Published in AJSE, Vol:19, Issue: 02
Received on 16th March 2020
Revised on 16th June 2020
Accepted on 30th August 2020

AIUB JOURNAL OF SCIENCE AND ENGINEERING
**ISSN: 1608 – 3679 (print) 2520 – 4890 (Online)**

# YOLO-Based Enhancement of Public Safety on Roads and Transportation in Bangladesh

Anjir Ahmed Chowdhury, Sabrina Kashem Chowdhury, Md. Hanif, Sadia Noor Nosheen, Md. Saniat Rahman Zishan

***Abstract*— In order to upgrade the efficiency level of multiple tracking like face, actions, characters, a deep learning method is introduced to reduce the accidents occurred in roads for carelessness and also to capture the criminals in Bangladesh. This paper presents a faster processing multiple detection method with the best possible outcome under the framework of YOLOv2 algorithm in the event of car accident, crossing foot over bridge and using the zebra crossing in Bangladesh. Different layers are added to the YOLOv2 algorithm to pass the information in various convolutional layers to detect multiple objects with actions. In this paper YOLOv2 algorithm under DarkFlow framework is used to achieve higher ratio of confidence value as the max convolutional layers reorganize the feature map so that other layers feature map can be matched with the bottom layers to achieve the expected output of the indicated events. By removing the noise from the unrelated area, the detections of the training video and test video adopt quite parallel confidence ratio.**

***Keywords— YOLOv2, training dataset, test dataset, confusion matrix, Precision rate, Accuracy ratio, Recall, Null error rate.***

## I. INTRODUCTION

Deep learning is making a way to make a huge advancement in the field of research. Recently, through the use of machine learning it is upgrading the field of detection of various targets including recognizing the actions also in the field of computer vision. The goal of such detection techniques is to enable higher security for the general people. The events also while using the footpath, using zebra crossing or having an incident like car accident. Applying YOLOv2 algorithm it not only specifies the face and objects but also provide require information that is preset on database. Using this it is creating the ability to recreate a positive awareness and simplest method of detecting culprits. Through this vision technique. The location of the objects & faces in the events are predicted by the bounded boxes which are labeled to ensure the higher detection ratio [1]. Again, there are others lots of parameters which also need to taken in concern such as precision rate, accuracy ratio, confusion matrix, average loss in terms of epochs etc. Using this model, it has the ability to accommodate all the features that need to be augmented. YOLOv2 model is robust and efficient one in terms of fastest detection technique [2]. By applying light weight detection technique model, it will train the frames of the videos for longer hours comparing to the YOLOv2 algorithm [3]. In this paper the algorithm is used to detect those persons who are crossing the roads during green signal or not again the persons who are using foot over bridge. In case of car accident, the Bangla number plate will be detected showing the information of the registered person of that particular vehicle's. Maximum amount of GPU power is used to get higher accuracy rate for this model [4]. The precision rate is also varying according to the distinction of GPU power. Having higher GPU, it will tend to transfer the detection ratio in real time environment. Regression problem that is face to detect multiple objects or face that is solved while overlapping by applying this model and the probabilities to associate the class domain is also minimized applying DarkFlow framework. Higher computing computation is required to obtain expected confidence ratio by using in-expensive device. Different classifiers are used to evaluate the frame's feature to attain detection technique [5]. Using the classifiers, it locates the various positions to detect through the system.

In this paper, a custom dataset is used for the learning of the training video. Again, test video is also performed to fill the confusion matrix and gather the data for the precision rate, accuracy rate and F-value. By applying the enhanced method and the YOLO network model it helped to detect four types of targets. Comparing with the other methods this detection method is quite improved in the part of detection speed and recall rate. Similarly, the qualitative analysis is performed on the basis of the sample size and as for the small sample lower limit of the samples is adapted [6].

This paper demonstrates the following sections like the introduction part covers the application of YOLOv2 algorithm under the DarkFlow network and its accuracy in determining multiple targets. Secondly, architecture of the whole detection process is illustrated in this portion by giving brief ides how the video will be captured and the description of showing the result in the display. Then, in the next part the flow chart of the whole process of YOLOv2 model is narrated and the simulation of training video outcome. After that the statistical analysis of the testing video and the result analysis through different graphs is stated. Finally, the conclusion part provides the overall the scenario of the detection and recognizing process of the three different events.

