Gun Detection in Surveillance Videos using Deep Neural Networks

JunYi Lim*, Md Istiaque Al Jobayer†, Vishnu Monn Baskaran†, Joanne MunYee Lim*, KokSheik Wong†
and John See‡
*School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia
†School of IT, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia
‡Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
Email: *jun.lim@monash.edu, †mial2@student.monash.edu, †vishnu.monn@monash.edu,
*joanne.lim@monash.edu, †wong.koksheik@monash.edu, ‡johnsee@mmu.edu.my

***Abstract*—The ongoing epidemic of gun violence worldwide has compelled various agencies, businesses and consumers to deploy closed-circuit television (CCTV) surveillance cameras in attempt to combat this epidemic. An active-based CCTV system extends this platform to autonomously detect potential firearms within a video surveillance perspective. However, autonomously detecting a firearm across varying CCTV camera angles, depth and illumination represents an arduous task which has seen limited success using existing deep neural networks models. This challenge is in part due to the lack of available contextual hand gun information from CCTV images, which remains unresolved. As such, this paper introduces a novel large scale dataset of hand guns which were captured using a CCTV camera. This dataset serves to substantially improve the state-of-the-art in representation learning of hand guns within a surveillance perspective. The proposed dataset consist of 250 recorded CCTV videos with a total of 5500 images. Each annotated CCTV image realistically captures the presence of a hand gun under 1) varying outdoor and indoor conditions, and 2) different resolutions representing variable scales and depth of a gun relative to a cameras sensor. The proposed dataset is used to train a single-stage object detector using a multi-level feature pyramid network (i.e. M2Det). The trained network is then validated using images from the UCF crime video dataset which contains real-world gun violence. Experimental results indicate that the proposed dataset increases the average precision of gun detection at different scales by as much as 18% when compared to existing approaches in firearms detection.**

***Index Terms*—active video surveillance, gun detection, deep neural networks**

I. Introduction

The present work is motivated by two recent and conflicting trends: 1) the prevalence of gun crime in nations such as the United States (as evidenced by the APHA report [1] and the American Journal of Medicine [2]), and 2) the increased presence of closed-circuit television (CCTV) video surveillance systems globally [3]. These two trends are at odds with each other since video surveillance is supposed to deter crimes. A typical passive based CCTV system requires a human operator to monitor multiple cameras, at times requiring one operator overseeing 78 cameras simultaneously [4]. Sociological studies have indicated that operators typically suffer from human failures, with some operators playing hide-and-seek with patrolmen on the ground [5]. Taken together, these observations suggest a need for an active based video surveillance system with automated firearm (or handgun) detection algorithms. These algorithms complement an operator's role in surveilling video content in identifying the presence of firearms in real-time. Such a solution could potentially save lives by allowing authorities to detect and respond quicker to anomalies.

The rapid rise of deep learning approaches [6] demonstrates the potential for a data-driven solution to this problem. State-of-the-art image classification results exceeding human performance come from deep neural networks [7]. A look at the body of work in gun detection reveals that most of them suffer from one of two limitations: hand-crafted feature engineering that does not scale with available data, or the datasets used do not capture the multifarious settings in which a gun detector would be useful (e.g. varying lighting conditions and viewpoint angles). As such, this paper attempts to put forward a solution to the latter. Specifically, in this paper:

- We construct a surveillance based gun dataset consisting of 250 recorded CCTV videos with a total of 5500 images in a multifarious setting.
- Using this annotated gun dataset, we then train a single-stage object detector using a multi-level feature pyramid network. The trained network is then validated using images from the UCF crime video dataset which contains real-world gun violence.

Our experimental results show that our dataset significantly improves gun detection accuracy in surveillance footage compared to other datasets. The rest of the paper is organized as follows: Section II reviews the related work. Section III describes the object method used in detecting a gun from surveillance images. Section IV describes the construction of a gun dataset within a video surveillance perspective. Section V describes the experimental setup, results and discussion. Finally, Section VI concludes this paper.

