Tracking Suspicious Entities Using UAVs in Critical Urban Areas: A R-CNN Approach

Mathias A. G. de Menezes
Computer and Systems
Graduate Program
Military Institute of Engineering
Rio de Janeiro, Brazil
E-mail: mathiasdemenezes@ime.eb.br

Paulo F. F. Rosa (PhD)
Computer and Systems
Graduate Program
Military Institute of Engineering
Rio de Janeiro, Brazil
E-mail: rpaulo@ime.eb.br

Erick Menezes Moreira (Dr Eng)
Computer and Systems
Graduate Program
Military Institute of Engineering
Rio de Janeiro, Brazil
E-mail: emenezes@ime.eb.br

***Abstract*—This paper proposes a tracking application that integrates object detection with a Region-based Convolutional Neural Network as the object detector and the Discriminative Correlation Filter with Channel and Spatial Reliability as the tracking algorithm for the tracking method. Our approach has the objective and motivation of assisting the operational actions of the security forces in Rio de Janeiro, especially the Military Police, in deflagrated regions. The results of the generated model showed an average accuracy of 86% for the object detector and an average of 74% for the object tracker when applied to the video sequences of our dataset.**

***Keywords*—object detection; object tracking; r-cnn; surveillance.**

I. Introduction

Unmanned Aerial Vehicles (UAVs) are commonly used in police and military operations [1], [2], especially for monitoring and tracking suspicious mobile entities in risky urban areas such as slums, war zones and terrorist-controlled regions. This approach enables learning and understanding the opponents tactics, providing actionable intelligence to anticipate and act on insurgent activities relating vehicles, places, and routes of locomotion [3], [4]. Monitoring and tracking are particularly challenging in low visibility enviroments where objects are ocluded [5].

In the city of Rio de Janeiro, police forces find it difficult and resisting to operate incursions into slums dominated by criminal groups. Police actions are usually accompanied by helicopters that support ground troops. However, this action reveals the intents of the police in those regions, putting an end to the surprise factor, as criminals manage to contain the advance of the police forces, as shown in Figure 1. As a consequence, the Military Police of Rio de Janeiro (PMERJ) has a high casualty rate, with 198 agents being killed in action against criminal gangs only in 2020. To reduce these casualties, the operation of UAVs in these scenarios is convenient. UAVs have reduced size, and models can no longer be seen at a height of 100 meters, which prevents them from being targeted by firearms and destroyed [6].

Figure 1. Military Police helicopter, shot down by criminals during an operation in 2009. In the fall, three police officers died and five were injured.

In 2012, the Military Institute of Engineering (IME), with support from other institutes to overcome these deficiencies, built six UAVs to be used in missions by the Rio de Janeiro security forces. The UAVs were used for surveillance, security and remote sensing and in security in stadiums and at major events, such as the 2014 Soccer World Cup and the 2016 Olympics.

The PMERJ also use other UAVs to monitor slums regions and assist in planning operations against drug trafficking. In one of the most violent slums in the city, images captured by the UAV shown in Figure 2 led police to discover two camps used by drug dealers hidden in a place of difficult access [6]. Without aerial imagery, it would be difficult to progress within these locations. The equipment shown in Figure 2 is operated by two people, one responsible for moving the UAV and the other for operating two attached cameras. However, none of the equipments operated by the military police has an autonomous application for detection and tracking of targets on board. With the motivation to contribute to solving this problem, we propose a method for detection and tracking suspicious entities, we use a Region Based Convolutional Neural Network (R-CNN) and the CSRT tracker, a python implementation of the Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR).

The rest of the paper is organized as follows. The Related Work is described in Section II, the problem formulation is described in Section III and our proposed tracking method is described in Section IV. The valuation of the proposed approach is carried out in Section V, and we conclude the paper in Section VI.

