illustrates the potential of YOLO for additional computer vision problems and shows its efficacy for face detection. The experiment's findings offer insightful information about the variables that affect how well deep learning-based object detection systems function. In the year 2019 [9]. The authors employed the YOLOv3 object detection system for the investigation. In this study, the scientists tested YOLOv3's performance at identifying emotions after training it on a dataset of facial expressions. This study makes a case for using YOLOv3 for facial expression-based emotion identification and offers some preliminary findings on its effectiveness in this job. Z Lu, J Lu et al in 2019 [10] The research introduces a multi-object identification technique that combines the benefits of two well-known ResNet and YOLO deep learning-based object detection systems. Yolo is a quick and effective real-time object detection system; however, it is not as accurate as other methods. Deep residual networks like ResNet are capable of obtaining great accuracy, although they move more slowly than YOLO. To obtain high accuracy while preserving real-time performance, the authors suggest a hybrid network that combines the advantages of YOLO and ResNet. On the Microsoft Common Objects in Context (COCO) dataset, the authors trained, assessed, and compared their hybrid network with several cutting-edge object detection methods. They discovered that their hybrid network maintained real-time performance while outperforming YOLO in terms of accuracy. The outcomes of the studies show how successful the suggested hybrid network is and emphasize its potential for real-world use.

Another work in 2020 [11] a deep learning-based method for predicting students' attention in a classroom context based on their outward behavior is presented. In order to evaluate video footage of pupils in a classroom, the authors used YOLOv3, a cutting-edge real-time object detection technology. The student gaze direction and head detection of YOLOv3 were trained. The students' gaze direction was then utilized to determine how attentive they were. On a collection of classroom video footage, the authors assessed their method and contrasted YOLOv3's performance with that of two existing gaze estimation techniques. They discovered that YOLOv3 performed well in both accuracy and speed. Kuldeep Mehta, Ashitosh Bhige et al in 2020 [12]. Provided a deep learning-based method for facial recognition and detection in multi-object pictures. A deep neural network was employed by the authors to conduct facial identification after using YOLO, a real-time object detection method, to find faces in photos. A picture dataset featuring faces and other objects was used by the authors to train and test their system. Even in photos with numerous objects and other distracting features, the authors' algorithm was able to recognize faces and reliably detect faces. Additionally, they evaluated the effectiveness of their system against other cutting-edge approaches for facial detection and recognition and discovered that it was competitive in terms of accuracy and speed. Research proposed in 2022 [13]. A deep learning-based method for real-time object detection was presented in the paper, with a focus on detecting micro vehicle objects. On a collection of tiny vehicle objects, the authors improved the real-time object recognition system YOLO-V2. They assessed the system's efficiency and accuracy and contrasted it with other cutting-edge approaches to object detection. The authors discovered that their system successfully detected minute vehicle objects in real time and performed well in terms of accuracy and speed. They also talked about the difficulties in finding little objects and the limitations of their method. The research paper [14] in 2021, Offers a real-time object detection system called YOLOv3 as the foundation for an enhanced object detection technique for remote sensing photos. The remote sensing image features, such as large-scale changes, complicated backdrops, and small objects, were improved upon by the authors by modifying YOLOv3. They tested the effectiveness of their upgraded system against other cutting-edge object detection techniques using a collection of remote-sensing photos. The scientists discovered that, when compared to previous methods, their modified algorithm performed better in terms of accuracy and speed, especially for recognizing small objects. The difficulties of object detection in remote sensing images and the limitations of their strategy were also explored. S. P. Rajendran, L. Shine et al in 2019 [15]. The study describes a deep learning-based method for real-time object detection utilizing YOLOv3 for traffic sign recognition. On a dataset of photos of traffic signs, the scientists improved YOLOv3, and they employed it as a detector to identify traffic signs in real time. They assessed the system's efficiency and accuracy and contrasted it with other cutting-edge techniques for recognizing traffic signs. The authors discovered that their system successfully recognized traffic signs in real-time while achieving good accuracy and speed performance. The difficulties of reading traffic signs and the limitations of their strategy were also explored.

A research work in 2018 [16]. On a dataset of face images, the authors improved YOLO and utilized it to quickly identify faces. They assessed the system's performance in terms of accuracy and speed and contrasted it with other cutting-edge face identification techniques. The authors discovered that their system successfully detected faces in real time and performed well in terms of accuracy and speed. They also talked about the difficulties with facial detection and the restrictions on their strategy.
Wu, Lv , Jiang, et al in 2020 [17]. The channel pruning technique considerably increased the accuracy and speed of the YOLOv4 algorithm, according to the authors' evaluation of the algorithm's performance in this study using a dataset of photos of apple flowers. Based on the findings, the YOLOv4 algorithm with channel pruning has a high success rate in detecting apple blooms in their natural habitats, making it a promising tool for precision farming. The paper's main point is the potential of deep learning algorithms, such as YOLOv4, for use in agriculture and other industries. The study's findings show how channel pruning might enhance deep learning algorithms' performance and make them more useful for real-world applications.

