Real Time Person Detection and Classification using YOLO

Tejas Rao C, Mohammed Zainuddin, Shrishail M Patil, Shashank G, Nimrita Koul

Abstract*: *A Convolutional Neural Network (CNN) is a class of deep neural network most commonly used in analyzing visual images. Various systems and applications have been built to detect and classify the objects in a faster way taking CNN as its foundation. In this paper, we introduce a model to identify and classify people wearing ID card.Our model uses an object detection system called YOLO (You Only Look Once) for detecting and classifying objects in real-time videos. In the YOLO algorithm, a single convolutional network predicts the bounding boxes and the class probabilities for these boxes. We aim to use our model for authentication, surveillance and security purposes at organizations, corporations and educational institutions to detect an unauthorized person at the premises or somebody without a valid identification document. Using the object detection and classification, we aim to build a model which would alert the respective authorities on the matter.

Index Terms*: *Convolutional Neural Network,Object Detection and Classification, You Only Look Once(YOLO).

I. INTRODUCTION

Humans are an intelligent species which recognize and identify an object upon looking. This is usually by experience. The human visual system is fast and accurate in determining the object. Similarly fast and accurate algorithms can help computers detect and classify objects. YOLO is one such object detection system that can used to classify objects.
YOLO or You Only Look Once is an object detection algorithm much different from the region based algorithms. In YOLO, we reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. Using this system, you only look once (YOLO) at an image to predict what objects are present and where they are. YOLO is extremely fast, reasons globally about the image when making predictions and learns generalizable representations of objects.[1]
However we are limiting the object detection to just two classes here i.e., person and Identification cards. The datasets for these two classes are collected and trained separately only to be later integrated together in the system.

Revised Manuscript Received on April 24, 2019
Tejas Rao C, CIT, REVA University, Bangalore, India.
Mohammed Zainuddin, CIT, REVA University, Bangalore, India.
Shrishail M Patil, CIT, REVA University, Bangalore, India.
Shashank G, CIT, REVA University, Bangalore, India.
NimritaKoul, CIT, REVA University, Bangalore, India.

Perhaps the most critical part is the detection of an object within a bounding box. This requires extra convolutional layers for detection of ID cards.

Security is an important aspect in the smooth running of a corporation and educational institutions .The identification card serves as an important tool in providing and

implementing security. Hence for the collection of dataset various images of people wearing ID cards is collected from various angles.

Rest of the paper is organized as follows, Section I contains the introduction of object detection and classification system, Section II contains the related research work of Pedestrian Detection System Recognition, Section III explains in detail the methodology various steps involved, section IV gives the implementation detail of proposed methodology, Section V gives a brief discussion on results, Section VI concludes research work with future directions.

II. RELATED WORK

Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi [1] introduce YOLO a unified, real-time state-of-the-art object detection system where they reframe object detection as a regression problem to spatially separated bounding boxes and class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. This model can be used to train directly on full images and the loss function directly corresponds to detection performance of the entire model is trained jointly.

Joseph Redmon, Ali Farhadi [2] introduce YOLO9000 improved model of YOLO[1], the fastest real-time object detection system that can detect over 9000 object categories and can run at variety of image sizes to provide tradeoff between speed and accuracy. It uses darknet-19 classification network for feature extraction. It uses anchor boxes to predict bounding boxes, k-means clustering for IOU and all fully connected layers are removed to improve accuracy.

Retrieval Number: E10080585S19/19©BEIESP
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Joseph Redmon, Ali Farhadi [3] introduce YOLOv3, which is the updated version of YOLOv2 [2] a real-time object detection system based on darknet-53, and better at detecting smaller objects. For class prediction, logistic classifiers are used for multilabel classification and during training cross-entropy loss for object confidence and class predictions are used.

Sergey Ioffe, Christian Szegedy[4]introduces an algorithm that addresses the internal covariate shift problem in deep neural networks. This algorithm is used for constructing, raining, and performing inference with batch-normalized networks. The resulting networks can be trained with saturating nonlinearities, are more tolerant to increased training rates, and often do not require Dropout for regularization. By adding batch normalization to image classification model reduces the training time.

Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele [5]discusses about the human baseline for pedestrian detection and by manually clustering the recurrent errors of top detectors. The background-versus-foreground is the main source of errors in pedestrian detection. The convnets that have better scores in image classification and general object detection, but have low scores in localized detection around small objects can be improved by Bounding box regression (and non-maximum suppression).

Wenbo Lan, Jianwu Dang, Yangping Wang, Song Wang [6] improves the network structure of YOLOv2 algorithm and proposes a new structure YOLO-R which has better ability at detecting shallow pedestrians. YOLO-R consists of three additional passthrough layers used for extraction of shallow layer pedestrian features and the shallow layer features extracted from the Route layer of the original algorithm are improved from the 16th layer to the 12th layer, combine shallow layer features with deep layer features to extract more fine-grained features. Experiments show that this algorithm improves the accuracy of pedestrian detection.