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-918-8

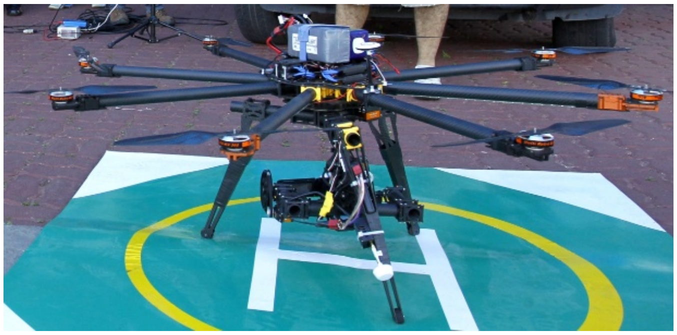

Figure 2. The multirotor has a camera with an infrared viewfinder attached for night monitoring, and a transmission link to enable real-time monitoring of images.

II. Related Work

Persistent tracking of targets in urban environments using UAVs is a difficult task due to the limited field of view, obstructed visibility of obstacles and uncertain target movement. The vehicle must be properly positioned, allowing visibility to the target to be maximized. In [7], an approach to target pursuit is presented, which constitutes a deep reinforcement learning technique based on Deep Q-Networks, with a curriculum training framework for the UAV to persistently track the target in the presence of obstacles and movement uncertainty. The results show that the UAV persistently tracks the target in diverse environments, avoiding obstacles in trained environments as well as in visible environments.

Another big challenge is tracking the object under occlusion conditions. The TensorFlow object detection API was used in [8] to detect moving objects. The location of the detected object is passed to a new CNN-based that was used for robust object detection. The approach proposed by [8] is able to detect the object in different illuminations and occlusions, reaching great accuracy in self-generated image sequences.

Occlusions and interactions between different objects are expected and common due to the rugged nature of these urban areas. In [9], a tracking framework employing classification label information from a deep learning detection approach was used to associate the different objects in addition to the objects positions and appearances. The results showed that object labels improve tracking performance, but that the output of object detectors is not always reliable.

In a persistent surveillance task, UAVs sometimes cannot independently complete the task and need to be supported by ground equipment. Thus, [10] presents a system of UAVs and UGVs to perform surveillance tasks, and the goal is to generate circular paths for UAVs and UGVs, respectively, to increase the operating time to complete the coverage of the environment. Keeping a circular flight in an area helps to circumvent the object's occlusion. In their approach, [10] integrates a distribution estimation algorithm (DEA) and a genetic algorithm (GA) to solve the problem. The advantages of DEA and GA in global and local search fully consider the demands in the different phases of the iterative process. This way, one can scan and determine the ideal sequence of passage of the open points. Then, an online site adjustment strategy is also applied to deal with changing land area coverage requirements. Simulation results demonstrate that UAV and UGV systems can increase surveillance efficiency.

The detection and tracking problem for reconnaissance and surveillance of UAVs requires that they fully cover an area of interest along their trajectories. Thus, [11] presents a two-phase strategy to solve this UAV recognition problem with a specified altitude. First, an easily implementable estimation algorithm is developed at a given altitude, and the minimum and how to targeted number of cameras is determined to provide complete coverage of the target area. The second phase deals with the distribution of achievements in one or more UAVs and creates the paths for them to recognize a corresponding area of interest. The results reported support the feasibility of their proposed solution.

Another work [12], proposed an approach to detect moving objects in wide area motion imagery, in which the objects are small and well separated. The approach was based on background subtraction as an efficient and unsupervised method capable of producing object shapes. To reliably detect small, low-contrast objects, they set up background subtraction to extract foreground regions that might be objects of interest. Although this dramatically increases the number of false alarms, the CNN, considering spatial and temporal information, is then trained to reject false alarms. In high-traffic areas, background subtraction produces mixed detections. To reduce the complexity required of tracking multiple targets, they trained another CNN to predict the positions of multiple moving objects in an area.

III. Problem Formulation

We emphasize that the motorcycle is the vehicle used by criminal organizations in the slums of Rio de Janeiro. Generally, they work in groups or pairs and mingle with citizens who travel through the alleys of the environment. Therefore, it is difficult to identify a suspicious vehicle, either by a large aircraft such as a helicopter, or by a small UAV without an embedded computer application. Thus, it is necessary to create the model of the target to be tracked, preserving its aspects in order to classify it.

An ideal approach to solve this problem is the use of Region-based Convolutional Neural Networks (R-CNN). These networks use as input data regions cut out of the image to detect whether a number of objects of a certain category are present, as well as detecting where each object is located in the image. This type of network is very robust and can discern several clustered objects simultaneously, regardless of the occlusion of parts of the target object.