## II. ARCHITECTURE OF THE SYSTEM

### A. *Block diagram*

The method of detecting the multiple objects including the face and actions works in the deep learning form by enhancing the samples of the custom dataset. At first video is

**Anjir Ahmed Chowdhury** is currently doing Master of Science in Computer Science (CS) at American International University- Bangladesh, Dhaka, Bangladesh.
Email: anzira431@gmail.com

**Sabrina Kashem Chowdhury** completed B. Sc in Computer Engineering (CoE) from American International University-Bangladesh, Dhaka, Bangladesh.
Email: sabrinachowdhury708@gmail.com

**Md Hanif** completed B. Sc in Computer Engineering (CoE) from American International University-Bangladesh, Dhaka, Bangladesh.
Email: aiub.hanifmd@gmail.com

**Sadia Noor Nosheen** completed B. Sc in Computer Engineering (CoE) from American International University-Bangladesh, Dhaka, Bangladesh.
Email: safanoor795@gmail.com

**Md. Saniat Rahman Zishan** is an Associate Professor at the Department of EEE & CoE & Head of Department of CoE in American International University-Bangladesh (AIUB).
Email: saniat@aiub.edu

recorded by a camera then features are extracted from video frames. Again, for creating dataset frames of video is stored in a particular folder by defining the path. The information regarding the labeled image is stored in the database to show the information on the display. If the extracted video and the labeled image matched together then it checks if the features are matched or not. Simultaneously it checks the database to match the labeled image and the provided information against it. If both the database and features matched then the display will show the result. By migrating the dataset, it will obtain the initial weight of the class of filter. Sending the training data to the YOLO network it enhanced the matched classifier through loading the weight of the training samples. At the end, the test samples are also sent to the previous classifier and the weight of the previous trained data is also loaded to achieve the test output.

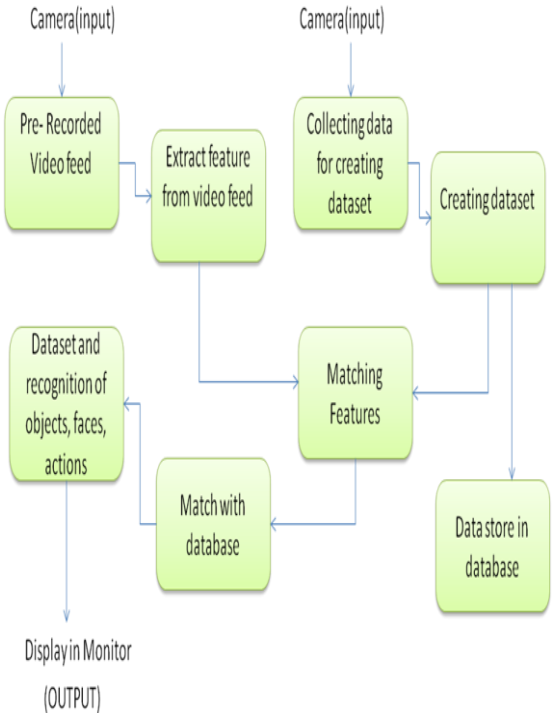


Fig. 1. Block diagram of the system.

### B. *Flow chart*

The YOLOv2 algorithm under DarkFlow framework can take into account the detection speed and also the higher accuracy rate underlining the regression method. Here it formats the convolutional layers in 3x3 portions by comprising 24 individual convolutional layers having 2 fully connected layers. The 6 maximum pooling layers also work as deep layers to transfer the features of the sampled data. After setting the camera in particular angle the video is recorded and divided into hundreds of frames. The frames are labeled then with Label Image toolkit. Then training video is also processed through Dark Flow model. Then applying the YOLOv2 algorithm it detects multiple faces, objects, actions within the bounded locations.

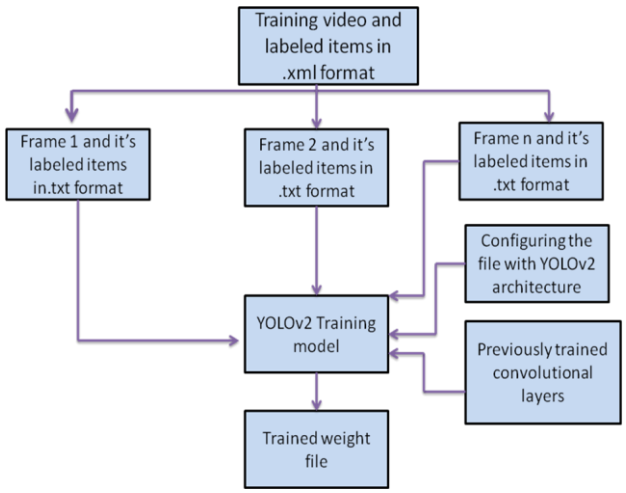


Fig. 2. Flow chart of YOLOv2 model.

