Face Recognition Analysis Based on the YOLO Algorithm

Fang Sun

Rutgers University, New Brunswick, NJ 08901, USA

fs434@scarletmail.rutgers.edu

Abstract. Algorithm YOLO consists of a discrete CNN method to achieve end-to-end detection of the target. The detected target is acquired by processing network prediction results that match up with the R-CNN algorithm; it is an integrated system with faster speed and an end-to-end mechanism for training. The first kind of method is slower but more accurate, and the second kind of algorithm is faster but less accurate. This paper introduces the Yolo algorithm. Face recognition analysis is an important task, basically summarizing the properties of this algorithm: Firstly, You Only Look Once which depicts that a single CNN operation is necessary, and Unified portrays that it is a combined system that gives end-to-end results of the prediction, while "real-time" says that the Yolo algorithm is quick. The YOLO 4 algorithm is slower than the SSD algorithm, but Yolo has continuously evolved, resulted in YOLO 9000. This paper basically portrays the principle of the Yolo3 algorithm, mainly the details related to the training, and detection, and finally gives how to use TensorFlow to realize the Yolo algorithm.

Keywords: computer vision, deep learning, face recognition, yolo detector.

1. Introduction

Face recognition main algorithm principle of the mainstream facial detection technology is categorized into:

1) This method consists of geometrical features, template-based method, and model included method.

2) The geometric features method is the oldest and traditionally used process, which is generally integrated with alternate methods to achieve good results.

3) Methods using templates can be categorized into methods based on relative matching, eigenface process, linear discriminant analysis process, singular value decomposition process, neural network process, dynamic connection matching process, etc. [1]

4) Methods using models consist of the hidden Markov model, active appearance model, and active shape model.

1.1. Geometric functions method

A Face consists of organs like eyes, nose, chin, mouth and other organs of different sizes and the structure of these organs is very different for different people all around the world so the geo values of the structures of these parts can be an important feature for the facial recognition [2]. Description and recognition of face profiles were the first uses of Geometric features. First step is to determine the number of salient points using a profile curve, and a set of properties for recognition such as length and inclination were calculated from the respective salient points.

1.2. Face analysis using local method

Atom space is small and compact for main atoms which mostly decreased the dimensions, it is a local, but expanding in the whole coordinate space, at the same time with the help of the kernel function of support which is topology, The axial projection neighboring points and the method for the center point of the real image space has no impact tin the process, and topology and local analysis model and segment method is the ideal properties, it is to be in more accordance with the central neural information. Based on this consideration, Atick used local features only to propose a facial feature extracting and recognition method, this method has got better solutions in real time application, and this becomes the core software of Facet facial detection.

1.3. Eigenface method (PCA)

The Eigenface method is among the most important algorithms and was designed by Turk and Pentland in the 1990s. It is basic and efficient. Also known as the Principal Component Analysis (PCA) it uses a face recognition process. Core idea of eigensub-face technology is referencing a statistical viewpoint, searching for the simple features of face detection like eigenvectors of the covariance matrix of face features set of examples for training, and characterizing face images approximately. The eigenvectors determined are called Eigenface; it is to get the properties like position, distance, size of the parts and other attributes like contour of the iris, nose, and mouth, and then determine their geometric feature quantities, which describes the facial image by forming a feature vector. The core of its technology consists of local human feature methods for analysis and recognition algorithms using graphics and neural methods [3]. That particular algorithm is a process that utilizes facial parts, distinct organs of the body. For instance, the attributes used for the identification of the geometric comparison of multi-data development are differentiated and judged with every genuine attribute stored in the data storage. Turk and Pentland proposed the feature face method, which constructs the principal element subspace according to a group of face images for training. As the principal element is the shape of a face, also known as the featured face, the test feature is integrated to the main element subspace in the time of recognition which results in a pair of projection coefficients, which are then differentiated with the features for the face of every known person for facial detection [4]. Pentland et al. reported fairly good results, 95% positive perception rate for 3000 assets of 200 distinct people, on the other side negative perception rate was 1 in 150 face assets in the FERET database. However, a lot of pre-processing like normalizing must be done prior to the eigenface algorithm being implemented.