In 2019 authors Tian, Y, Yang, et al [18]. In their study the results demonstrated that the enhanced YOLO-V3 model was faster than some of the other object identification models that were assessed in the study and were able to detect apples with good accuracy at various growth stages. The research emphasizes the promise of deep learning algorithms for agricultural item detection, particularly for orchard apple detection. The study's findings show how well the upgraded YOLO-V3 model performs in this application and imply that deep learning techniques may be able to increase the effectiveness and precision of fruit recognition in orchards. Another work published in 2020 [19]. The YOLO model's performance was assessed using a different test dataset after the authors trained it on a sizable collection of face photos that had been gender-labeled. The findings demonstrated the YOLO model's excellent gender recognition accuracy and speed, making it an attractive option for real-time applications. Additionally, the authors compared YOLO's performance to that of other cutting-edge techniques and discovered that it outperformed them in terms of accuracy and speed.

III. Image Classification and Localization.

The objective of image classification in computer vision is to give a label or class to an input image. To determine whether an emotion of a face is happy, sad or disgust, for instance, would be the task of emotion classification. To accomplish this, a deep learning model is trained using a sizable dataset of labeled photos, where each image belongs to a particular class. The appropriate class can subsequently be assigned to recently found photos using the trained model. On the other hand, the process of pinpointing an object's location within an image is known as image localization. Image localization seeks to identify the precise location of each object in the image, which is commonly represented as bounding boxes, as opposed to image classification, which merely aims to identify the class of objects present in an image [20]. Image localization is a crucial computer vision problem that is utilized in processes including object detection, segmentation, and tracking.

A. Challenges in Image classification and emotion detection

The difficult task of emotion identification in computer vision entails identifying human emotions from facial expressions or body language. However, there are a number of difficulties with YOLO-based emotion recognition. Several of these difficulties include:

1. *Emotional complexity:* Because emotions are so complex and multidimensional, it can be difficult to recognize and categorize them in visuals.
2. *Lack of annotated data:* When training emotion detection models, there is sometimes a dearth of annotated data, which results in models that are poorly generalized.
3. *Cultural and individual variations:* Because emotions can vary depending on culture and individual, it is challenging to create models that can be used by a wide range of people.
4. *Ambiguity in facial expressions:* Because facial expressions can represent several emotions at once, it might be difficult for models to recognize emotions.
5. *Environment dependence:* The context in which facial expressions and emotions are expressed can affect how those expressions are understood, making it challenging for algorithms to reliably predict face expressions and emotions.

When developing and refining a YOLO-based emotion detection system, it is crucial to carefully analyze and address these issues because they may have an impact on the effectiveness and precision of emotion detection models.

B. Stages in Emotion Detection.

The stages in emotion detection can be summarized as follows which are shown in the below Fig.1.

1. *Face detection:* The first step in emotion detection is to detect the face in the image. This can be done using a variety of techniques, including Haar cascades, HOG features, or deep learning-based methods.
2. *Face alignment:* The next step is to align the detected face so that it is consistently positioned and sized, regardless of the position and orientation of the face in the original image.

3. *Feature extraction:* After face alignment, the next step is to extract relevant features from the aligned face that can be used to detect emotions. This may involve extracting features such as the shape of the face, the appearance of the eyes, or the texture of the skin.
4. *Emotion classification:* The extracted features are then used to classify the emotion present in the face. This can be done using a variety of machine learning algorithms, such as decision trees, support vector machines, or deep neural networks.
5. *Model evaluation:* Finally, the performance of the emotion detection model is evaluated on a separate dataset to assess its accuracy and robustness. The model may be fine-tuned and improved based on the results of the evaluation.

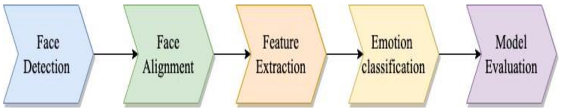


Figure 1. Emotion detection stages.

These steps can be used to create an accurate emotion detection system that can discern emotions from a variety of sources by carefully choosing the right algorithms and methodologies for each stage.

IV. Research Methodology.

A comprehensive description of various YOLO models is covered in this section. The model description includes the architectural updates in YOLO series models with their merits and limitations. The focus of this section is to describe the use of YOLO models for face and emotion detection.