## III. METHODOLOGY AND OUTPUT

Yolo divides the frame into S X S grids as whenever the center of an object falls into the grid then Yolo is responsible to detect those targets. The resolution of the images that are applied to be detected is 1920 X 1080 [7]. In YOLOv2 the S is stated with the value 13. By grouping the bounding boxes, it predicts its confidence value. Again, the bounding boxes show the information of translocation (x, y, and w, h) which represents the width, height and the center of the location of the bounding grids of the particular object [8]. And so that the labeled bounding boxes give information of 4 particular positions including the score of confidence value. The confidence score represents the efficiency of the selected location to be detected of the particular object, face and actions and it also refers the retention of the object. YOLOv2 holds the probabilities in C category to predict the confidence ratio. In DarkFlow framework, it overviews the feature map by generating the primary information of the targets. Although the regression portion of the bounding box provides the specific position of the targets that is labeled by referring the object classification. When it connects with the database it passes the information of the search that is selective for the individual through the convolutional layers. While training the frames the max pool layers statistically map the size of inputs that is distinct from one to another and convert it into fixed scale vector. The feature is extracted for each bounding region [9]. After that it passes the information to the other max convolutional layers to start the process of recognition.

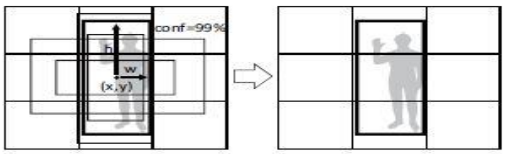


Fig. 3. Bounding Box [10].

The YOLOv2 model is trained using the frames that contain the actions that need to be recognized. In this paper, approximately 200 frames are generated for the purpose of the detection. It trained the file of custom dataset that is created by us. Again, for the purpose of testing around 35

frames are selected from a video for the intention of recognition and localization of the actions, objects, face and characters. The frames are selected after a minimum number of frames depending on the total number of frames corresponding to that video. All the paths of these frames of individual events are stored in a text file which is used as an input of the model. The output of the recognized file is stored as csv file format.

After applying the YOLOv2 model each frame are recognized with face, actions, objects or characters as per required. All these features are detected in more than 100 frames with a confidence threshold of 0.3 over 50 frames that are identified after the conclusion of the video frames. The total confidence score is calculated by averaging the previous confidence score.

In this paper, the YOLOv2 model takes the input of an image and then resizes it into 416 x 416. After that, the images are transfer to the convolutional layers and output is formed into the 7 x 7 x 30 tensor. By maintaining the thresholding ratio, the confidence value removes the label of class lesser than 25%.

This model applied on various locations on multiple events. The detections of higher confidence value are estimated as successful detections of that event. Different classifiers are used to classify the objects feature and localization of that specific portion. In this process the FPS of these frames are approximately 13-15 having slight latency in a few milliseconds. The Epochs value is measured is 300 with the learning rate is $10^{-3}$. This applied model is 100 times faster than YOLOv1 model [11].

By providing the number of total action classes with labeled images the path of the all frames are stored in text file for training. The action classes are also named with the text file and then the path of trained weight files is also indicated. The .cfg file represents the value of the filters for the second last layer in the max convolutional layer which is not an arbitrary one. But it depends on the number of classes that given by the filters. The number of layers in the trained file are 5.

$$\text{Filters} = \text{Second to last layer number} * (\text{classes} + 5) \qquad (1)$$

### *A. Training video output*

The annotations that are provided are formed in XML format. The dataset is exemplary for the detection of frames in videos in surveillance camera. There is multiple visual annotated actions which are detected and recognized simultaneously. The dataset contains 3 different actions from 3 different individual videos with extra 3 testing videos. The specifications required for this system are Ubuntu 18.04LTS with 64 bit and GPU Nvidia GTX 1060 6 GB including 16 GB RAM.

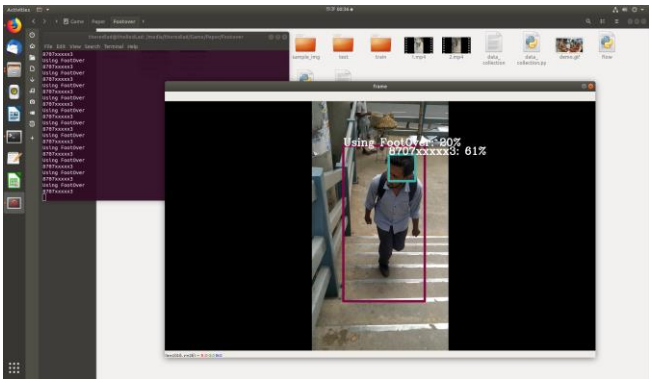

Fig. 4. Recognition and detection of trained dataset for Using Foot over bridge.

The action recognized in this event is 85% whereas the confidence value of face detection is 98%. In this training process the total iteration took place around 6 s. The whole model for this event is trained up to 150 iterations and again average loss is calculated approximately 0.65 for the batch size of 16 having 8 subdivisions.