2. Method

2.1. Introduction of YOLO detector

2.1.1. PNet. Proposal Network (P-NET) is the Network structure which mainly determines the regression vector of the candidate window and the boundary box of the face region and makes regression with the boundary box, performs calibration on the candidate window to merge highly overlapping candidate boxes using non-maximum suppression (NMS) Since the actual image sizes are different, PNet is a fully convolutional network, and the size of the input image can be any value. Before entering the image into PNet, there is a loop, each loop will scale the image, and then enter PNet; In this way, an image pyramid is formed. Each scaling factor of the image is 0.80(the initial value of the paper should be 0.709). When both the dimentions of the image are less than 12, the corresponding cycle of the image ends.

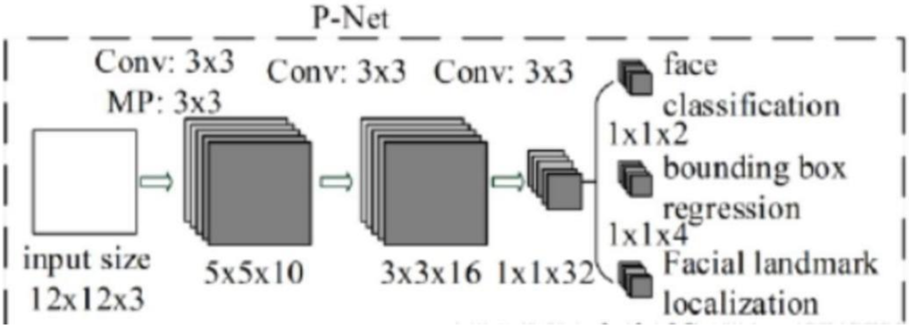

Figure 1. Framework of PNet.

As can be seen from the above picture, the final output result of a 12x12x3 picture is a 1x1x32 feature picture, which is then divided into three branches for face classification frame regression face feature point positioning.

The loss function of the three branches is the sum of squares of cross entropy (commonly used in dichotomy problems). The sum of squares of the five characteristic points and calibrated data is the total loss of the three losses multiplied by their weight ratios. In PNet, the weight of the three losses is 1:0.5:0.5 After I input the image into PNet, I get cls_cls_map, The two arrays of REG, where CLs_CLS_map is a two-dimensional array of (H, W,2), are non-human faces and the probability of human faces. The boundary box directly output by PNet is not the boundary coordinate in traditional regression, but the position difference between the predicted human face position and the input picture, namely reG.

The probability of face in CLS_CLS_map is compared with a preset threshold value. If it is greater than this threshold value, the predicted value in the corresponding REG array of this picture is extracted. The original pixel coordinate input for X * y is obtained through inverse operation, which will produce a size of [(x 12)/2+1] [(y Since the step size of the pooling layer is 2, the denominator of the above formula is 2. By using this method, the coordinate of REG can be restored to the pixel coordinate of the predicted border value in the original picture Finally, the returned array is (x1,y1,x2,y2, score, reg), where (x1,y1,x2,y2) is the pixel coordinates of Bbox in the original picture and score is the face probability corresponding to CLs_cls_map.

After completing this step, some duplicate boxes will be removed using the non-maximum suppression method (NMS). The principle of this algorithm is to extract the element in the row with the largest score value of the array returned in the previous step, compare the score of all remaining elements with a set threshold value, discard the elements whose score value is greater than the threshold value (0.5), and then discard the remaining elements Under the element to repeat before the extraction of the maximum and comparison of the operation until the end, so that the initial abandonment of those high degrees of convergence of the face box Because (x1,y1,x2,y2) are the pixel coordinates in the original image, and reg is the deviation of the candidate box area relative to the pixel coordinates. In this way, the candidate box coordinates are obtained by adding the original pixel coordinates to the deviation value. After the initial screening of the Bbox following the refine above, the PNet section is finished. The output is the candidate box's four coordinates plus the corresponding score value.