A. *YOLO Overview*

The object-detection field was dominated by YOLO, which swiftly ascended to become the most extensively used algorithm because of its speed, precision, and learning ability. The modern deep learning framework for real-time object detection is called YOLO. It is an improved model than the region-based detector and outperformed standard detection datasets like PASCAL VOC and COCO datasets. This model can operate at various resolutions, providing quick and accurate results. The photos can be enlarged to a random scale to enhance performance toward scale invariant. The YOLO frameworks for detection are getting faster and more accurate with the help of neural networks. YOLO is also computationally more effective than previous two-stage object identification systems because it performs object classification and bounding box regression using a single convolutional neural network (CNN). Figure 2 below displays the bounding box predictions.

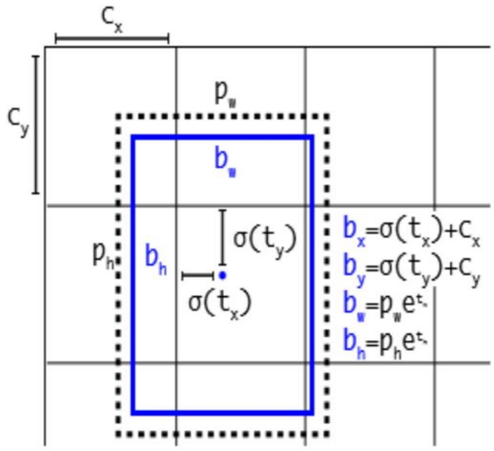


Figure 2. Architecture for bounding box prediction Illustration adopted from [19].

The key components of YOLO are as follows: Performance in real-time: YOLO is built to be quick, enabling real-time object detection on standard hardware. This makes it appropriate for a variety of applications, including robots, self-driving cars, and video surveillance. Single-shot detection: YOLO detects objects in a single forward pass of the network, as opposed to conventional object detection systems that employ a two-stage process. Because of this, YOLO is computationally faster and more effective than its two-stage competitors. End-to-end training: YOLO is trained from the beginning to end, enabling it to learn bounding box regression and object classification in the same network.[21] Compared to systems that demand separate training phases for object classification and bounding box regression, this makes it simpler to learn and utilize. Grid-based approach: YOLO accomplishes object detection by taking a single look at an image and dividing it into cells. This helps it handle objects of different sizes and makes it less sensitive to item scale than other object detection systems.

B. *YOLO V1.*

The original YOLO system was introduced in 2015 and used a single neural network to make predictions about the object categories and bounding boxes in an image. The network processed the entire image in a single forward pass and made predictions at multiple scales to handle objects of different sizes.

Fig.3 demonstrates the network architecture of YOLO. The architecture of YOLO v1 consists of the following main components:

Input layer: The input to the network is an image of fixed size (448x448 in YOLO v1).[22]

Feature extraction: A series of convolutional and max-pooling layers are used to extract features from the input image. The output of the feature extraction layer is a tensor with multiple feature maps.

Detection layer: This layer is responsible for making predictions about the objects in the image. The detection layer is a fully connected layer with output neurons equal to the number of grid cells in the feature map multiplied by the number of classes plus five (for the x, y, width, height, and confidence of each bounding box prediction).

Non-maximum suppression: After object detection, overlapping bounding boxes are removed using non-maximum suppression to produce the final object detections.

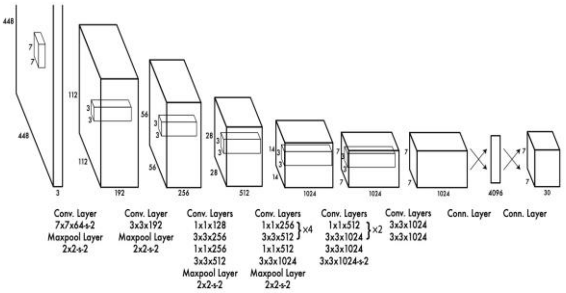


Figure 3. Network Architecture YOLO Illustration adopted from [22]

C. YOLO V2.

In comparison to the original YOLO system, YOLOv2, which was released in 2016, makes a number of advancements, including the usage of anchor boxes, a new loss function, and a higher-resolution feature extractor. Compared to YOLOv1, these advancements produced better accuracy and quicker processing times [23]. The YOLOv2 architecture is intended to overcome some of the drawbacks of the original YOLO architecture, including object scale sensitivity and challenges with small object detection. It is often faster than its predecessor and increases detection accuracy using anchor boxes and multi-scale predictions.[24]The following are the key elements of the YOLOv2 architecture: Input layer: The input to the network is an image of fixed size (416x416 in YOLOv2).