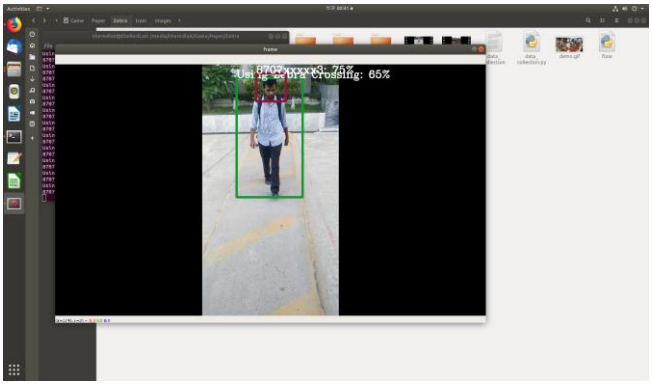

Fig. 5. Recognition and detection of trained dataset for Zebra crossing.

The action recognized for the event of zebra crossing is 65% and the confidence value of face recognition is 75%. Similarly, the whole iteration took place around 8s and around 180 iterations took place for this zebra crossing event. With 16 batch size including 8 subdivisions the average loss is estimated 0.54.


Fig. 6. Recognition and detection of trained dataset for the event of Car accident.

For the action of car accident, the confidence value of this event is 49% with 79% of face detection. The total iteration

took place around 10s with 16 batch size. Like previously around 175 iterations took place with average loss 0.48.

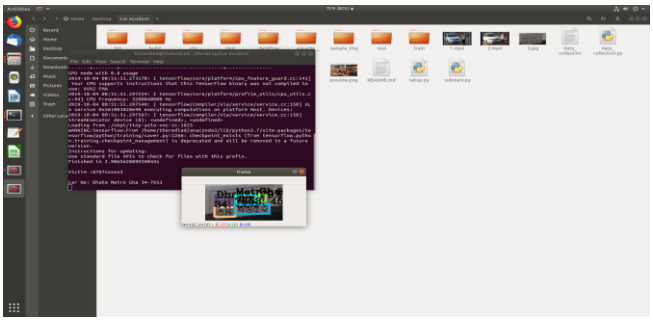

Fig. 7. Bangla number plate detection.

By training the number plate; the character detection is achieved by using custom made dataset. Here 6-digit number plate is detected by applying the higher algorithm. The character detection is augmented by training the YOLOv2 model maintaining the stability and taking into concern of not getting over fitted. Bangla number plate detection is augmented with features like contrast modification and random notation. All images are resized to 150 X 150 by maintaining the scale ratio. The character detection is about 89%.

*B. Test video output*

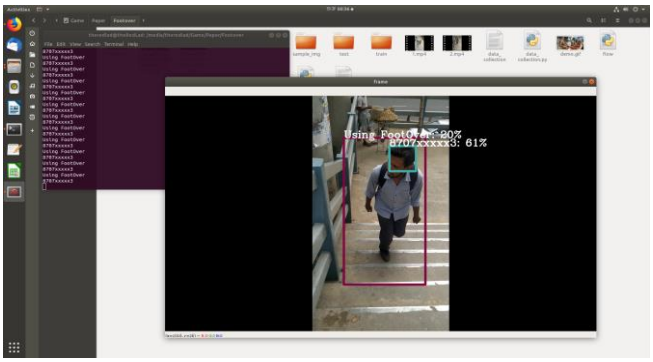

Fig. 8. Recognition and detection of test dataset for Using Foot over bridge.

For testing of foot over bridge event the average recognition that has been accounted for this event regarding action in per frames is taken around 60 ms with 12-15 FPS. The mAP is detected around 48.76% with FLOPS 7.04. The action of using foot over is recognized around 20% and face recognition is done with 61% confidence value.

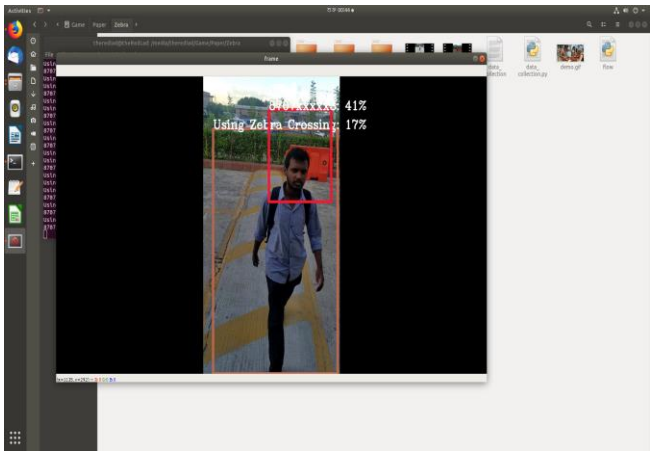


Fig. 9. Recognition and detection of test dataset for Zebra crossing.