2.1.2. PFLD. The main network is surrounded by the yellow curve, which helps in predicting the position of feature points. The auxiliary network is surrounded by the green curve which can predict face posture in training (some literature shows that adding this auxiliary task to the network can improve positioning accuracy, please refer to the original paper for details). This part is not needed in the test. For the above challenges affecting accuracy, the loss function is modified to focus on those rare samples during training, while the lightweight model is used to improve the calculation speed and reduce the model size

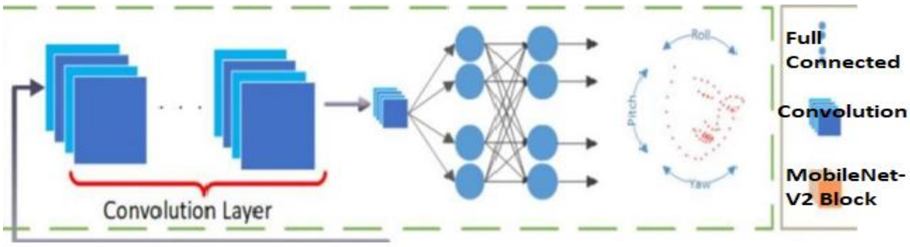

Figure 2. A framework of PFLD.

The loss function is used for the shape predicted by the neural network and the error of labeled shape during each training. Considering the imbalance of samples, the author hopes to assign a higher weight to those rare samples.

2.2. Algorithm design

The introduction to the YOLO algorithm.

Even before YOLOv1 was published, the R-CNN algorithm was the best in target detection series. R-CNN is highly accurate in detection. However, the two-stage structure networking has a low recognition speed which is difficult to implement real-time processes and is widely judged on it. To terminate this deadlock situation, which is an unavoidable trend to develop a faster target recognition.

Joseph Redmon, Santosh Divvala, Ross Girshick, et al. published the one-stage target recognition thread in 2016 and the recognition speed is high, it would go through 45 frames per second, and it can easily run instantaneously [5]. Due to the high speed and different process method, the author called it "You Only Look Once" (also known as YOLO) and published the results in CVPR In 2016, the basic ideology of YOLO is converting object recognition to a regression function by using a complete graphical data to input into the network and obtaining its locations and groups of bounding boxes from a neural network.

YOLOv1 operates on a divide-and-rule method, dividing an image into 7 or 7 networks, and here each network is important for target prediction that the middle point would fall on in that grid. Recall that in Faster In R-CNN, an RPN is used to obtain the area of interest of the target. This method has high accuracy, but it needs to train an additional RPN network, which undoubtedly increases the burden of training. In YOLOv1, 7 or 7 grids are obtained through partitioning, and the 49 grids are the same as the area of interest of the target. In this process, we do not need to design and develop an additional RPN network, which is why YOLOv1is simple as a one-stage network.

In 2018, Redmon improved on YOLOv2. For the feature extraction process the darkNET-19 network was replaced by darknet-53, and the feature pyramid networking structure was used from which multi scale detection was implemented [6]. Softmax was replaced by the logistic regression from classifying the method, by which accuracy and instant performance for real time target detection was ensured. From YOLO version1 to YOLO version3, for every generation the performance is mainly linked to the improvement of the Backbone (backbone network). In YOLOv3, the author provides Darknet-53 and lightweight Tiny-Darknet, when you want both precision and speed, darknet-53 would act as the backbone [7]. If you want to achieve faster detection speed and compromise accuracy, tiny-Darknet is a great choice for you. In short, YOLOv3's flexibility makes it popular in practical engineering.

Introduce the face recognition task:

The specific implementation process is as follows:

1. Image is segregated into S*S grid cells. The network grid is important for detecting the object when the middle of an object drops in this grid.

2. Every bounding box is predicted with B bounding boxes, and each box is predicted with 5 parameters (x, Y, W, H) and CONFIDENCE.

3. Every grid also finds a sector of information, denoted as C classes.

4. In general, S*S grids, and every grid is required to detect B bounding boxes and C network-like outputs is S*S (5 B+C) tensor.

Loss is composed of three parts, namely: coordinate prediction loss confidence prediction loss category prediction loss using difference square and error. It should be noted that the square root of W and H is taken in error calculation because bounding boxes of different sizes are predicted, compared with big bounding for reducing this issue, writer takes the ingenious method, that is, taking the square root of W and H of bounding boxes instead of the original W and H The positioning error is greater than the classification error, so increase the penalty for the positioning error to λcoord =5 in each image, many grid cells do not contain any targeted training will put the confidence of the box in these grids The score is pushed to zero, which often exceeds the gradient of the frame containing the target and may lead to model instability. The training diverges early so that the loss of confidence prediction for the frame without the target is reduced to λnoobj=0.5 The convolution layer is a simple and crucial layer in CNN. It calculates the generated matrix of pixels for a particular image to generate an activation map of a provided image.