Feature extraction: A series of convolutional and max-pooling layers are used to extract features from the input image. The output of the feature extraction layer is a tensor with multiple feature maps.

Darknet-19: This is the base network architecture in YOLOv2, consisting of 19 convolutional layers interleaved with max-pooling layers. It serves as the backbone for the feature extraction process.

Detection layer: The detection layer is responsible for making predictions about the objects in the image. The detection layer is a convolutional layer that generates a 3D tensor with multiple channels, where each channel corresponds to a grid cell in the feature map.

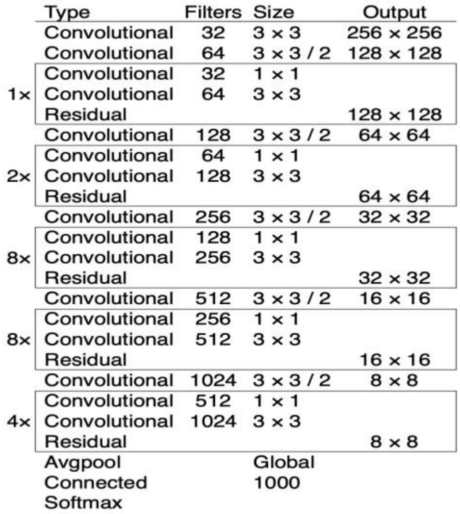
Unsampling layer: The detection layer output is upscaled to the original image size to provide more accurate bounding box predictions.

Non-maximum suppression: After object detection, overlapping bounding boxes are removed using non-maximum suppression to produce the final object detections.

D. YOLO V3.

YOLOv3, which was released in 2018, improved upon YOLOv2 by using a deeper network architecture, better training methods, and a new network module for handling complex objects. To more effectively handle objects of various sizes, YOLOv3 has added a multi-scale prediction system. The dark-net 53 architecture can be seen in the below Fig.4. To boost efficiency and accuracy, it combines numerous scales, FPN, anchor boxes, and better architecture [25]. YOLOv3's architecture includes three crucial elements that make it more precise and effective than earlier iterations. Multiple scales: To handle items of various sizes in an image, YOLOv3 employs multiple scales. This increases accuracy and makes the network more resistant to objects of different sizes. YOLOv3 uses a feature pyramid network (FPN) to extract features from an image's various scales. This increases the network's capacity to recognize objects of various sizes and contributes to increased accuracy. To handle various item scales and aspect ratios, YOLOv3 uses anchor boxes. For the purpose of object detection, bounding boxes called "anchor boxes" are employed as a reference.[26]

Improved architecture: YOLOv3 has a deeper and wider network than its predecessors, which enables it to learn more complicated features. More accurate detection of smaller objects is made possible by YOLOv3's use of a greater resolution for feature extraction.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 4. Dark-net 53 .Illustration Adopted From [25].

E. YOLO V4.

The fourth iteration of the YOLO method was released in April 2020 by authors Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao in their study "OLOv4: Optimal Speed and Accuracy of Object Detection." They present the self-adversarial training method and the mosaic data augmentation methodology (SAT) mosaic data enhancement [27] Utilizing specific ratios of mosaic data augmentation, four training images are blended into one. A brand-new data augmentation technique called Self-Adversarial Training (SAT) operates in two stages—forward and backward. In the first stage, the neural network alters the original image rather than the network weights. In the second stage, the neural network is trained to recognize an object on this altered image using standard methods. It is optimized for GPU performance, making real-time object detection applications possible.[28] The algorithm divides the image into an S × S grid of cells, where each cell is responsible for predicting the existence of objects, and employs a single convolutional neural network to recognize items inside the image.

F. YOLO V5.

Glenn Jocher released the fifth iteration of the most well-known object detection system. The PyTorch deep learning framework was used for the first time by YOLO. The official paper could not be published. The real-time object detection system YOLOv5 is used for computer vision tasks. It is an enhanced variation of the object detection models from the YOLO family created by the writers at Ultralytics LLC. The algorithm first creates a grid out of an image, after which it predicts whether or not there are items in each grid cell. A deep neural network that has been trained on substantial datasets of tagged images is used to make the predictions. A set of bounding boxes and associated class probabilities are produced by the network, and these are post-processed to provide the final detections. The modular architecture of YOLOv5 is one of its important characteristics since it makes customization and extension simple. The YOLOv5 architecture is also made to be scalable, which makes a variety of input resolutions possible. A feature extraction backbone network, numerous convolutional layers, and a prediction head make up the architecture of YOLOv5.[29] CSP ResNet or MobileNetV3 is the backbone network used in YOLOv5, and it is responsible for collecting high-level features from the input image. After being refined by a number of convolutional layers, these features are then used to predict the existence and positioning of objects in the image. [30] For each object in the image, the prediction heads, numerous convolutional layers, and prediction layers produce bounding box predictions. Anchor boxes, which are pre-defined bounding boxes utilized as references for making predictions about the locations of objects in the image, are used to make the predictions. YOLOv5 does multiclass object detection because it predicts the class of each object in addition to the existence and location of each object. YOLOv5's architecture is made to be effective and quick.