The test video of zebra crossing takes up on the average rate of frame action and face detection which is 67ms with 13-15 FPS. The mAP rate that is calculated for this event is 30.76% including the flops 7.02B. The average loss for test dataset is 58%. The action zebra crossing is recognized is 17% and face recognition confidence value is 41%.

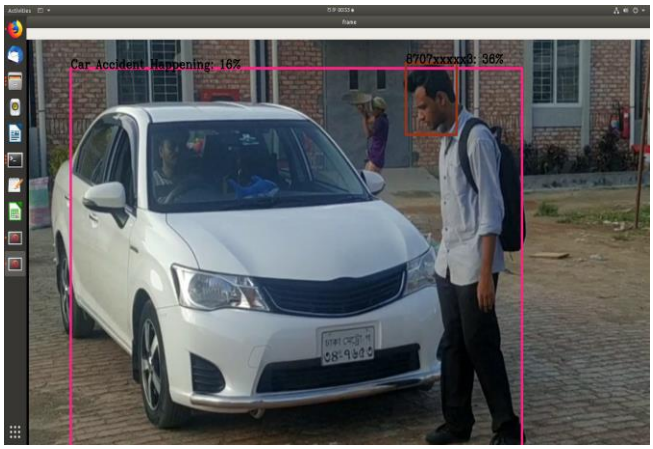


Fig. 10. Recognition and detection of test dataset for the event of Car accident.

The test dataset for car accident has proven the mAP rate 23.7% with 10-12 FPS. The FLOPS value is gained around 7.07B with average loss 48.87%. The data augmentation for transferring the image through random scaling is around 25% of the original trained data. In this event the car accident action is recognized around 16% along with the confidence value of the face recognition is 36%.

## IV. Result Analysis

The model of YOLOv2 algorithm trained for 300 epochs using DarkFlow framework. The accuracy rate is stable around 84% for 245 epochs and assigning the learning rate .0001. Having higher value of leaning rate, the accuracy and loss of the event recognition also varies.

*A. Rate of Average Loss*


Fig. 11. Loss vs Epoch in the event of foot over bridge.

On the event of foot over bridge the highest loss at 0 Epoch is achieved which is .105. When the frames are ended at 300 ephocs the loss came near to 0.2164. the average loss is calculated around 32%.

==

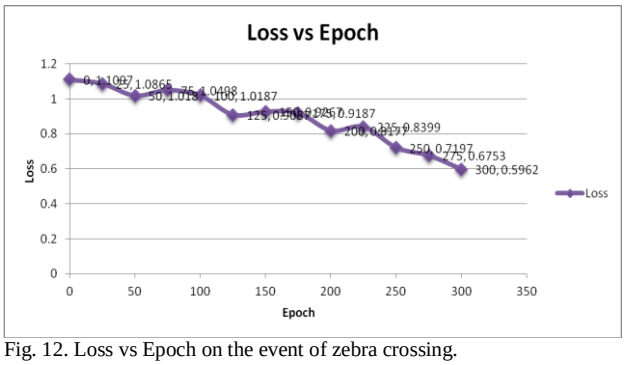


Fig. 12. Loss vs Epoch on the event of zebra crossing.

Again for the event of zebra crossing the average loss value randomly changed but at 300 Epoch the loss is gained 0.5962. The average loss for this event is obtained about 27% by keeping the same batch size and learning rate.

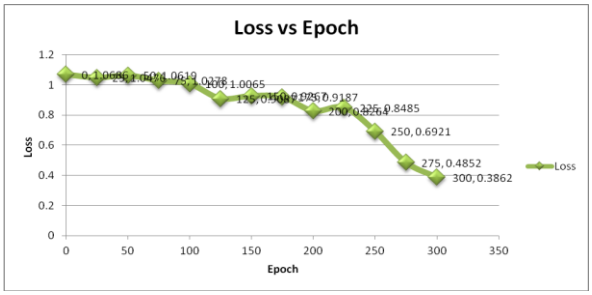


Fig. 13. Loss vs Epoch on the event of car accident.

Fig.13 shows that minimum 0.38 loss is achieved at 300 epoch while between the epochs of 0-100 the loss is about 1.00. The average loss is calculated near about 24% because coming close to the 300 epoch the loss came across to 0.00.

*B. Confusion matrix*

TABLE I.
CONFUSION MATRIX TABLE FOR CAR ACCIDENT

| | | | |
|---|---|---|---|
| TN: | 19 | FP: | 33 |
| FN: | 0 | TP: | 35 |

The total number n=87 and the total sum of individual row and column is 87. As, TN+FP=52 and FN+TP= 35. And sum of them is equal to 87. The sum of (TN+FN) + (FP+TP) = 87.