The important use of an activation map is to reduce the data to be processed and store different attributes of a given asset. Convolution with the data of the matrix is a feature detector, it is basically a machine compatible with a set of values using different characteristics detector to generate different versions of the image convolution model using back propagation training, to determine the minimum error of each layer [8].

According to minimum error Settings, assign the depth and fill in the following figure 1 shows the working principle of convolution the process has the convolution consisting image matrix, and characteristics of the detector, it provides us with an activated figure or features in the convolution, the value of the data and features in the same location (that is, the value of 1 or greater than 1) will be preserved, while the rest of the value to be deleted from the image data of a 3 x3 comparison matrix The size of the feature detector changes depending on the type of CNN.

Finally, we describe the Neural network structure detection. Complete the introduction to the data set: the quantity of pictures/quantity of face categories, and display of the face features in the data set. Facial recognition algorithms are the basis of face detection and recognition software. Experts, these algorithms can be divided into two core methods: geometric methods focus on the distinguishing features of luminosity; statistical methods are to extract value from the image data. Then these values are matched and judged with the template to avoid differences These algorithms can also be categorized into two basic categories Feature-based models and holistic models focus on facial markers and analyze their space parameters and correlations with other features, but holistic approaches treat human faces as one. And convolution neural network (CNN) is the artificial neural network (ANN) and one of the breakthroughs for the development of artificial intelligence [9]. It is among the most popular algorithms in deep learning, deep learning is a kind of machine learning, model learning on the image video text or voice classification tasks The model in multiple areas showed an impressive result: Computer vision - Natural Language Processing (NLP) and the huge Image classification data(Image Net) CNN is an ordinary neural network with fresh layers, convolution layer and pooling layer CNN can has many of such layers, and each layer learns to identify different imaging feature.

The Passthrough layer is same as the Shortcut of the ResNet network, it is input with the high resolution feature graph of the front and then connected to the lower resolution feature graph of the rear which has twice the dimension of the latter and the Passthrough layer extracts every 2 of the front layer For feature maps of 26*26 and 512, the new feature maps of 13*13 and 2048 are done by Passthrough layer (the size of feature maps is reduced by 4 times, while the size of channels is increased by 4 times), which can be compared with the following 13 and 13 The feature images of 1024 are connected to form feature images of 13*13*3072, and then convolved to make a prediction based on the feature images, which is shown in Figure 3.

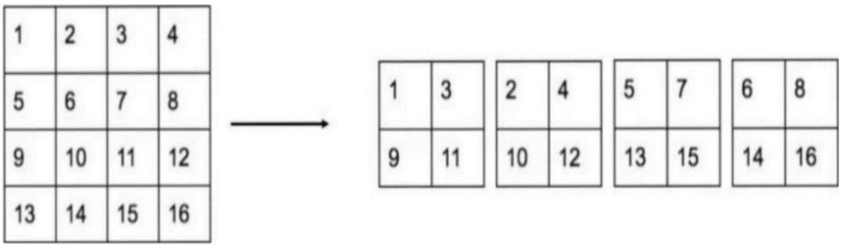

Figure 3. The illustration of the convolution operation.

Yolo utilizes a convolutional grid to get required features and then get the predicted values by using a complete connection layer. The networking points to the GooLeNet model, which contains 24 convolutional layers and 2 full connection layers, as presented in FIG. 8. For the convolutional layer, 1x1 convolution is important for Channel reduction, with 3x3 convolution for the convolution layer and fully connected layer, the Leaky ReLU activation function is applied: [formula] But for the last layer, the linear activation function is taken.

We can see the end of the frame-network's outputs of $30 \times 7 \times 7$size formula. And as the previous discussions we have consisted. To which means of the tensor represents as presented in fig 4 For every cell, the top 20 attributes is class probability value, and the other two elements are the bounding box confidence, confidence multiply will get category, the last eight elements are bounding box (x, x, x, h) we wonder whether the confidence degree and (, , , h)are arranged separately in addition to the boundary boxes, rather than in accordance with (, , , h). In fact, it is purely for the ease of calculation, in other words, the 30 elements are corresponding to a cell, and the arrangement can be random. But by separating the arrangement, each part can be easily extracted. To explain it here, first of all, the determined value of the network is a two-dimensional tensor P, whose dimension is [h, 7 × 7 × 30].