G. YOLO V6.

The YOLO (You Only Look Once) line of algorithms includes the cutting-edge object identification algorithm YOLOv6. It can be utilized in real-time applications because it is made to be quick and effective.[31] The YOLOv6 architecture is made up of several parts, which can be seen in Fig.5 including, extracting features from the input image is the responsibility of the backbone network. The core of YOLOv6 is a deep convolutional neural network with Res-Net inspired architecture. The neck is made up of several convolutional layers that enhance the features the backbone network extracted. Bounding boxes and class probabilities of the items in the image are predicted by the fully connected head layer. Anchor boxes are also included, which are utilized to align the predicted boxes with the ground truth boxes. Loss Function: The localization loss, objectness loss, and classification loss are some of the losses that make up the loss function utilized in YOLOv6. Non-Maximum Suppression (NMS): YOLOv6 uses NMS to eliminate overlapping and redundant detections, which enhances the algorithm's overall accuracy.

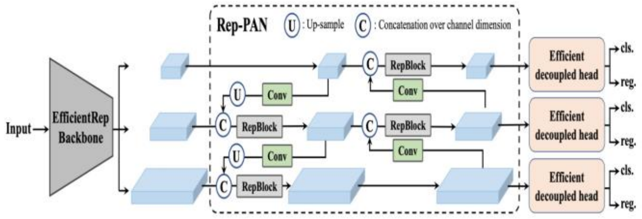


Figure 5. YOLOv6 Framework adopted from [31]

H. YOLO V7.

A real-time object detection system called YOLOv7 (You Only Look Once version 7) is based on the idea of a single convolutional neural network (CNN) architecture. The YOLOv7 architecture is made to be computationally efficient while still achieving great speed and accuracy. The following are the key parts of the YOLOv7 architecture: Backbone network: The backbone network is in charge of taking the input image's features and extracting them. ResNet is a deep convolutional neural network (CNN) type that has been trained on huge image datasets like ImageNet and is used by YOLOv7 as the backbone network. The features extracted from the backbone network are refined using the neck network. Convolutional and pooling layers are employed in YOLOv7's neck network to enhance the number of feature channels while reducing the spatial dimensions of the feature maps.[32] Head network: The head network analyzes the feature maps the neck network produced and conducts object detection on them. The final prediction is made using a number of Convolutional and Fully Connected (FC) layers. The head network predicts the bounding boxes of objects in an image using anchor boxes. In order to identify items in the image, anchor boxes, which are pre-defined boxes, are employed as reference points. YOLOv7 handles various object shapes by using anchor boxes with various aspect ratios. Loss function: The YOLOv7 architecture employs a multi-task loss function that seeks to reduce the variation in class probabilities, confidence intervals, and bounding boxes between predictions and the real world.

V. Proposed Model.

We have used a pre-trained YOLO v5 model for emotion detection. The input consists of a 48x48 pixel human face image with a text file containing its label.

Three components make up the suggested model architecture: the YOLO layers, the neck, which is made up of PANet Head for feature fusion, and the backbone, which is CSP-Darknet for feature extraction. A single forward pass of the network is employed in the model's single-shot multibox detection (SSD) technique to conduct object detection. In our previous research work [33] we used the YOLOv5 model for emotion detection from facial expressions. And we were able to achieve an accuracy of 50% with few images from the FER2013 dataset. The proposed model is depicted in Fig. 6 below.