Accuracy, A= $\frac{TP+TN}{TP+FP+TN+FN}$ = 62% (2)

Prevalence = $\frac{FN+TP}{n}$ = 40.22% (3)

Precision, P= $\frac{TP}{TP+FP}$ = 51.47% (4)

Recall, R= $\frac{TP}{TP+FN}$ = 100% (5)

F score= $\frac{2 \times Precion \times Recall}{Precision+Recall}$ = 67.97% (6)

Null error rate= $\frac{TN+FP}{n}$ = 59.77% (7)

Overall agreement probability, Pe = $\frac{actual\ yes \times predicted\ yes}{n^2}$ + $\frac{actual\ no \times predicted\ no}{n^2}$ = 44.4% (8)

Cohen's Kappa, K = $\frac{Accuracy-Pe}{1-Pe}$ = 31.5% (Fair Agreement) (9)

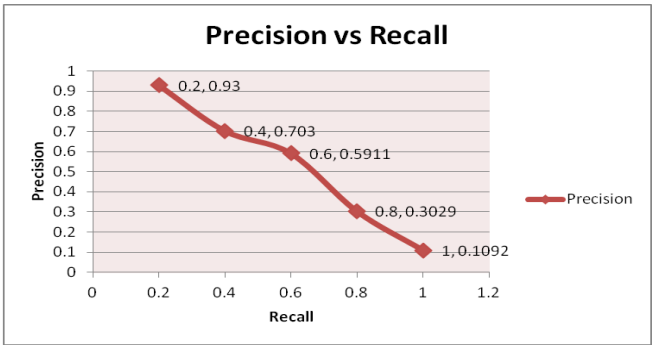


Fig. 14. Precision vs Recall on the event of Car accident.

TABLE II.
CONFUSION MATRIX TABLE FOR FOOT OVER BRIDGE

| | | | |
|---|---|---|---|
| TN: | 51 | FP: | 28 |
| FN: | 0 | TP: | 5 |

The total number n=84 and the total sum of individual row and column is 84. As, TN+FP=79 and FN+TP= 5. And sum of them is equal to 84. The sum of (TN+FN) + (FP+TP) = 84.

Accuracy, A= $\frac{TP+TN}{TP+FP+TN+FN}$ = 66.67% (10)

Prevalence = $\frac{FN+TP}{n}$ = 5.95% (11)

Precision, P= $\frac{TP}{TP+FP}$ = 15.15% (12)

Recall, R= $\frac{TP}{TP+FN}$ = 100% (13)

F score= $\frac{2 \times Precion \times Recall}{Precision+Recall}$ = 26.31% (14)

Null error rate= $\frac{TN+FP}{n}$ = 90.80% (15)

Overall agreement probability, Pe = $\frac{actual\ yes \times predicted\ yes}{n^2}$ + $\frac{actual\ no \times predicted\ no}{n^2}$ = 0.594 =59.44% (16)

Cohen's Kappa, K = $\frac{Accuracy - Pe}{1 - Pe}$ = 17.83%(Slight Agreement) (17)

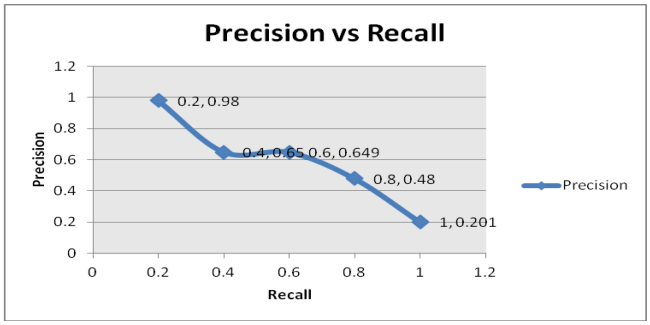


Fig. 15. Precision vs Recall on the event of Foot over bridge.

TABLE III.
CONFUSION MATRIX TABLE FOR ZEBRA CROSSING

| | | | |
|---|---|---|---|
| TN: | 13 | FP: | 21 |
| FN: | 42 | TP: | 95 |

The total number n=171 and the total sum of individual row and column is 171. As, TN+FP=34 and FN+TP= 137. And sum of them is equal to 171. The sum of (TN+FN) + (FP+TP) = 171.

