International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I2P226
Volume 69 Issue 2, 188-193, February 2021

Modeling Abandoned Object Detection And Recognition In Real-Time Surveillance

WWWWTWKMRNDB Weliwita, JAP Isuru and SC Premaratne

Faculty of Information Technology,
University of Moratuwa, Katubedda, Moratuwa, Sri Lanka.
nilnuwan19811@gmail.com, jayasooriyaisuru@gmail.com, samindap@uom.lk

Abstract: *This paper presents the model's accuracy of abandoned object detection and recognition in real-time surveillance. There basically focused three models call Faster Region Convolutional Neural Network (Faster RCNN), Single Shot Multiple Detector (SSD), and You Only Look Once Version 3 (YOLO-Version 3). The research tested under MXnet Framework used the GluonCV Library for object detection, OpenCV used for pre-processing, and other preliminary adjustments of captured video inputs in the Python 3.8 Platform. The objectives of the research listed as acquiring knowledge on abandoned object detection, algorithms, different frameworks, and neural network. Identifying significant parameters, determining accuracy performances of the different models, and finalizing the most accurate model in real-time recognition and detection of an object. The research focused on the use of practical readings and calculation of the accuracy from 'Confution Matrix.' It suffices to obtain the maximum accurate results of each model separately. Python program used to obtain the input videos to decide the abundancy very sensitively. Then those reading tested and received the results with percentage value to decide the accuracy. Finally, the Confusion Matrix could be able to provide the results separately. Those results revealed that YOLO-V3 gave the most accurate results; secondly, SSD and third place goes to Faster RCNN.*

Keywords: *Faster RCNN, SSD, YOLO-V3, MXnet, GluonCV, OpenCV, and Python 3.8.*

I. Introduction

The fear of "Terrorism" again raised with the Easter Sunday day attack in Sri Lanka almost after one decade of ruthless experience of the civil war in the country. This caused a huge impact on finding technical solutions to detect such activities beforehand the disasters. Therefore, government and non-government authorities kept their maximum effort to find solutions to identify such incidents in advance to eradicate future calamities in the state. The use of Close Circuit Television (CCTV) was done during the last two-three decades even more than that. But those were monitored in manual operational rooms by men, and further, those records were used to post investigations of any incident that occurred only. Expertise looked on this aspect and thought to use the resources more efficient way to obtain real-time solutions rather than the post-incident investigations.

The use of a different number of neural network frameworks, compatible number of their algorithms, tools, models, and libraries …etc. Were deeply used last year in different countries on different platforms to improve the concept to overcome the barriers. Eventually, those researchers have developed lots of different successful solutions in different directions such as face recognition, suspicious activities of humans and animals, abandoned object detection and object detections …etc. Out of those, most of the research focused on using neural network platforms with image processing techniques to obtain suspicious activities and suspicious objects, which can cause a disaster. This paper focused on abandoned object detection and recognition in real-time surveillance. Most of the researchers have used the Tensorflow framework to do their researches. In our research, we focus on the MXnet framework because, most of the time, lots of other libraries, tools, and datasets used during the time of research compatible with the aforesaid framework. It was really easier to continue the research; not only that but the results also so amazing and accurate.

The research selected three models to determine the accuracy of those models of Faster RCNN, SSD, and YOLO-V3[1]. This highly important to use in the real-time environment because the accuracy must be 100%; otherwise, the wrong reading or the results will make wrong actions then it will badly affect every aspect. Therefore, the finding of the accuracy of the particular model is very much important. The background and foreground subtraction[2] are used to grab the images of the particular video input during the image subtraction process. The evolution of computer vision technology the segmentation, classification, and pose estimation [3-7] eventually done with the development of modern science and technology.