The UAV that PMERJ uses is a multirotor with an integrated infrared camera and a radio link for communication. Two human operators manage the drone during the mission: where one operates the camera and the other moves the aircraft. Basically, the application we propose will help the camera

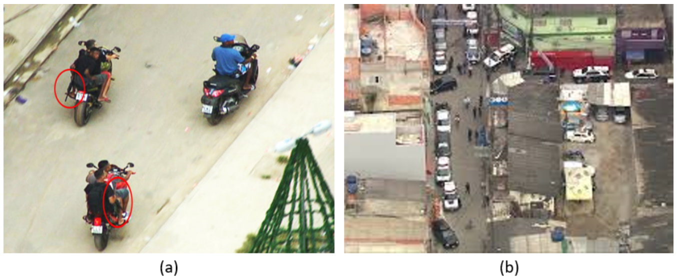

Figure 3. Men armed with rifles (a) move on a street in a slum (b).

operator to detect and track suspicious entities and monitor them within the camera's field-of-view, as shown in Figure 3.

IV. Proposed Tracking Method

Our proposed object tracking method contains two parts: the trained object classifier and the generation of the object detection model from the training result. The selective search algorithm that takes the regions and delivers them to the object detection model to predict the object location and classify it as a object class. The detection output predictions feed the tracking algorithm to track the predicted box of the object class. Any other circumstance of the object prediction change tracker will trigger the object classifier and restart the process from the beginning.

The tracker, known as Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR), uses spatial reliability to define the filters support for a portion of the selected area of the frame for tracking. This expands and locates the selected zone and tracks non-rectangular areas or objects. This tracker practices two standard features, HoGs and Colornames. Furthermore, it works at frames below 25 fps [13].

A. R-CNN for Object Detection

The R-CNN method in [14], is a machine learning model that performs segmentation based on the results of object detection. The R-CNN initially uses the selective search algorithm to extract a large amount of proposed regions and then calculates the characteristics for each one of them through a Convolutional Neural Network (CNN). Finally, it classifies each region using a specific linear classifier, typically a Support Vector Machine (SVM). The R-CNN is capable of performing more complex tasks, such as object detection and coarse image segmentation.

B. Extracted Regions

The R-CNN initially generates around 2000 proposed regions using the Selective Search algorithm [15], which is based on simple traditional computer vision techniques. The process is as follows: first, each proposed Region of Interest (RoI) is deformed into a square image of standard size; second, the image is fed to a CNN that generates a array with 4096 dimensional features as output; and finally, a SVM classifies the feature array producing two outputs: a classification, and an indication of deviation (offset) that can be used to adjust the bounding box.