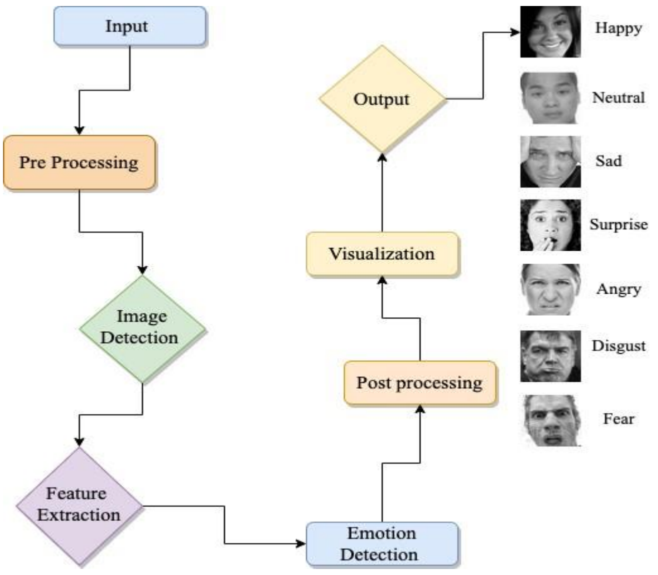


Figure 6. Architecture of proposed model

54 convolutional layers make up the entire structure including three fully connected layers and three output layers. The data labeling approach is used to determine the bounding box for the human face image. The determined bounding box's coordinates are kept in a separate text file. There are seven nodes in the output layer. The leaky ReLU activation function and the stochastic gradient descent optimization approach were utilized in the output layer and hidden layer, respectively, to provide trustworthy findings. Non-Maximum Suppression (NMS) was used to keep only the bounding box with the highest probability in order to eliminate overlapping bounding boxes. Based on the features that were retrieved from the image data, predictions are made using fully linked layers. The activations of the network's neurons are normalized using batch normalization layers. The model can learn more quickly and perform better during generalization by incorporating batch normalization layers. The final detection employs the sigmoid activation function. To enhance the performance of our model images from the FER 2013 dataset was used. The dataset includes more than 35,000 photos of faces in grayscale that have been annotated with seven different moods (angry, disgusted, fear, happy, neutral, sad, and surprised). This dataset is frequently employed in the creation and evaluation of emotion recognition models.

VI. Results.

The YOLOv5 architecture was employed in this work to identify emotions. Performance was assessed using the FER 2013 dataset. According to the results,

YOLOv5 performs better than several approaches in terms of recognition rate. With the pilot run of 32 photos in typical computing conditions, our model achieved an accuracy of 48.7%. We can increase the model evaluation by increasing the size of the dataset and number of epochs. An overview of YOLOv5's performance in terms of its capacity to correctly identify various emotions in photos is provided by a confusion matrix. The confusion matrix for our model is shown in Fig.7. Each emotion that the model had been trained to recognize has its row in the matrix, and each anticipated emotion has its column. The number of instances where the model predicted one emotion and the actual feeling was another is represented by each item in the matrix.

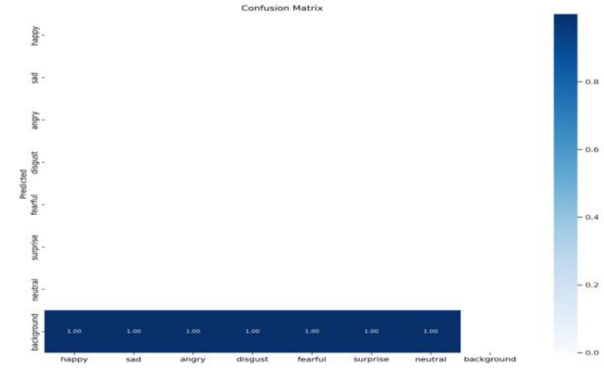


Fig.7. Confusion matrix of the proposed model

A graphical representation of the trade-off between the precision and confidence of the emotion identification predictions provided by the YOLO model is known as a precision-confidence curve for YOLO emotion detection. While confidence is the probability score that the model gives to each prediction, precision is the percentage of true positive predictions made by the model. The accuracy confidence curve of the proposed model is shown in figure 8. The evaluation's findings are then plotted to produce the curve, which can reveal information about the model's overall performance, its advantages and disadvantages, and the trade-offs involved in making predictions with a high degree of confidence. The number and quality of the training data, the complexity of the network architecture and the evaluation methods will all have an impact on the precision-confidence curve's form and location for YOLO emotion detection.