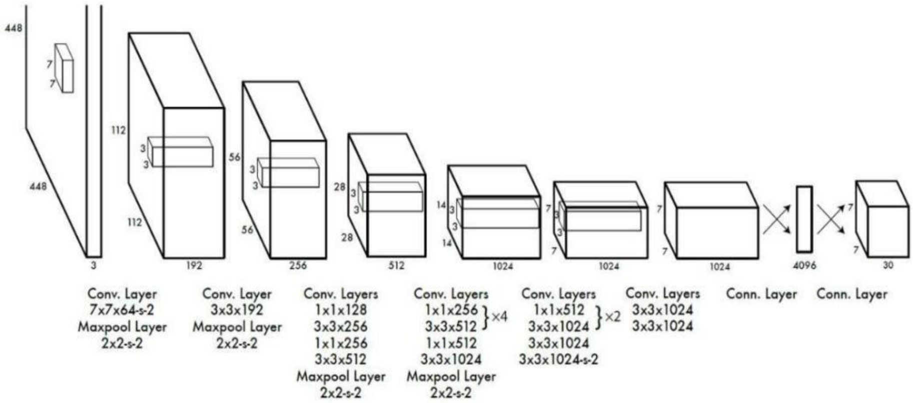

Figure 4. Framework of YOLO detector.

At the other point, since every cell predicts many boundary boxes but only the related category during the training process, if there is indeed a target in this particular cell, only the boundary box with the highest IOU related to basic truth is said to be responsible for predicting this target, when other boundary boxes think there is no other target One solution for this setting will be to professionalize the bounding boxes respective to a cell, and implemented to targets of various sizes and aspect ratios, thus gaining model performance You might think that there are multiple targets within a single cell, whereas Yolo algorithm can only train one, where the Yolo algorithm shortcoming Pay attention to the point, for there is no corresponding target bounding box, the error term is only the confidence, coordinate error is not calculation. The error in classification can be determined only when the target is in a cell, otherwise it cannot be determined.

Prior to the detection method of the YOLO algorithm, a non-maximum suppression algorithm was initiated here. NMS), this process is not just applicable to the Yolo algorithm, but all detection methods will be using the NMS algorithm mainly to solve the problem of multiple detection of a target. Face detection can be seen for multiple detection, but we could output only the best prediction boxes at last. For instance, a beautiful woman, let's say we require the recognition result in red Therefore, NMS algorithm would be utilized in achieving such effects:

Firstly, determine the box which has the most confidence among other recognition boxes, and then find the IOU for the same and for the other boxes one by one. If its quantity is larger than a particular reference point (the coincidence degree is too high), the box would get deleted. Then the same process repeats for the remaining frames until all detection frames are checked. The Yolo prediction process thus needs an NMS algorithm.

Preparation of all data has been received, so we say the basic strategy first is to obtain the result of the text box, I think that is most normal and natural. First, the box for each prediction confidence is selected according to the classes that category as its forecasts, after the layer handles, we get different forecast box of prediction categories and corresponding confidence value, its size is [Formula] In general, the confidence threshold is set, that is, the boxes whose confidence is lower than this threshold are cleared out. That is why after this part of processing, the remaining boxes with high confidence are finally utilized for these prediction boxes, and the detection results are left, at last its worth understanding whether NMS takes all prediction boxes equally or uses NMS separately for each category Ng says in deeplearning.ai that NMS should be used separately for each category, but after referring to several implementations, every box is treated equally [10]. I estimate that the probability of objects of various categories appearing in the particular location is very less. The above stated prediction method should be basic, simple, and straightforward, but in the case of Yolo algorithm, another approach is adopted (according to the C source code). The main difference is to do NMS and then find the categories of each box. For the 98 boxes, the values below the confidence threshold are first returned to 0, and NMS is applied to the confidence value separately. The NMS does not eliminate the boxes, rather changing their confidence value to 0, and then determines the category of each box. When the confidence value is not 0, the detection result is output This strategy is not straight, but Yolo main code is direct. Yolo's paper said that the NMS algorithm is a great influence on Yolo's evolution, so maybe this strategy is good for Yolo, but I tested the ordinary picture detection, and the two strategies are the same result.

3. Experiments

3.1. Dataset description

Figure 5. Samples from our training dataset.