II. Related Works

Real-time detection and the recognition of an abandoned object and defining a static object as an abandoned is critical [2] study. The identification of a piece of abandoned luggage in real-time surveillance is done by the background and foreground modeling. Further examined whether the candidate region contained the abandoned object by analyzing the back-traced trajectories of luggage owners. The researchers have used that the PETS2006 and AVSS2007 datasets during the research. Instead of using the robust automated system rather than using the old concepts of manual monitoring of the CCTV

camera footages researchers have focused their attention on the interest of deep learning neural network models [8]such as Faster Regional Convolutional Neural Network(Faster RCNN), Single Shot Multibox Detector(SSD) and You Only Look Once Version 3 (YOLO-V3).

Wentong Liao and et al. [9] explained that the use of security-based implementation of any application or concept needs to have accuracy in three aspects. Those are; the accuracy of input data, the accuracy of the output results, and the accuracy of the detection. These aspects needed to clarify before getting into the further steps; otherwise, everything will be a waste. Researchers have found certain difficulties that obstruct obtaining maximum successful outcomes, such as illumination changes, shadows, and the high density of the moving objects. Aditya Gupta and et al. [10] explained how to do the dual background subtraction method to find out-static objects. Researches assumed in the research that a particular object is a part of the foreground for a long time, then eventually turned in to the background, finally the background modeling done with approximate median modeling. The research used the PETS2006 dataset for testing the algorithms. [11]explained that the most important factors, especially in the Temporally Static Objects (TSO), in robust detection of the abandoned object detection and recognition. The research is basically based on the use of Finite State Machine (FSM) through the technique of Background modeling. During the research was disturbed its performances due to lighting changes, occlusions and the cluttered backgrounds.

Another different angle of research [12] is illustrated by the use of a novel adaptation of the Binocular Information Reconstruction and Recognition (BIRR) algorithm. It was based on the road traffic environment to minimize road accidents. There were initial primary input videos from the street monocular cameras then through the foreground segmentation algorithm to detect the recognition. After that, those identified input like the road plain equation and the height of the abandoned objects transferred to the aforesaid three-dimensional (3D) algorithm. Then could be able to finalize the decision properly.[13] The robust foreground and abandonment analysis for large-scale abandoned object detection in complex surveillance videos in the more crowded area is explained properly. The use of Confusion Matrix is basically used to finalize the background and foreground modeling readings input, whether abandoned or not. During the process, researchers have tried to minimize the false-positive outcome and improved the detection accuracy. The lighting changes, low textures, low contrast and the cluttered background were disturbed to obtain best results during the research.

[14] Illustrated the use of Gaussian Mixture Model (GMM) detection of an abandoned object in the high ways. This model could not be able to update every frame for keeping the abandoned objects in the foreground. Researchers had used an edge statistic feature-based approach to minimize the noise from sunshine and the wind. Not only that object tracking model is also integrated into the model. Finally, the expectation could not reach the desired goals, but it could obtain high accuracy results for practical applications.

II. System Overview

The modeling of the abandoned object detection and recognition in real-time surveillance has been tasked and organized according to the following figure 3.1 of the activity diagram. It has explained clearly and deeply step-by-step functions very clearly.

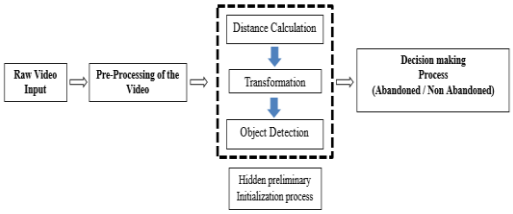


Figure 3. 1 An activity Diagram

From the beginning, the clear input CCTV camera footages have been taken through the installed static surveillance cameras. Then, the preliminary pre-processing of the input video was done using an algorithm before entering the next step of the hidden preliminary initialization process. Under the third step, the input signal may go through the distance calculation, Transformation, and object detection processes eventually. Finally, the object detection and recognition process are completed, then it needs to be decided whether the particular object is abandoned. This has been planned to do in the last stage. The whole process runs under the python 3.8 platform. It has been written to each, and every input video needs to be tested from three models (Faster RCNN, SSD, and YOLO-V3) separately by keeping the other parameters stable.