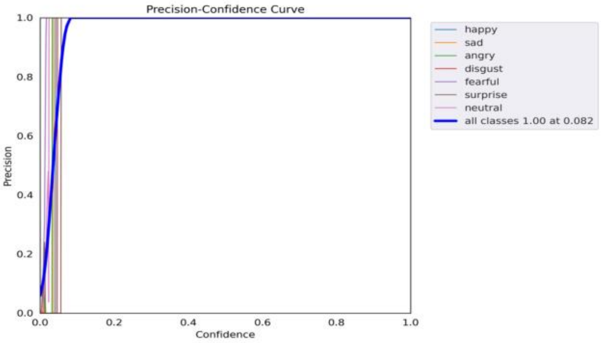


Fig.8 Represents the Precision Confidence Curve of the proposed model

VII. Conclusion.

Deep learning-based object detection techniques have recently attracted a lot of research attention. Consequently, the YOLOv5 algorithm for emotion detection is proposed in this study. Yolo v5's detection has a faster detection time than the conventional algorithm, which can lower the miss rate and error rate. In a complex setting, it can still ensure a high test rate, and the speed of detection can satisfy the need for real-time results quickly. The experiment has demonstrated that further training on the wide dataset can greatly improve the YOLO model's performance on emotion detection. Using a dataset of 40 photos to train our model, we were able to get a 48.4% accuracy with a batch size of 32 and for 60 epochs. We want to learn more about how various factors, such as the object's distance from the camera and the time of day, affect detection performance. For each of these circumstances, we want to assess the model's performance. The following conclusions can be drawn from the analysis done using the proposed model. First, the size of the network and the object also affect the learning rate. The learning rate should be kept low if the network is medium or large and the size of the image is considered to be less. Future comparisons between the performance of this model and that of its successors, YOLOv6 and YOLOv7, will help us better understand our model.

References

[1] Burić, M., Pobar, M., & Ivašić-Kos, M. (2019). Adapting Yolo Network for ball and player detection. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*.

[2] Luh, G.-C., Wu, H.-B., Yong, Y.-T., Lai, Y.-J., & Chen, Y.-H. (2019). Facial expression based emotion recognition employing yolov3 deep neural networks. *2019 International Conference on Machine Learning and Cybernetics (ICMLC).*

[3] Shinde, S., Kothari, A., & Gupta, V. (2018). Yolo based human action recognition and localization. *Procedia Computer Science*, *133*, 831–838.

[4] Birogul, S., Temur, G., & Kose, U. (2020). Yolo object recognition algorithm and "buy-sell decision" model over 2D candlestick charts. *IEEE Access*, *8*, 91894–91915.

[5] Aiswarya, P., Manish, & Mangalraj, P. (2020). Emotion recognition by inclusion of age and gender parameters with a novel hierarchical approach using Deep Learning. *2020 Advanced Communication Technologies and Signal Processing (ACTS).*

[6] Deepa, R., Tamilselvan, E., Abrar, E. S., & Sampath, S. (2019). Comparison of yolo, SSD, faster RCNN for real time tennis ball tracking for Action Decision Networks. *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*.

[7] Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., Zhang, X., Huang, Z., Liu, M., & Yang, T. (2021). Heu emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, *33*(14), 8669–8685

[8] Garg, D., Goel, P., Pandya, S., Ganatra, A., & Kotecha, K. (2018). A deep learning approach for face detection using Yolo. *2018 IEEE Punecon*.

[9] Luh, G.-C., Wu, H.-B., Yong, Y.-T., Lai, Y.-J., & Chen, Y.-H. (2019). Facial expression based emotion recognition employing yolov3 deep neural networks. *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*

[10] Lu, Z., Lu, J., Ge, Q., & Zhan, T. (2019). Multi-object detection method basedon Yolo and ResNet Hybrid Networks. *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*.

[11] Mindoro, J. N., Pilueta, N. U., Austria, Y. D., Lolong Lacatan, L., & Dellosa,R. M. (2020). Capturing students' attention through visible behavior: A prediction utilizing yolov3 approach. 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC).

[12] Kuldeep Mehta, Ashitosh Bhinge, Aditya Deshmukh, & Alka Londhe. (2020). Facial detection and recognition among heterogenous multi object frames. International Journal of Engineering Research And, V9(01).

[13] Deng, P., Wang, K., &amp; Han, X. (2022). Real-time object detection basedon Yolo-V2 for Tiny Vehicle Object. SN Computer Science, 3(4).

[14] Wu, K., Bai, C., Wang, D., Liu, Z., Huang, T., & Zheng, H. (2021). Improvedobject detection algorithm of Yolov3 Remote Sensing Image. *IEEE Access*, *9*, 113889–113900.

[15] Rajendran, S. P., Shine, L., Pradeep, R., & Vijayaraghavan, S. (2019). Real-time traffic sign recognition using yolov3 based detector. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.

[16] Yang, W., & Jiachun, Z. (2018). Real-time face detection based on Yolo. *2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII).*

[17] Wu, D., Lv, S., Jiang, M., & Song, H. (2020). Using channel pruning-based Yolo V4 deep learning algorithm for the real-time and accurate detection of Apple Flowers in natural environments. Computers and Electronics in Agriculture, 178, 105742.

[18] Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., & Liang, Z. (2019). Apple detection during different growth stages in orchards using the improved Yolo-V3 model. *Computers and Electronics in Agriculture*, *157*, 417–426.