3.2. Comparison results

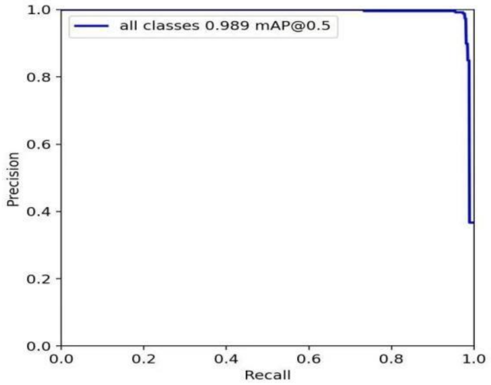

Figure 6. Result in mAP.

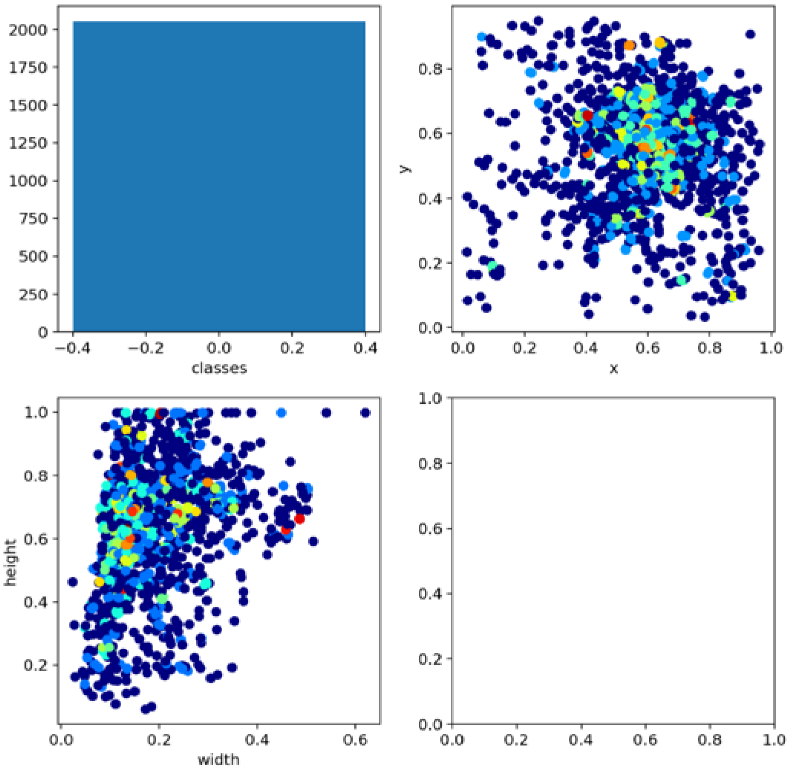

Figure 7. Result in scatter plot.

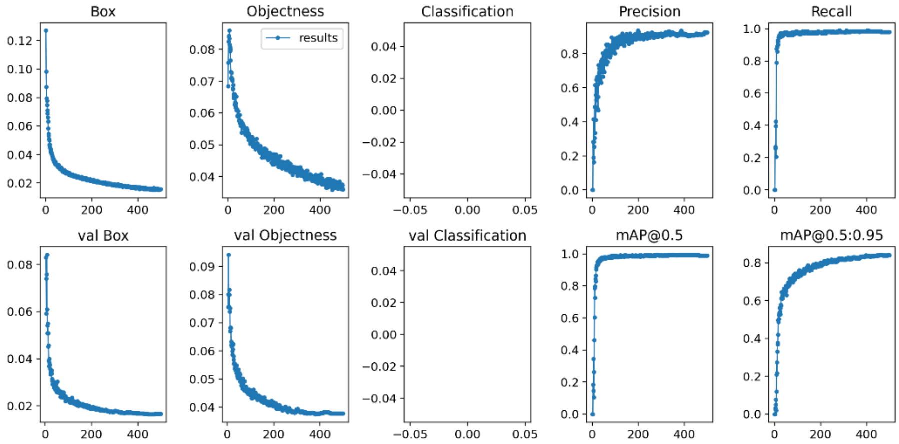

Figure 8. Result comparison.