IV. Implementation

The whole research program is written on the Python 3.8 platform and MXnet neural network framework run on the python platform. MXnet contains the OpenCV and GluonCV libraries. MXnet is comparatively fast and more accurate in the Graphical Processing Unit (GPU) and Central Processing Unit (CPU) environments. Initially, the acquired raw input video clips from the static CCTV video cameras have absorbed by the written python program into the MXnet framework. Indicated in figure 4.1 Input video program.

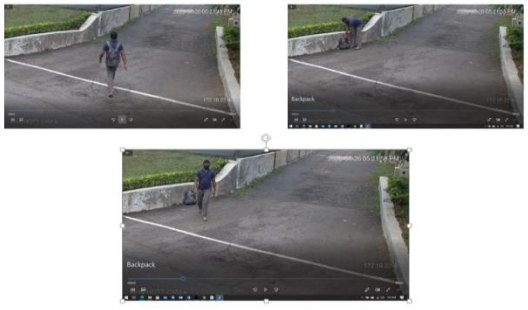


Figure 4. 1 Input video program

Figure 4. 2 Input video samples

Once the program runs, the system may absorb the following sample input video clips to the program illustrate under figure 4.2 input video samples.

A. Pre-Processing of the Input Video clips

Secondly, absorbed input video clips adjustment of the color transformation from BGR to RGB and the taking to birds' view of the images occurred during this step. Figure 4.1.1 pre-processing of the input video is illustrated as follows.

Figure 4.1.1 Pre-Processing of the input video

B. Distance Calculation

The research has planned to do the accuracy among the models to obtain a successful outcome; therefore, keeping other parameters such as the Threshold value, Minimum distance, and frame size constant in each video input allows to do the test correctly. The following source code figure 4.2.1 distance calculation illustrates the function of the process.

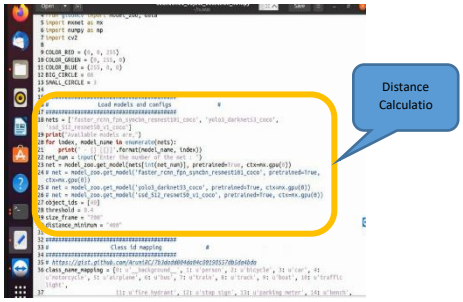


Figure 4.2.1 Distance calculation

C. Transformation

The OpenCV handled the transformation of the parameters such as corner points, matrix, width, image and list download. The 'numpy' and 'cv2' packages used to the transformation of the aforesaid parameters. Further, 'transform.py' and 'cv2.getperspectivetransform' are used to proper transformation of the corner points of the images as shown in the source code of the figure 4.3.1 transformation of the corner points of images.

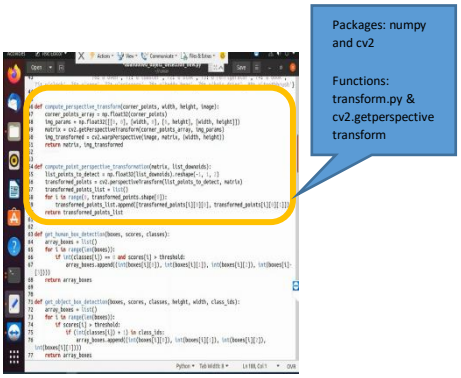


Figure 4.3. 1 Transformation of the images and corner points.

The numpy used to numerical processing and cv2 used to OpenCV binding. cv2.getperspectivetransport used to obtain the maximum accurate image.

D. Object Detection

From the beginning to until shaping the input video according to the required styles and endo of the third stage of transformation, an operation occurs to create, and an operation occurs to create an image that is the same as the original input image. Then activate the recognition and detection task to be done. These created images may be compared and used with the GluonCV model zoo, and pre-trained datasets finally created the output video with indicating the proper detection and non-detection according to the following figure 4.4.1 (a) and (b) of object detection and recognition source codes.