II. Related Work

There is little body of work directly concerned with gun detection in a surveillance setting. Looking at both manual feature engineering and deep learning approaches, we find that they are: a) confined within artificial lab settings, or b) use

978-988-14768-7-6©2019 APSIPA

deep models on datasets that fail to capture the rich context in surveillance images, or c) use fixed features independent of the dataset that cannot scale with available data or d) overestimate performance by ignoring measures of localization performance. Although standard object detection datasets (e.g. Pascal VOC [8], MS COCO [9]) exhibit volume and variety of examples, they are not suitable for gun detection as they annotate a set of object categories that does not include guns.

Feature engineering. In [10], a training/test split of 12,000 images each was used for gun detection in a lab setting. The authors applied foreground extraction followed by edge detection and a sliding window on this image to extract patches that are reduced using PCA and classified into regions of interest (RoIs) using a neural network. These RoIs are matched against a gun template using MPEG-7 region descriptors. Details on detection evaluation are sparse, and the reported scores are a *sensitivity* of 95% for baseline, and 35% after tuning to reduce false positives. The pipeline is complex, uses fixed features and needs hand-tuning to adapt to different environments.

In [11], image segmentation feeds segmented instances for SURF feature computation. These features are matched against a predefined gun template, using a nearest neighbor algorithm. The same authors, in [12], report a similar work with SURF replaced by FREAK descriptors. Their best reported performance is a *true positive rate* of 86%, but the implementation is slow and assumes guns to be of uniform color.

Deep feature learning. In [13], the authors use Overfeat trained on a subset of the IMFDB dataset[1]: 2535 for training and 218 for testing. They reported an overall *accuracy* of 93%, with 58% on revolvers and 46% on rifles. Non-standard metrics are used so it is not indicative of localization performance. Furthermore, the implementation is slow and requires 1.5 seconds per image. In [14], the authors use Shi-Tomasi corners to find RoIs to feed to a Mobilenet V1 classifier. The training images are taken from various online sources. They report a *true positive rate* of 86% (on 30 images of guns). This is highly likely to be an overestimate of true performance (due to the small test set), and again the implementation is slow due to the use of sliding windows.

Recently, Olmos et al. [15] utilized the Faster R-CNN detector on their collected dataset, which has 3,000 training images, and 608 testing images—collected from various online sources. They reported a *precision* of 84.2%. The evaluation protocol is unspecified, but examining their dataset reveals a lack of rich context that would be representative of surveillance settings—figure 1 shows some typical examples from their dataset.

It is evident from the preceding review that there is a clear gap in the existing body of work in gun detection on visual images. Specifically, we note that the lack of available contextual handgun information from CCTV images for representation learning within a surveillance perspective, and the fact that issue remains unresolved. Therefore, the following sections describes the object detection method used for handgun detection and the dataset development which consist solely of images of guns from a surveillance perspective.

Fig. 1. Guns without rich context

III. Gun Detection Method

In handgun detection in an image, the location of the gun in the image must first be found followed by gun recognition. To do so, we generally use object detectors and apply them in the context of guns. There are many object detectors with different convolutional neural network architectures. Each architecture is unique and has its own advantages. Current state-of-the-art object detectors exploit feature maps of various scales to improve detection accuracy as well as reducing memory used and computational cost. Object detectors that exploit these feature maps of different resolutions are Mobilenet-SSD, Faster R-CNN with FPN, RefineDet and M2Det. Mobilenet-SSD [16] uses the later-layers of a VGG16 backbone and depthwise separable convolutions for multi-scale detection. This allows Mobilenet-SSD to detect accurately and execute in real-time as it is a single-stage detector. Faster R-CNN with FPN [17] uses a combination of region proposals from the Faster R-CNN architecture and lateral connections with a top-down pathway to produce a feature pyramid with more representative features. However, this causes Faster R-CNN with FPN to be very computationally heavy and consumes much more resource due to its two-stage process but achieves higher accuracy than Mobilenet-SSD. RefineDet [18] is also a two-stage architecture which uses two-step cascade regression for accurate object localization to achieve higher accuracy than Faster R-CNN with FPN and retains the efficiency of Mobilenet-SSD.