Zhong Hong, Lei Zhang, Pengfei Wang [7] introduces YOLO-D an improved model of YOLOv2 to improve the pedestrian detection accuracy in crowded scenes. In YOLO-D, the low-level feature maps of the YOLOv2 to the higher layers in turn and introduces a head-shoulders model to solve occlusion problem. The results show that YOLO-D has good generalization ability, and better accuracy than YOLOv2 under the public pedestrian dataset.

III. PROPOSED WORK

The primary objective of this study is to build a real-time system for efficient detection of persons and the presence of ID card on them. This system is used for surveillance purposes in locations where there is less risk involved. For example, in large schools the security person mainly validates every person by the ID tag worn by students or employee of that school. For a person it can be wearing and may require multiple people to monitor depending on the crowd. So we introduce a system that can detect persons in real-time and classify them as authorized and unauthorized person. The proposed methodology is shown in Fig 1. The modules used in methodology are Image acquisition, image pre-processing and annotation, training and testing YOLO, evaluation, construction of ID card Detector, training and testing of ID card detector, integration of YOLO and ID card detector.

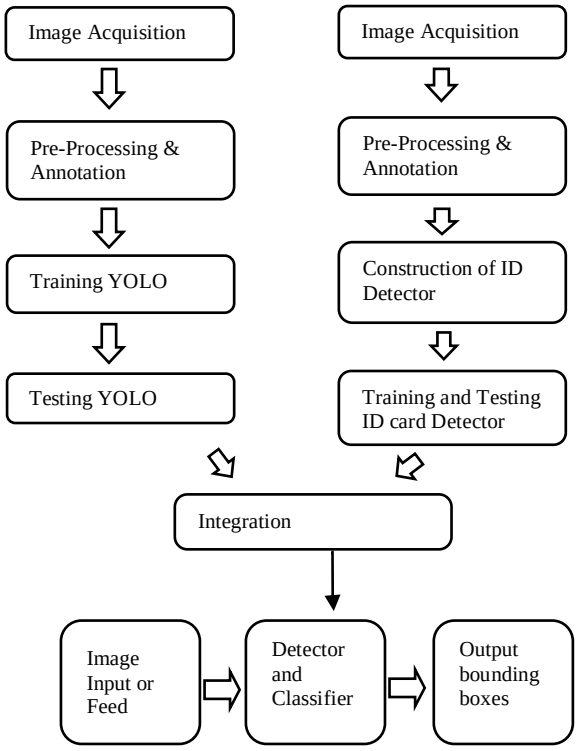

Fig. 1 Work flow of the project

A. *Image Acquisition*

Image Acquisition involves collection of images from different sources. We require two different image datasets one containing persons and another dataset of persons wearing ID cards. We are using the person dataset gathered from PASCAL VOC [8] and the ID card dataset gathered from capturing pictures of many students wearing ID from different position, angle, resolution and contrast.

B. *Image Pre-Processing and Annotation*

Image Pre-Processing involves processing or cleaning of images. This step focuses on removal of noise and distortion, sharpening, intensity normalization, etc. The VOC dataset is refined with only person images and annotated according to the format of YOLO model. A text file is created for each image in the same directory with the same name that contains object number and object coordinates on this image, for each object in new line. The object number is an integer number of object from zero to total number of classes – 1, and object coordinates are float values relative to width and height of image, it can be equal from (0.0 to 1.0]. The ID card images are only pre-processed.

Retrieval Number: E10080585S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

C. Training and Testing YOLO model

After pre-processing and annotation, the person dataset is dividedinto training and testing datasets. We train the YOLO model using training dataset until we get a better mean Average Precision (mAP). After the training, it is tested with testing dataset. YOLO is a full convolutional network consisting built using darknet-53. It detects objects at three different strides (8, 16 and 32) which helps to detect smaller objects. The input image is divided into S*S uniform grid, and each cell is composed of (x, y, w, h) and confidence C(Object). The coordinates (x, y) represent the position of the center of the detection boundary box relative to the grid. (w, h) is the width and height of the detection boundary box. Each grid predicts the probability of C categories. The confidence score is the probability of the model to include the target objectand the accuracy of the prediction detection box. Pr(Object) stands for whether there is a target object falling into this cell. If there is confidence, it is defined as:

$$C(Object) = \Pr(Object) * IOU(Pred, Truth)$$

If the cell does not have a object, the confidence score is zero C(Object) = 0. IOU is the overlapping rate of the generated candidate bound and ground truth bound, that is, the ratio of their intersection and union.

$$IOU(Pred, Truth) = \frac{area(box_{trut\,h}) \cap area(box_{pred})}{area(box_{trut\,h}) \cup area(box_{pred})}$$

After obtaining the confidence of each prediction box, the low-score prediction box are removed by setting the threshold value, and then non-maximum suppression is performed on the remaining bounding box.