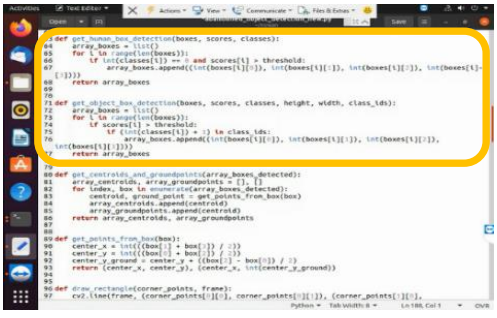


Figure 4.4. 1 (a) Object detection and recognition

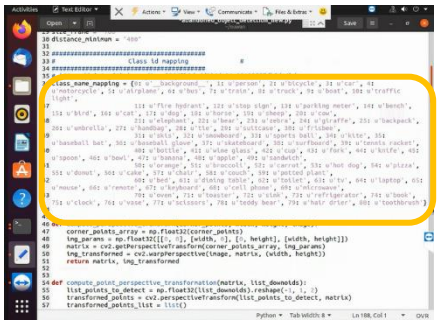
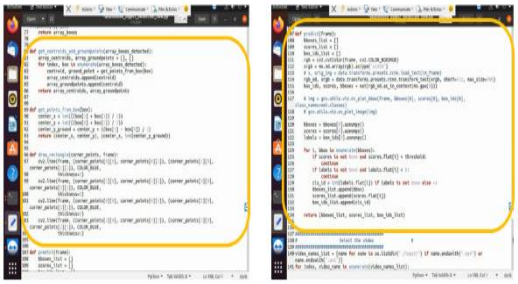


Figure 4.4.1 (b) Class variable

E. Decision Making Process (Abandoned / Non-Abandoned)

During the previous steps, the object's detection was done correctly, then under this stage, needed to identify or decide the detected or recognized object whether abandoned or not. Then program had directed to produce the output video by indicating if the object is without been particular distance away from the owner, then the subject-object bounding with a red color box which indicates that the item or the object is abandoned or if the object and the owner both within the predefined distance then that object considered as the non abandoned then the out video will cover the particular object with blue color bounding box which means subjected object consider as nonabandoned. During the process, the GluonCV toolkit directly dealt with the MXnet framework to obtain the desired results. The following figures of 4.5.1 illustrate the particular source codes which relate to the decision-making process.



(a) (b)

Figure 4.5.1 Decision-making source code (abandoned /non abandoned)

The following sample video images may evidence to illustrate the abandoned and non abandoned objects, which relate to the input video samples images as figure 4.5.2 of input/output video clips indiction of abandoned and non abandoned.

Figure 4.5.2 Abandoned and non abandoned detection of input and resulted in output videos

V. Evaluation

The implementation occurred during the whole execution and activation process, especially during the distance calculation process. Actually, distance calculation happened then it happened in the three-dimensional plane but in the CCTV images, calculate the distances in a two-dimensional plane. There we needed to do the maximum accuracy of the conversion. This was done by the use of the 'Equiladian Equation.' The following figure 5.1 may explain the simple process of calculation clearly. This caused to transfer of the three-dimensional distance calculation to the bird's eye view of the two-dimensional plance with maximum fewer errors.

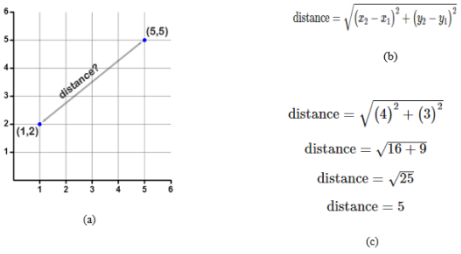


Figure 5.1 The Equiladian Equation for distance calculation

Basically, the abandoncy of the recognized object was calculated by the camera screen pixels. Therefore, the owner and the object's distance initially may have decided as such. If the particular distance exceeds both objects (owner and the object), it will be considered abandoned, and if both of them were less than the predefined distance as the nonabandoned. The calculation of the distance was done by the Pixels. The research has decided to keep the 400-pixel value as the static distance between the owner and the object. Once the limit exceeds or the owner goes out from the object keeping the object alone, the object would be indicated in the ‘Red’ 'Red' color-bounding box. Object and owner less than the said distance the object may indicate in ‘Green’ 'Green' color-bounding box. The following figure 5.2 Distance calculation pice of source code illustrated the calculation of the distance and instructions.