Accuracy, A= $\frac{TP+TN}{TP+FP+TN+FN}$ = 63.16% (18)

Prevalence = $\frac{FN+TP}{n}$ = 80.11% (19)

Precision, P= $\frac{TP}{TP+FP}$ = 82% (20)

Recall, R= $\frac{TP}{TP+FN}$ = 69% (21)

F score= $\frac{2 \times Precion \times Recall}{Precision + Recall}$ = 75% (22)

Null error rate= $\frac{TN+FP}{n}$ = 19.88% (23)

Overall agreement probability, Pe = $\frac{actual\ yes \times predicted\ yes}{n^2}$ + $\frac{actual\ no \times predicted\ no}{n^2}$ = 60.74% (24)

Cohen's Kappa, K = $\frac{Accuracy - Pe}{1 - Pe}$ = 6.164% (Slight Agreement) (25)

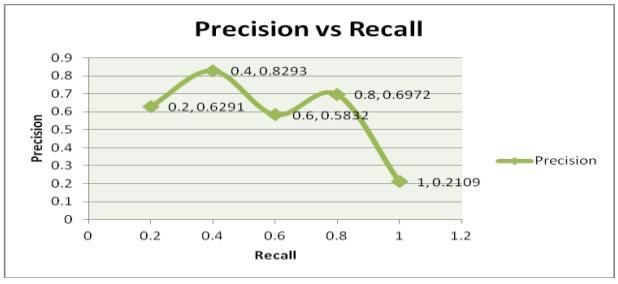


Fig. 16. Precision vs Recall on the event of Zebra Crossing.

## C. IoU Accuracy

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \quad (26)$$

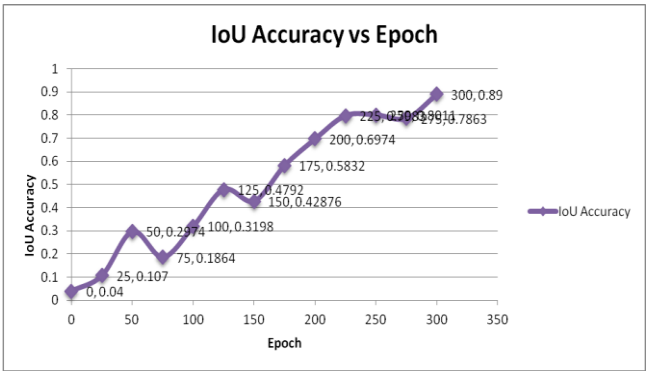


Fig. 17. IoU Accuracy vs Epoch on the event of foot over bridge.

The average IoU gained for the event of foot over is near about 89% for the dataset of foot over bridge.

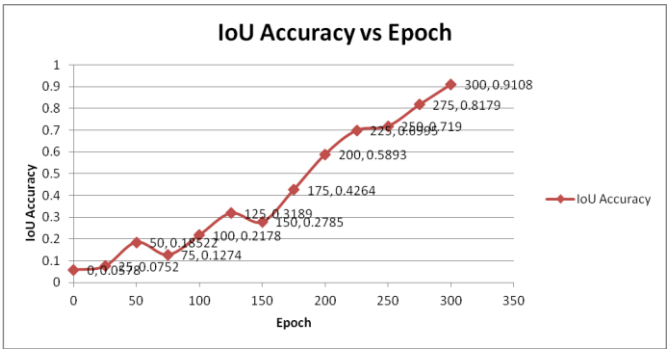


Fig. 18. IoU Accuracy vs Epoch on the event of car accident.

The average IoU gained for the event of car accident is near about 72% for the custom made dataset which is analysed to gain higher IoU accuracy.

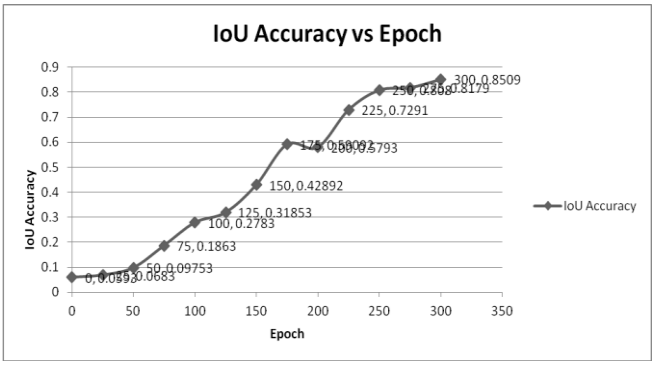


Fig. 19. IoU Accuracy vs Epoch on the event of zebra crossing.

The average IoU gained for the event of zebra crossing is near about 72% for the custom made dataset.

## V. Conclusion

A vast amount of experiment in this paper is done by giving efficient and robust analysis of the detection and recognition of foot over bridge crossing, zebra crossing, occurrence of car accident to detect the wounded person and also the information of the car owner by detecting the Bangla number plate. A custom made dataset is analysed to detect and recognized those events. This Yolo v2 model under the DarkFlow framework processed the whole dataset to achieve higher IoU, Precison rate and Recall rate.The YOLOv2 architecture is higher efficient to detect those events accuratly by ensuring the utmost safety to the people of Bangladesh.