[19] E.K., V., &; Ramachandran, C. (2020). Real-time gender identification from face images using you only look once (YOLO). 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184).

[20] Chen, H., He, Z., Shi, B., & Zhong, T. (2019). Research on recognition method of electrical components based on Yolo V3. *IEEE Access*, *7*, 157818–157829.

[21] Chen, W., Huang, H., Peng, S. *et al.* YOLO-face: a real-time face detector. *VisualComputer* 37,805–813 (2021).

[22] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[23] Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Shi, B., Li, X., Nie, T., Zhang, K., & Wang, W. (2021). Multi-object recognition method based on improved YOLOV2 model. *Information Technology and Control*, *50*(1), 13–27.

[25] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement.

[26] Won, J.-H., Lee, D.-H., Lee, K.-M., & Lin, C.-H. (2019). An improved yolov3-based neural network for de-identification technology. *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*.

[27] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection.

[28] Degadwala, S., Vyas, D., Chakraborty, U., Dider, A. R., & Biswas, H. (2021). Yolo-V4 deep learning model for medical face mask detection. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*.

[29] Jung, H.-K., & Choi, G.-S. (2022). Improved Yolov5: Efficient object detection using drone images under various conditions. *Applied Sciences*, *12*(14), 7255.

[30] Castellano, G., De Carolis, B., Marvulli, N., Sciancalepore, M., & Vessio, G. (2021). Real-time age estimation from facial images using Yolo and EfficientNet. *Computer Analysis of Images and Patterns*, 275–284

[31] Li, Chuyi & Li, Lulu & Jiang, Hongliang & Weng, Kaiheng & Geng, Yifei & Li, Liang & Ke, Zaidan & Li, Qingyuan & Cheng, Meng & Nie, Weiqiang & Li, Yiduo & Zhang, Bo & Liang, Yufei & Zhou, Linyuan & Xu, Xiaoming & Chu, Xiangxiang & Wei, Xiaoming & Wei, Xiaolin. (2022). YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications.

[32] Wang, Chien-Yao & Bochkovskiy, Alexey & Liao, Hong-yuan. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.

[33] Shaikh, A., Kanojia, M., & Mishra, K. (2022). *Emotion Detection Based on Facial Expression Using YOLOv5*.

Author Biographies

Awais Shaikh Currently enrolled in Sheth L.U.J. and Sir M.V College in Mumbai, Maharashtra, India, where he is pursuing a bachelor's degree in Computer Science. Artificial Intelligence, Deep Learning, and Machine Learning are the focus of Awais' research. During his undergraduate studies, he worked on a research project on Emotion detection based on facial expression using YOLOv5 and a comprehensive study on Emotion Detection with Facial Expression Pictures using YOLO Models, which sparked his enthusiasm for research. His interest in research was sparked by this event, which motivated him to look for more career options in the area. He is appreciative of the chances and encounters he has had thus far and eager to see what the future brings.

Keshav Mishra Currently pursuing a Bachelor's Degree in Computer Science from Sheth LUJ & Sir MV College, Mumbai University, Mumbai, Maharashtra, India. His current research focuses on deep learning, machine learning, machine translation for Sanskrit language, facial emotion detection using YOLOv5, transformer models for Sanskrit to English translation, machine learning techniques for detecting mental health issues. In addition, he is also exploring the field of data science and data analytics. This event sparked his interest in research, which led him to explore more career options in this area. He feels grateful for the opportunities and experiences he has had so far and looks forward to seeing what the future holds.

Pradnya Kharade received a Master's degree in Computer Science from the University of Mumbai, Maharashtra, India, in 2012. She is been teaching computer science for more than ten years. At the moment, she works as an Assistant Professor at Sheth L.U.J. and Sir M.V. College in Mumbai, India. Her current research focuses on A comprehensive study on Emotion Detection with Facial Expression Images using YOLO Model. She is gradually establishing herself as a computer science research scientist with expertise in multiple disciplines.

Mahendra Kanojia Currently employed as I/C Principal and HOD of the Department of Computer Science and Head of the Computer Education Centre in Sheth L.U.J. and Sir M.V. College, Mumbai, India. He completed his Ph.D. in year 2020 and M.Phil in Computer Science in year 2017. His current research of interest focuses on the detection of cancer using machine learning and deep learning techniques. He is also interested in the paradigm of medical diagnosis using digital image processing and AI approaches. He is exploring the field of data science and data analytics after receiving his PhD in Computer Science on Breast cancer detection using deep learning methods. Studies of IoT and chatbots are also part of his current projects. He is emerging as a multidisciplinary computer science research scientist.