Figure 5.2 Pixel distance calculation

The other important point was once any video input tested through any model, and we needed to keep the threshold value stable or static for the three models; otherwise, we can not calculate the accuracy reasonably among those three models.

The lengthy process can be explained simply step by step as follows;

(a). Once the whole video clip runs through any model particular model, suppose to cut the whole video input into one-second video slots.

(b). Then, above one second time slot, each video clip is used to create five photo frames each and every second video input, and those created five frames in folders separately.

(c). Meantime, above step (b) occurs same video slot send through the algorithm to mark the decided bounding box (if abandoned 'Red' if not 'Green') and the same manner then freeze to folders which contains the five photos of the one-second video clip.

(d). Finally, the source code may create input one second of five photo frames include folders call as 'Ground truth folders' other folder call as 'Truth' contained with the result colored bounding boxes.

Following figure 5.3, Ground truth and Result folders may illustrate the above steps clearly.

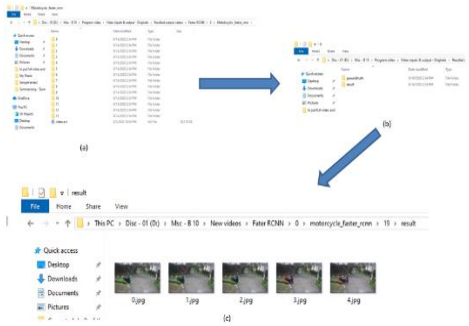


Figure 5.3 Ground truth and Result folders

A. Calculation

The assigning the binary values for the input (Ground truth binary values) and output (Result binary values) may be assigned as follows.

(a). If the Ground Truth (GT) / Result (R) each folder's photo frame have three (3) out of five (5) same indication (as an example, if any box contains the majority of Green color bounding boxes, then that consider as the nonabandoned but if it is red colored then considered as an abandoned).

(b). When the decision is taken according to the above (a), binary value = 1 (one) assign to the identification of an abandoned, then binary value = 0 (zero) assign to the identification of a nonabandoned decision.

(c). The research has planned to conduct with three input videos, and those three have different threshold values but each video tested under static threshold value from three different models. Then only can we measure the accuracy reasonably?

(1).Bike video – threshold value 0.004
(2).Bottel video – threshold value 0.06
(3).Laptop video – threshold value 0.39

(d). The calculation of the models' accuracy may have done by using the 'confution matrix' this equation basically use to calculate the binary values. The following figure 5.1.1 indicates the function of the confusion matrix.

The factors of the equation are as follows;

	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

- TP- True Positive
- FN – False Negative
- FP – False Positive
- TN- True Negative

Figure 5.1.1 Confusion matrix

The calculation of the accuracy as follows;

The accuracy= [(TP + TN) / (TP + FP + FN + TN)] x 100%

Except for calculation of the accuracy, it gives the facility of calculation of 'Precision' and 'Recall.' Those equations are as follows

Precision = [(TP) / (TP + FP)] x 100%

Recall = [(TP) / (TP + FN)] x 100%

Then all the parameters and metrics have been streamlined. The steps for measuring the results are as follows.

Example
Video – Motorbike (Threshold value 0.004)
Model – Faster RCNN

Total readings = 20 nos
TP= 5, TN=12, FP=1, FN=2
Then
Precicion= [(5) / (5 + 1)] x 100% = 83.33%

Recall = [(5) / (5 + 2)] x 100% = 71.40%

Accuracy = [(5 + 12) / (5 + 1 + 1 + 13)] x 100%
= 85%

The following figure 5.1.2 shows the results of the three models individually as follows.

Ser No	Video input Description	MODELS								
		Faster RCNN			YOLO – V3			SSD		
		P(%)	R(%)	A(%)	P(%)	R(%)	A(%)	P(%)	R(%)	A(%)
01	Motor bike	83.3	71.4	85	100	71.4	90	83.3	71.4	85
02	Bottle	61.5	80	61.1	100	80	88.8	72.7	80	72.2
03	Laptop	60	100	78.9	85.7	100	97.7	75	100	89.5

P – Precision R – Recall A - Acurracy

Figure 5.1.2 Results of the models

According to the calculations above, figure 5.1.2 illustrated the results of those three models separately. Out of those results, the maximum accuracy acquired by the YOLO-V3 for all the three video inputs secondly, SSD and thirdly Faster RCNN have performed.

VI. Conclusion and Future Works

This research mainly focused on the modeling of the abandoned object detection and recognition in real-time surveillance. Their research was mainly focused on three main models, which are mostly used in the field. Furthermost of the time, researchers used to use Tensorflow as the Framework, but in this research used MXnet because it contains other libraries, toolkits, datasets, etc. therefore, the operation and continuation of the process was quite easier and free.

The literature really helped to research to find the gaps in the subject area to find the solutions through the research. Further, obtaining the ground truth input reading was taken closely as a nacked eye, but if a further deep method to take input readings, the accuracy may be more perfect than this level. There were lots of limitations like the CCTV camera performances acquiring angles of the readings and the power of videos' quality cause some kind of limitation to obtain the readings except natural difficulties such as lighting effects and the shaking due to winds.

This research had focused on and evaluated the accuracy of the three models. This function is using in the real
-time operation and object detection has very little time to finalize the abundance; therefore, accuracy is very much important. To take a decision is 100% correct. Another important factor is the speed because, for security purposes, even those the decision accuracy though the 100% the time taken to finalize the result then no point because the degree of surprise may have gone due to huge time consumption to take the decision. Therefore, this abundance recognition must be very quick; otherwise no point in taking late actions. Therefore a future work needs to find the speed performances of the models separately with considering more parameters.

References

[1] Rohith Sri Sai, M., S. Rella, and S. Veeravalli, OBJECT DETECTION AND IDENTIFICATION A Project Report. (2019).

[2] Tripathi, R.K., A.S. Jalal, and C. Bhatnagar. A framework for abandoned object detection from video surveillance. in 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). (2013). IEEE.

[3] Girshick, R., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2014).

[4] Girshick, R., et al., Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence, **38**(1)(2015) 142-158.

[5] Krizhevsky, A., I. Sutskever, and G.E. Hinton, Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017. **60**(6)(2017) 84-90.

[6] Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015).

[7] Toshev, A. and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2014).

[8] Lin, K. et al., Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance. IEEE Transactions on Information Forensics and Security, **10**(7)(2015) 1359-1370.

[9] Liao, W., et al., Security event recognition for visual surveillance. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4 (2017), Nr. 1W1, **4**(1W1)(2017) 19-26.

[10] Gupta, A., et al. Real-Time Abandoned Object Detection Using Video Surveillance. in Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing. (2016) Springer.

[11] Fan, Q. and S. Pankanti. Modeling of temporarily static objects for robust abandoned object detection in urban surveillance. in 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). (2011) IEEE.

[12] Zeng, Y. et al., A novel abandoned object detection system based on three-dimensional image information. Sensors, **15**(3)(2015) 6885-6904.

[13] Fan, Q. and S. Pankanti. Robust foreground and abandonment analysis for large-scale abandoned object detection in complex surveillance videos. in 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance. (2012) IEEE.

[14] Fu, H., et al. Abandoned object detection in highway scene. in 6th International Conference on Pervasive Computing and Applications. (2011) IEEE.