Recently, a newly designed state-of-the-art architecture (i.e., M2Det [19]) integrates a multi-level feature pyramid network (MLFPN) into Mobilenet-SSD. M2Det is a single-stage object detector based on a multi-level feature pyramid network, which is capable of overcoming limitations in existing single-stage and two-stage feature pyramid networks. Specifically, feature maps in existing pyramid networks are not accurately represented for object detection tasks due to the layers being constructed from a backbone which was intended for object classification tasks. Each feature map in the pyramid is mainly constructed from single-level which only contain low-level information. M2Det solves this limitation by using an architecture that fuses multi-level features extracted from a backbone as base features, and then feeding it into an

[1] http://www.imfdb.org

alternating joint thinned u-shape module (TUM) and feature fusion module(FFM). These modules are used to construct feature maps which contain rich multi-scale and multi-level features. Hence, the purpose of using M2Det is to essentially train a feature pyramid network to learn the multi-scale and multi-level spatial representation of objects at varying resolutions, object rotations and angles. Such features would be advantageous in detecting firearms as an object within the perspective of a video surveillance environment. Apart from this, M2Det demonstrates substantially higher detection accuracy and higher effiency compared to Mobilenet-SSD, Faster R-CNN with FPN and RefineDet. Hence, given the performance advantages of M2Det, this detector is used for the remainder of the paper.

IV. Dataset Development

A. Existing datasets

In this section, we review existing datasets which contain images of guns from a CCTV perspective. The **UCF Crime** dataset [20] consists of surveillance videos which are data mined from YouTube and LiveLeak. These videos are classified into different classes which are: Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting and Vandalism. The videos in this dataset are down-sampled to 240×320 pixels and have a frame rate of 30 fps. However, not all videos from each class contain gun footage, and only approximately 26 videos in this dataset have clear representation of guns. A majority of the videos in this dataset contain weak representations of gun/s in part due to the effect of double compressing a CCTV video (i.e., the video frames were compressed during the recording stage and then compressed again during upload stage into YouTube or LiveLeak). The current amount of videos are simply insufficient for an accurate representation learning of a gun in a multifarious setting. To this end, Olmos et al. from the University of Granada [15] constructed a dataset consisting of 3000 images which contains handguns in various contexts and scenarios. These images were mined from the world wide web. In this paper we refer to this dataset as the **Granada** dataset. The context of guns in this dataset is rich. However, only 48 out of 3000 images of guns are from a CCTV perspective which also makes this dataset insufficient for representation learning of a gun within a surveillance context.

B. Our dataset

Due to the absence of a well defined dataset which consist solely of images of guns from a surveillance perspective, we have constructed a dataset to overcome this issue. The idea here is to have a dataset which contains a clear representation of handguns within a video surveillance context. Based on this idea, we implemented role playing activities with a person/s brandishing a gun and these activities were recorded using a CCTV camera. At the time of this writing, the constructed dataset currently has 5,500 images and is expected to grow further. The recorded images of guns in this dataset are of different scenarios to ensure for a diverse representation learning.

Data collection. Prior to constructing CCTV based videos containing the presence of guns, we first analyzed certain factors which contribute to the diversity of our dataset:

- Lighting conditions — day/evening/night
- CCTV configurations — height/angle/footage depth
- Environment — indoor/outdoor/background clutter

TABLE I
Number of videos for each scenario in our dataset

Lighting/Environment	Indoor	Outdoor
Day	1	39
Evening	16	105
Night	3	55

TABLE II
Number of frames for each scenario labeled for training

Lighting/Environment	Indoor	Outdoor
Day	3984	228
Evening	456	456
Night	104	203

We simplify the data collection process by ensuring that the collected dataset contains examples that cover variations in each of the factors above. To cover varying illumination conditions, we filmed scenes under both daylight and night-time settings. Typically, CCTV mountings do not follow a particular standard in terms of camera elevation and viewing angles. A CCTV is mounted to provide the best possible viewing angles for a particular surveillance scenario. As such, during the collection process, we subjectively analyzed real-world surveillance camera footage and compared it with our CCTV setup to determine a suitable viewing angle and depth. To account for scene variety, we covered both indoors and outdoors environment, and included scenes with background clutter such as trees, shrubs and helifans. The gun used in the video recordings is a replica of a hand-held pistol with a uniform black color for its body. The recorded videos were then annotated for representation learning, which is be expanded in the following section.

V. Experiment

A. Implementation details

All images from the Granada dataset, UCF Crime dataset and our dataset were re-scaled from 1080×1920 pixels to 512×512 pixels as this resolution is required for M2Det training. We start training using M2Det's warm-up strategy for 5 epochs and initialize the learning rate as 4×10^{-3}, then decrease it to 2×10^{-3} at 100 epochs, 4×10^{-4} at 130 epochs, 4×10^{-5} at 150 epochs and 4×10^{-5} at 200 epochs and stop it at 300 epochs. The experiment is conducted on a machine with PyTorch 1.0.1, CUDA 9.0, cuDNN 7.3.1, two

TABLE III
COMPARISON OF GUN DATASETS

Dataset	# of Frames	# of Frames in CCTV context	Scenario
Granada	3000	48	Mostly frontal, side view of guns
UCF Crime	7247	419	Blurry CCTV footage
Ours	**5500**	**5500**	**Clear CCTV footage, all viewing angles of guns covered**

NVIDIA Quadro P5000 graphics processing units (GPUs) with a combined video memory of 32 GBytes. The batch size is set to 32 images (with 16 images for each GPU). The total training time using a VGG-16 backbone took approximately 2 days and 6 hours.

B. Experimental results and analysis

In this experiment, we extracted 45 frames from the UCF Crime dataset to act as a baseline test set for evaluating the accuracy of each dataset. In these 45 frames, all guns are distinguishable by the human eye from a CCTV point of view. We evaluate the performance of each dataset by testing the trained neural network models using MS COCO metric.

We first train a neural network model, *Model 1*, using the Granada dataset which consists of 3000 images of guns, then train a second model, *Model 2*, using the combination of our dataset and the Granada dataset. The assumption is the addition of gun images from a CCTV perspective in the dataset will increase the distribution of guns with small resolutions and enhance gun detection in surveillance footage.

TABLE IV
COMPARISON OF DETECTION ACCURACY BETWEEN DATASETS IN TERMS OF MAP PERCENTAGE

Model	Dataset	Avg.Precision, IoU			Avg.Precision, Area		
		0.5:0.95	0.5	0.75	S	M	L
1	Granada	0.114	0.281	0.053	0	0.110	0.800
2	**Ours**	**0.223**	**0.442**	**0.202**	**0.180**	**0.224**	0.717

From Table IV, it shows that *Model 1* cannot detect low resolution guns from the test set, but is capable of detecting medium resolution guns with 11% mAP and high resolution guns with 80% mAP. This is because the Granada dataset contains a large portion of guns with high resolutions and a small portion with a mixture of low and medium resolution of guns. We recognize this weakness of the Granada dataset and try to solve it by combining the Granada dataset with our dataset. The result shown in Table IV is evidence that our assumption is true. *Model 2* is capable of detecting low and medium resolution guns more accurately with an increase in mAP of 18% and 11.4% respectively compared to *Model 1*. Even though the dataset is still in an unpolished state with much room for improvement, we argue that our approach in this work is justified by the above observations. Namely, an increase in dataset volume and variety leads to significant performance improvements relative to the Granada baseline. We justify this by the observation that a ∼1.8 fold increase in dataset volume (from 3,000 images in Granada to ∼8,500 in our combined dataset) yields a minimum of 11% increase in mAP—with no frills or tricks applied during training. With careful regularization and hyperparameter tuning, we suspect that much more performance may be gained. Thus, our dataset can potentially improve the accuracy even further with higher diversity of guns from different viewing angles and background.

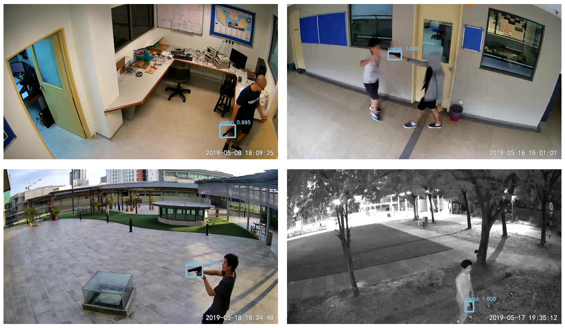

Fig. 2. Examples of guns from various scenarios in our dataset

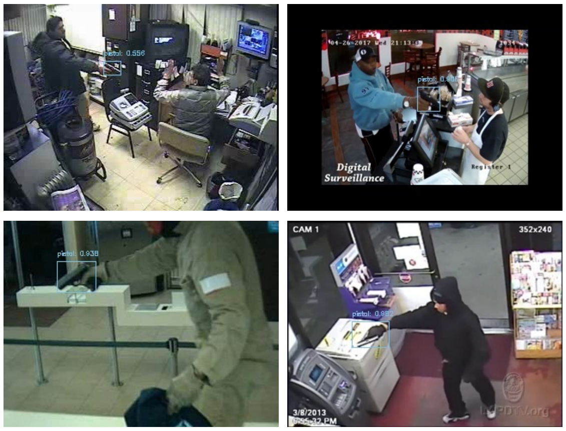

Fig. 3. Validation results on UCF Crime dataset using our dataset

VI. CONCLUSION AND FUTURE WORK

In this paper, we constructed a dataset which contain guns from a CCTV perspective as a solution to the absence of a surveillance based gun dataset for representation learning using deep neural networks. In this dataset, we take into account factors such as lighting conditions, CCTV configurations and environments which contribute to the gun detection accuracy from a CCTV perspective to increase the diversity of the dataset. To validate the accuracy of our dataset, we trained it using M2Det as a single stage object detector using MLFPN. Experimental results indicate that our dataset improves the accuracy of gun detection in real-world scenarios against that of existing datasets. For future work, we intend to add more classes and different types of guns to the dataset to enrich it further. We also intend to explore the use of other neural network architectures with a focus towards small scale object detection in attempt to further improve the accuracy in detecting firearms from a surveillance video.

VII. Acknowledgement

This work was funded by the Malaysian Ministry of Education's Fundamental Research Grant Scheme (Grant Number: FRGS/1/2018/ICT02/MUSM/02/3) under the purview of Monash University Malaysia.

References

[1] R. Smith, "Preventing Gun Violence," *Ned for CT*, p. 1, 2018.

[2] E. Grinshteyn and D. Hemenway, "Violent Death Rates: The US Compared with Other High-income OECD Countries, 2010," *American Journal of Medicine*, vol. 129, no. 3, pp. 266–273, 2016.

[3] A. Doyle, R. Lippert, and D. Lyon, *Eyes everywhere: The global growth of camera surveillance*. 2013.

[4] H. M. Dee and S. A. Velastin, "How close are we to solving the problem of automated visual surveillance?," *Machine Vision and Applications*, vol. 19, pp. 329–343, 10 2008.

[5] G. J. Smith, "Behind the screens: Examining constructions of deviance and informal practices among CCTV control room operators in the UK," *Surveillance and Society*, vol. 2, no. 2-3, pp. 376–395, 2004.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," tech. rep.

[8] M. Everingham, L. Van Gool, C. K. I Williams, J. Winn, A. Zisserman, M. Everingham, L. K. Van Gool Leuven, B. CKI Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *Int J Comput Vis*, vol. 88, pp. 303–338, 2010.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," tech. rep., 2014.

[10] M. Grega, A. Matiolański, P. Guzik, M. Leszczuk, M. Grega, A. Matiolański, P. Guzik, and M. Leszczuk, "Automated Detection of Firearms and Knives in a CCTV Image," *Sensors*, vol. 16, p. 47, 1 2016.

[11] R. K. Tiwari and G. K. Verma, "A computer vision based framework for visual gun detection using SURF," in *2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)*, pp. 1–5, IEEE, 1 2015.

[12] R. K. Tiwari and G. K. Verma, "A Computer Vision based Framework for Visual Gun Detection Using Harris Interest Point Detector," *Procedia Computer Science*, vol. 54, pp. 703–712, 2015.

[13] J. Lai and S. Maples, "Developing a Real-Time Gun Detection Classifier," tech. rep.

[14] M. Singleton, B. Taylor, J. Taylor, and Q. Liu, "Gun Identification Using Tensorflow," 2018.

[15] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, 2018.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector,"

[17] T.-Y. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[18] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network," in *AAAI*, 2019.

[20] W. Sultani, C. Chen, and M. Shah, "Real-world Anomaly Detection in Surveillance Videos," tech. rep., 2018.