D. Construction of ID detector

A new CNN model has to be constructed for the detection of ID card in a given image. The image should contain a person wearing an ID card. The person image can contain only one person. The output is probability of belonging to ID class. This model is trained until we get a better accuracy. If the model is over-fitted the hyper parameters are tuned else more and better dataset is required.

E. Integration

After the training of both models, the YOLO layers and ID card detector are combined together where the input is sent to the YOLO layers initially which localizes the persons in an image and forwards the coordinates to the ID detector model where it classifies the image as wearing ID or not wearing ID.

F. Testing Phase

After integration, the model should output bounding boxes of persons with person class probability and ID class probability for each person. This testing phase involves finding accuracy for a new dataset containing people belonging to both classes and testing it in real-time.

IV. RESULTS AND DISCUSSION

To evaluate the performance of our model we have prepared images taken under different environments such as rainy, fog, sunny, etc. with persons present at longer distances and wearing similar colour dress as ID cards. Our model is evaluated separately for YOLO layers and the layers for detection of ID card. The YOLO layers accuracy is measured using mAP and the ID detector layers are measured for classification accuracy. Along with the accuracy score, the FPS rate is also an important metric used for evaluation. The FPS of our model can be increased by using Tiny version of YOLO which has less convolutional layers than the darknet-53. As the convolution layers (higher resolution images) increases the FPS decreases. Fig.2 and Fig. 3 show the results of the proposed project in detecting the students as persons with bounding boxes and labels.

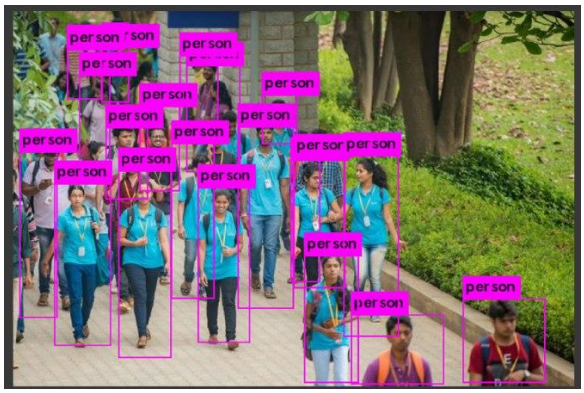

Fig. 2 – Person Detection and Tagging by the proposed system

Fig. 3- Person Detection and Tagging by the proposed system

V. CONCLUSION AND FUTURE WORK

We have studied about YOLO object detection system and detection of object within an object. This model requires high computational capabilities for higher resolution images. This time can be traded with accuracy and vice-versa. In the future work, we can introduce better approaches for real-time detection systems and work on nested object detections which can useful in many fields.

Retrieval Number: E10080585S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

[1] S. Divvala, R. Girshick, A. Farhadi, J. Redmon, "You Only Look Once: Unified, Real-Time Object Detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] A. Farhadi, J. Redmon, "YOLO9000: Better, Faster, Stronger," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[3] A. Farhadi, J. Redmon, "YOLOv3: An Incremental Improvement," 2018.

[4] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv:1502.03167, 2015.

[5] R. Benenson, M. Omran, J. Hosang, B. Schiele, S. Zhang, "How Far Are We From Solving Pedestrian Detection?," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1259-1267, 2016.

[6] J. Dang, Y. Wang, S. Wang, W. Lan, "Pedestrian Detection Based on YOLO Network Model," in The IEEE Conference on Mechatronics and Automation (ICMA), 2018.

[7] L. Zhang, P. Wang, Z. Hong, "Pedestrian Detection Based on YOLO-D Network," in The IEEE 9th International Conference on Software Engineering and Service Science, 2018.

[8] L. V. Gool, C. K. Williams, J. Winn, A. Zisserman, M. Everingham, "The Pascal Visual Object Classes (VOC) Challenge," in IJCV, 2010.

AUTHORS PROFILE

Tejas Rao Cpursuing B.Tech (Computer Science and Engineering) in REVA University, Bangalore. His subjects of interests are Data Structures, Design of Algorithms, Data Science, Image Processing and Deep Learning.
Mohammed Zainuddinpursuing B.Tech (Computer Science and Engineering) in REVA University, Bangalore. His subjects of interests are Data Structures, Design of Algorithms, Data Science, Image Processing and Deep Learning.
Shrishail M Patilpursuing B.Tech (Computer Science and Engineering) in REVA University, Bangalore. His subjects of interests are Computer Visualization, Artificial Intelligence and Image Processing.
Shashank Gpursuing B.Tech (Computer Science and Engineering) in REVA University, Bangalore. His subjects of interests are Image Processing, Computer Visualization, Web Development and Machine Learning.
Nimrita Koul is an Assistant Professor at School of CIT, REVA University Bangalore.

Retrieval Number: E10080585S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication