Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221322

Bank Robbery Detection System Using Computer Vision

Avinash Kumar Sharma, Pratiyaksha Mittal, Ritik Ranjan, Rishabh Chaturvedi
(Department of Computer Science and Engineering, ABES Institute of Technology, Affiliated to AKTU, Lucknow)

Abstract. We propose a computer-vision-based detection and recognition system which can be used in the banks to detect the anomalous activity of bank robbery. We propose CCTV based robbery detection along with tracking of thieves. We have used computer vision to detect theft and robbers in CCTV footage, without the use of sensors. This system concentrates on object detection. This detection of bank robbery is done based on detecting components, which are prohibited by using in the banks and are common in robbery, like handguns, a person wearing a helmet or a ski mask which comes under the object detection. Apart from these, recognition is done on the human postures like raising hands and kneel down which comes under the posture detection. The security official will be notified about the suspicious event by using Real-time analysis of the movement of any human from CCTV footage and thus gives a chance to avert the same, so that necessary action will be taken by the authority and prevent threat to bank as well as to the human life presents there.

Keywords – Yolo, OpenPose, Object Detection, Pose Estimation

1. Introduction

The crime in the finance sector has increased vastly in recent years and bank robbery is one of those crimes. The number of bank robbery gets increased every year and so does the amount lose by the banks. During Bank robbery employees and customers are subjected to violence, force, or threat of violence putting human lives at risk. During bank robbery guards or any other employee are kept under threat and hence they cannot contact the police to take appropriate action on time. Therefore, this project is made to detect the action of bank robbery with the help of surveillance video cameras. This project detects the presence of the substances used in a robbery like a gun, a person wearing a helmet, or a ski-mask. It also detects situations where people are forced to raise their hands or make them kneel down. These detections ensure that the situation in which these detections have been made is similar to or exactly a bank robbery. We have chosen the objects for the detection that are not commonly allowed inside the banks and also, they are very popular and common in bank robberies. Every object or feature used for detection will be given some weightage which on detection gets summed up and when this sum reaches a threshold value, the model will confirm that robbery is detected. The person raising their hands or kneeling are also used in the project but they are used as additional features to the detection. As, these are complex to detect due to various reasons and may give false positives, and therefore, they will

have less weightage compared to other object detections. In this project, we have used YOLO (You Only Look Once) (Joseph Redmon S. D., 2016) for object detection and OpenPose for pose recognition.

2. Event Detection

The event detection model works as follows. First, the model is fed frame by frame by the camera. These frames work as an input to the detection model. Then the model performs object detection and pose estimation algorithm on the frame, to detect the required conditions for the robbery.

A. Object detection within images

Analysis of the frames of images begin with the object detection model. There are numerous object detection algorithms in deep gaining knowledge of along with R-CNN, FasterRCNN, and single Shot Detector. Here, Figure 1 is given for the reference of the time taken to analyze one frame of image by different object detection model. We can clearly see that YOLO performs the best. As, YOLO (You best look once) is one of the fastest algorithms to locate gadgets in real-time, we've used YOLOv3 (You appearance most effective once model three) (Joseph Redmon A. F., 2018) on this task.

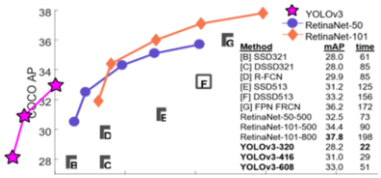

Figure 1. We have adapted this image from the Focal Loss Paper [6].

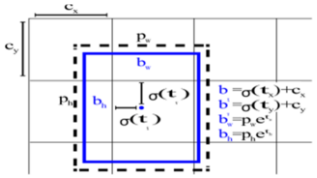

Figure 2. Bounding boxes with dimension priors and location prediction[9]

YOLO divides the body of the photo into areas and predicts bounding containers and probabilities for each location of the frame. Then, these bounding bins are weighted via the predicted probabilities.

YOLOv3 predicts 3 bounding containers for every mobile [2]. The network predicts four coordinates for each bounding box by the formulation (Farhadi, 2017):

bx = σ(tx) + cx

by = σ(ty) +cy

bw = pwet

bh = heth

Right here, bx, by way of are the x, y middle coordinates, bh, and bw are the width and peak of the prediction. The objects that require to be identified (firearms and knives) can also be in extraordinary orientations in the video footages available. So, the device wishes to gain knowledge of. with multiple feasible orientations of the unique items. To ensure that the object is detected even in variable distances we are able to want to exchange the scale of the sliding window as a consequence a good way to boom the general computation. The center coordinates are then run thru logistic regression to expect the objectness rating. This offers output as 1 if the bounding box

prior overlaps the ground fact object by way of more bounding bins else 0. Objectness score represents the opportunity that an item is contained internal a bounding container. Then multilabel classification is completed on those packing containers to predict the elegance that box may also comprise.

B. Human pose recognition in the image

We are performing Human 2D pose estimation to recognize the complex poses like people raising hands and people kneeling down. These poses act as a feature in this robbery detection project.

This is a very difficult problem. This is because there are various components such as small, barely perceptible details, large discrepancies in occlusion and pronunciation. A classic approach to assessing speaking posture is to use a picture structure system. The main idea here is to answer the question with a set of "parts" made up of highly deformable (non-flexible) settings. A "part" is an image format that matches an image. Posture assessment means determining the position and orientation of the body. This posture estimation is done by detecting key points in the human body. To discover these key points, we used an OpenPose model trained on the COCO dataset, which provides 18 human body key points.

These points of the numbered body are:

Nose – 0, Neck-1, Right Shoulder- 2, Right Elbow - 3, Right wrist -4, Left Shoulder -5, Left Elbow - 6, Left wrist–7, Right hip - 8, Right knee - 9, Right ankle - 10, Left hip - 11, Left knee - 12, Left ankle - 13, Right Eye - 14, Left Eye - 15, Right ear - 16, Left ear - 17, Background – 18.

The model takes an h x w color image as input and outputs a matrix containing the key point confidence map and partial similarity (PAF) for each key point pair. First, the image is passed through 10 Convolutional layers of VGGNet to create feature maps of the image. Then it is passed through a 2-branch Multi-stage CNN, where the first branch predicts a set of 2D confidence maps of the body part location and the second branch predicts the 2D vector of PAF (Part Affinities). The output contains 57 matrices in which the first 19 matrices are confidence maps of the key points and the rest matrices are the PAF matrices. A confidence map is a grayscale image in which every pixel represents a probability of a certain key point. It will give a very high value at pixel where the likelihood of a certain body part is high.

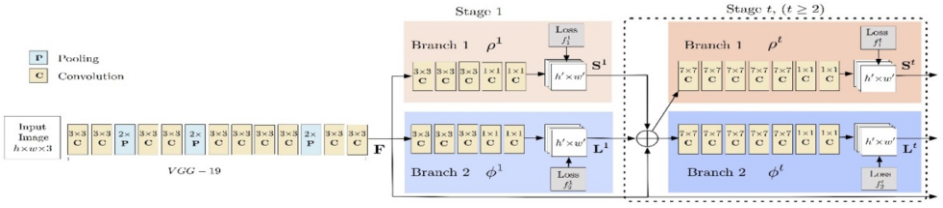

Figure 3. OpenPose Architecture [5].

There may be many people in the image therefore detecting only the key points is not enough as it may be difficult to understand which key points belong to whom. To solve this issue, PAF (Part Affinity) is calculated for each key point pair. PAF represents which key point belongs to which key point pair and ultimately to which person. For

example, A key point pair of neck and shoulder is taken, it will calculate the affinity of each shoulder in the image with the neck but only the neck which belongs to the same person will give high affinity and then it is considered as a pair.

Confidence Map: Confidence maps are a 2D representation of the idea that body parts are often located at specific pixels in a frame.

Part Affinity Fields: Part Affinity can be a set of 2D vector fields that encode the positions and orientations of various human limbs in an image. It encodes information in the way of pair-wise junctions between parts of the body.

Figure 4. PAF of neck and left shoulder.

We have used these key points to recognize and detect human poses in our project. Specifically, we detected people raising their hands using their left wrist, left elbow, right wrist, and right elbow. And the right knee, right ankle, left knee, and left ankle detect who is kneeling.

OpenPose has trouble estimating pose when the image in the frame is upside down, so you need to make sure the camera position is correct to capture the image. In high-load images where images of people in the frame overlap, the model tends to merge comments from multiple people, and overlapping PAFs cause greedy parsing of multiple people to fail, sometimes skipping others.

For detection of raising hands, we've taken coordinates of both the wrists and elbows. If the y-axis of the elbow is on top of y-axis of wrist, then the person is taken into account as raising a hand. this is often done on both the hands and if found this condition true on elbow-wrist pair then only person is taken into account as raising both hands.

For detection of person kneeling down, we have taken coordinates of both the knees and both the ankles. Now we have calculated the angle between the ankle and the knee with the formula:

$\Theta = \tan{-1}\ ((y2 - y1) / (x2 - x1))$

If this theta is found to be less than 45 degrees, then it is considered as bending of knee or kneeling down. Unlike the raising hands detection, person will be considered kneeling down even if only leg is bending [4].

3. Experiment

In order to perform these type of detection accurate data is required. Since, no standard dataset for scenario presented in this paper was found, and real recordings from banks are confidential are not available to us, we have trained this model on the self-made dataset. Self-made dataset contains images of helmets, guns and ski-masks of different orientation and scale in the image.

Verification was made using the object detection and pose estimation models. We have fed the project with a small robbery clip hand with few images, it showed a very good performance on them. From these, we can say that the idea presented in this paper works as expected. Detection of harmful objects was possible in each of the frame of video. However, the threshold value had to be carefully tuned in order to prevent false detections.

How our algorithm works??

In the above portion, we have explained how the models which we have used in our project works and what are the procedure which is going on within those models. Now, we will be mentioning how we have made our bank robbery model using these models.
First, we require a camera from which we get the image feed. For our project we will be using the security cameras which are already installed in the banks. We will be taking the taking the frames from those cameras and will be doing analysis on them. Once the camera is live and starts taking the images, our model will retrieve them and feed them to the algorithm one by one for the analysis. First, the image will feed to the object detection algorithm to detect the harmful objects in the image and then the image will be passed to the pose detection algorithm and then the required output will be generated from the outs of these models.
Object detection model will provide a matrix consisting of the results after detection. First value in the matrix will be the total number of helmets and sci-masks detected in the frame. Second value will be the total number of the guns detected in the frame. We will be using these values to detect either robbery is happening or not. If the number of guns is 2 or more than 2 and there is anyone with helmet or a mask then it will give output as robbery detected else it will then feed the image for pose detection. When the image is fed to pose detection model, we will get coordinates of the body parts and with those coordinates we will identify the number of persons raising their hands and number of persons kneeling down. And then with these numbers and the weightage for these instances we will further decide whether the goons are present in the bank or not. We have done the detection of the robbery using this formula:

score = hndup/num*thresh + knldn/num*thresh + object [0]*thresh//3 + object[1]*thresh//2

Here, score is the total detection score,
thresh is the threshold value,
hndup is the total number of people raising their hands,
knldn is the total number of people kneeling down,
num is the total number of people present in the frame,
object [0] is the total number of masks and hekmets present,
object [1] is the total number of guns present.
When the value of score will be greater than the value of threshold then the system will say that it has detected a robbery in the bank.
Keeping in mind the situations where sometimes security guard can also come in the bank with his gun and sometimes some people can also come with a helmet on and there may be some few chances when some of the people may raise their hands and we don't want our projects to raise false alerts therefore we have given different weightage to different conditions and objects in this algorithm and based on those weightages total score of the model is calculated by doing the sum of all the values.

4. Conclusion

Bank Robbery detection system should follow a complete framework, starting from very first trace of movement to decision making, of the detection process. The system should comprise customer and employee behavior analysis, and object recognition. However, most of the papers have worked only on objects or only on the behavior. And most of the papers have only worked on the weapons ignoring the other objects which usually a robbery carries like helmet or wear a mask. Most of the papers did not deal with different postures which might restrict the detection of robbery.