## References


[1] Zhu, C., & Yang, Y. (2018, September). Face Detection and Recognition Based on Deep Learning in the Monitoring Environment. In *International Conference of Pioneering Computer Scientists, Engineers and Educators* (pp. 698-705). Springer, Singapore.

[2] Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior research methods, instruments, & computers*, *32*(1), 93-110.

[3] Ko, K. E., & Sim, K. B. (2018). Deep convolutional framework for abnormal behavior detection in a smart surveillance system using YOLO. *Engineering Applications of Artificial Intelligence*, *67*, 226-234.

[4] Thiel, G. (2000). Automatic CCTV surveillance-towards the VIRTUAL GUARD. *IEEE Aerospace and Electronic Systems Magazine*,*15*(7),3-9.

[5] Sang, J., Wu, Z., Guo, P., Hu, H., Xiang, H., Zhang, Q., & Cai, B. (2018). An improved YOLOv2 for vehicle detection. *Sensors*, *18*(12), 4272

[6] Nakahara, H., Shimoda, M., & Sato, S. (2018, August). A Demonstration of FPGA-Based You Only Look Once Version2 (YOLOv2). In *2018 28th International Conference on Field Programmable Logic and Applications (FPL)* (pp. 457-4571). IEEE.

[7] Wang, L., Li, W., Zhang, Y., & Wei, C. (2017, October). Pedestrian detection based on YOLOv2 with skip structure in underground coal mine. In *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)* (pp. 1216-1220). IEEE.

[8] Wang, L., Yang, S., Yang, S., Zhao, C., Tian, G., Gao, Y., ... & Lu, Y. (2019). Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World journal of surgical oncology*, *17*(1), 12.

[9] Kadam, S., Hatalge, A., Balip, A., & Powar, A. (2019). Freight Analysis Using YOLOv2. *Available at SSRN 3420232*.

[10] Li, J., Ge, H., Zhang, Z., Wang, W., & Yang, Y. Traffic Multiple Target Detection on YOL.

[11] Yang, S., Bo, C., Zhang, J., & Wang, M. (2020). Vehicle Logo Detection Based on Modified YOLOv2. In *2nd EAI International Conference on Robotic Sensor Networks* (pp. 75-86). Springer, Cham.

[12] Liu, Z., Shi, Y., & Sun, M. (2018, November). A pedestrian detection algorithm based on improved YOLOv2. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 488-492). IEEE.

[13] Zhu, H., & Zhang, C. (2018, October). Real-time traffic sign detection based on YOLOv2. In *2018 International Conference on Image and Video Processing, and Artificial Intelligence* (Vol. 10836, p. 108361B). International Society for Optics and Photonics.

[14] Chen, J., Ni, Z., & Sang, N. (2018, November). Multi-Scale YOLOv2 for Hand Detection in Complex Scenes. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (pp. 1525-1530). IEEE.



**Anjir Ahmed Chowdhury** received Bachelor of Science in Computer Engineering (CoE) in 2019 & currently doing Master of Science in Computer Science (CS) from American International University-Bangladesh. His research interests focus on Artificial Intelligent.



**Sabrina Kashem Chowdhury** received Bachelor of Science in Computer Engineering from American International University-Bangladesh. Her research interest is Network Security and Cloud Computing.



**Md. Hanif** was born on 10th July, 1998 in Dinajpur, Bangladesh. He received his Bachelor of Science (B.Sc.) in Computer Engineering from American International University-Bangladesh. His research interest is focused on Artificial Intelligence.



**Sadia Noor Nosheen** received the Bachelor of Science (B.Sc.) in Computer Engineering from American International University-Bangladesh in 2019. Her research interest is focuses on Networking.



**Md. Saniat Rahman Zishan** received B.Sc. in Electrical and Electronic Engineering and Master of Engineering in Telecommunications degree from American International University-Bangladesh (AIUB). On September 2009, he started his teaching career as a lecturer in AIUB. At present he is serving as an Associate Professor at the Department of Electrical and Electronic Engineering (EEE) & Computer Engineering (CoE) of AIUB. He is also serving as the Head of CoE Department. He is enrolled for PhD Degree at Universiti Sultan Zainal Abidin,Malaysia. Mr. Zishan is a member of the Institute of Electrical and Electronics Engineers (IEEE) and Institution of Engineers, Bangladesh (IEB). His current research interest includes Wireless Communication, Signal Processing, E-Health System, Telemedicine, Robotics and AI.

BLANK