3.3. *Demo representation*

The following is the analysis of the prediction process Yolo. Here, we do not consider Batch and think that only one input picture is predicted. The analysis says that the final network can be divide it into three categories: the category probability part, the confidence part, and the boundary box part (for this part, don't forget to calculate its true value according to the e original picture). Then multiply the first two terms (matrix times can be done by adding one dimension each) and you get a category confidence value of, and there is a total of predicted bounding boxes.

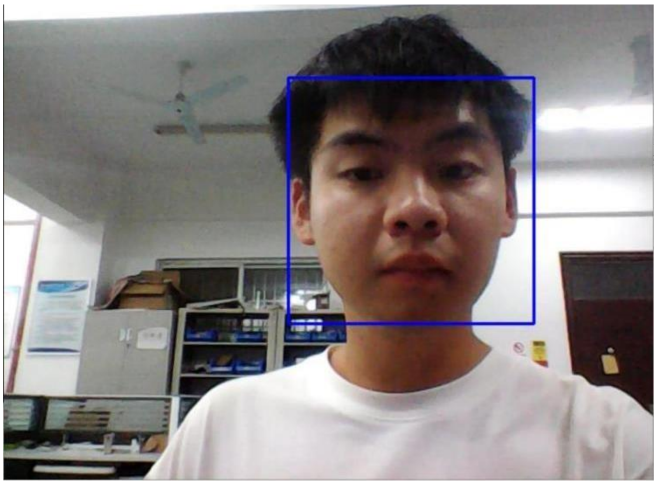

Figure 9. Demo illustration of the proposed YOLO detector.

4. Conclusion

This paper introduces the YOLO algorithm for face recognition analysis, which consists of a discrete CNN method to achieve end-to-end detection of the target. The first kind of method is slower but more accurate, and the second kind of algorithm is faster but less accurate. This paper also introduces the principle of You Only Look Once (YOLO). The article mainly focuses on how to use TensorFlow to realize this algorithm. Algorithm YOLO uses a discrete CNN approach to detect the target from beginning to end. The R-CNN method matches network prediction results to obtain the identified target; it is an intelligent approach with higher speed and an end-to-end training mechanism. The first type of approach is more accurate but slower, whereas the second type of algorithm is more accurate but faster. The Yolo algorithm is introduced in this publication. An essential task is face recognition analysis.

To summarize the characteristics of this algorithm: First, You Only Look Once suggests that only one CNN operation is required, while unified suggests that a combined system provides end-to-end results for the prediction, and "real-time" suggests that the Yolo process is speedy. Given a set of images and associated facial feature vectors, the goal is to detect the frames that contain faces with high confidence. This paper introduces the YOLO (You Only Look Once) classifier as a method for performing face detection and recognition. The framework for this classification is described, as well as several customizations that have been made to achieve best performance on various datasets. Finally, the Neural network structure detection has been described. Then the introduction to the data set: the quantity of pictures/quantity of face categories, and display of the face features in the data set are completed.

References

[1] K. Fukushima, S. Miyake. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. Competition and Cooperation in Neural Net(1982), pp. 267-285

[2] S.P. Khandait, P.D. Khandait and Dr.R.C.Thool, ―An Efficient approach toFacial Feature Detection for Expression Recognition, International Journal of Recent TrendsinEngineering, Vol 2, No. 1, November 2009,PP. 179-182

[3] B.B LeCun, J.S. Denker, D. Henderson, R.R. Howard, W. Hubbard, L.D. Jackel. Handwritten digit recognition with a back-propagation network. Proceedings of the Advances in Neural Information Processing Systems (NIPS) (1989), pp. 396-404

[4] Xiaofei He, Shuicheng Yan. Face recognition using Lapalcianfaces IEEE 10.1109/TPAMI.2005.55

[5] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[6] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21-26 July 2017, 7263-7271.

[7] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[8] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale face recognition. Proceedings of the International Conference on Learning Representations (ICLR) (2015).

[9] M.D. Zeiler, R. Fergus. Visualizing and understanding convolutional networks. Proceedings of the European Conference on Computer Vision (ECCV) (2014), pp. 818-833

[10] Gengtao Zhou, Yongzhao Zhan, Jianming Zhang, ―Facial Expression RecognitionBased on Selective Feature Extraction‖, Proceedings of the sixth International Conferenceon Intelligent System Design and Applications (ISDA'06) 2006 IEEE.