In this paper, we have suggested an approach that can be used efficiently for the detection of robbery in the bank. We have used the images from the security camera installed in the banks and performed analysis on those images. The proposed method includes two models: Object detection model and human pose estimation model. Object detection model detects the presence of harmful objects in the frame, while the pose estimation model detects the pose of human body. If the object detection model detects the presence of harmful objects and pose this paper, then the event is said to be a robbery. We have kept in mind some of the main events and objects used in the robbery detection and we have performed detection and analysis based on those events and objects. But main limitation of this project is the diversity of the robbery means there may be chances where robbers can use some other weapons or some other kind of tactics which cannot be detected by this model so to improve our project there can be some other cases and other objects that we can include in this model. So, to conclude we can say that this model is a good model based on the robberies that happen and the objects that are usually used by the robbers but we can definitely increase its accuracy and we can definitely increase area to detect the robbery and we can also diversify the areas on which the analysis is based on we can also increase the objects that is we are doing analysis on and we can make this project more advanced in the future.

References

[1] Joseph Redmon, Ali Farhadi, YOLOv3 An Incremental Improvement.

[2] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You Look Only Once: Unified, RealTime Object Detection.

[3] Derui Wang, Chaoran Li, Sheng Wen, Qing-Long Han, Fellow, IEEE, Surya Nepal, Xiangyu Zhang, and Yang Xiang, Fellow, IEEE, Daedalus: Breaking Non-Maximum Suppression in Object Detection via Adversarial Examples.

[4] Zhe Cao, Student Member, IEEE, Gines Hidalgo, Student Member, IEEE, Tomas Simon, Shih-En Wei, and Yaser Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.

[5] Zhe Cao Tomas Simon Shih-En Wei Yaser Sheikh The Robotics Institute, Carnegie Mellon University, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. (Z. Cao, 2017)

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. ´ Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017. (T.-Y. Lin, 2017)

[7] Afzal Godil Roger Bostelman Will Shackleford Tsai Hong M Shneier Performance Metrics for Evaluating Object and Human Detection and Tracking Systems Published 2014.

[8] Michał Grega , Andrzej Matiolanski, Piotr Guzik and Mikołaj Leszczuk (2016). AGH University of Science and Technology, Article Automated Detection of Firearms and Knives in a CCTV Image.

[9] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6517–6525. IEEE, 2017.

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.

[11] Rutvik Kakadiya, Reuel Lemos, Sebin Mangalan AI Based Automatic Robbery/Theft Detection using Smart Surveillance in Banks.

[12] M. S. Munagekar, "Smart Surveillance system for theft detection using image processing", International Research Journal of Engineering and Technology. Aug.-2018

[13] "Comparison of YOLO v3, Faster R-CNN, and SSD for Real-Time Pill Identification" Lu Tan1, Tianran Huangfu1, Liyao Wu1, Wenying Chen1

[14] D. M. Dinama, Q. A'yun, A. D. Syahroni, I. Adji Sulistijono and A. Risnumawan, "Human Detection and Tracking on Surveillance Video Footage Using Convolutional Neural Networks," 2019 International Electronics Symposium (IES), 2019, pp. 534-538, doi: 10.1109/ELECSYM.2019.8901603.

[15] M. Naveenkumar and V. Ayyasamy, "OpenCV for Computer Vision Applications", Proceedings of National Conference on Big Data and Cloud Computing (NCBDC'15), pp. 52-56, March 2015.

[16] M. A. Vinith, G. Pradeep and B. Priya, "An Approach for Detecting and Identifying Suspected Weapons Using YOLO Algorithm," 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 478-480, doi: 10.1109/ICSPC51351.2021.9451686.

[17] M. Putra, Z. Yussof, K. Lim, and S. Salim, "Convolutional neural network for person and car detection using yolo framework," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-7, pp. 67–71, 2018.

[18] M. C. Thar, K. Z. N. Winn and N. Funabiki, "A Proposal of Yoga Pose Assessment Method Using Pose Detection for Self-Learning," 2019 International Conference on Advanced Information Technologies (ICAIT), 2019, pp. 137-142, doi: 10.1109/AITC.2019.8920892.