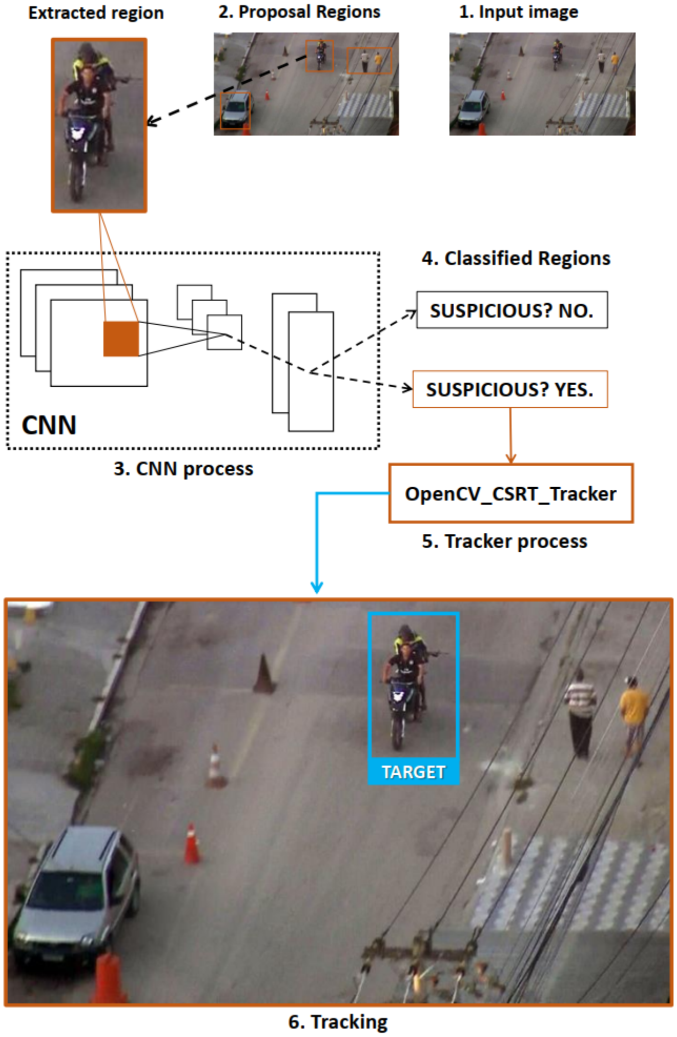

Figure 4. Proposed Tracking Method

C. Processing of Convolutional Characteristics

The extracted proposed regions will feed the CNN. We used the VGG-16 CNN [16]. Basically, the CNN will receive a proposed region passing through a series of convolutional, non-linear, clustering and fully connected layers to obtain two outputs. An output is a single class that best describes the proposed region. The CNN is structured in four layers, or stages: convolution layer, grouping layer, normalization layer, and fully connected layer.

D. Object Tracking

The basis of our proposed tracking method was taken from the DCF-CSR algorithm. Furthermore, this algorithm was implemented and integrated into the OpenCV library as a Deep Neural Network (DNN) module. We propose a tracking application that integrates object detection with R-CNN as the

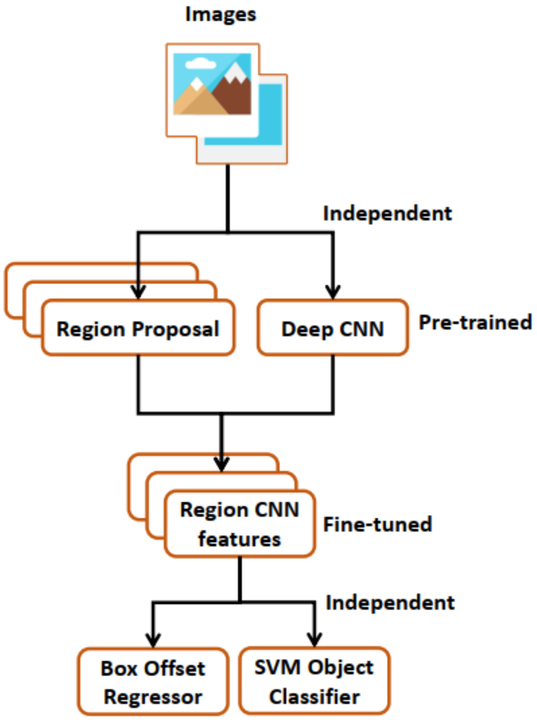

Figure 5. R-CNN Model Architecture

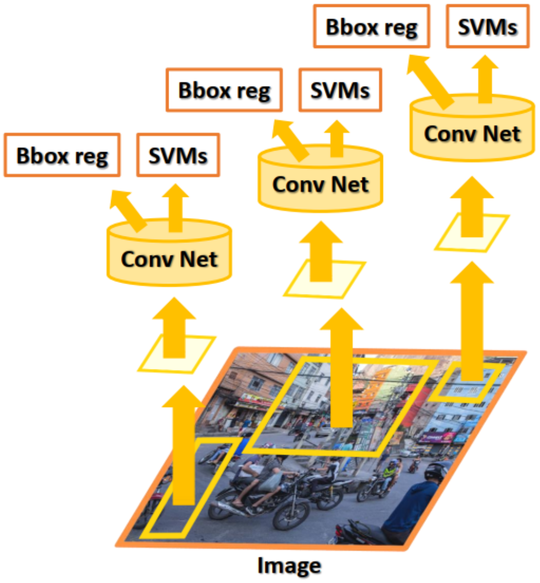

Figure 6. R-CNN architecture: Each proposed RoI is passed through the CNN to extract features then an SVM classifier

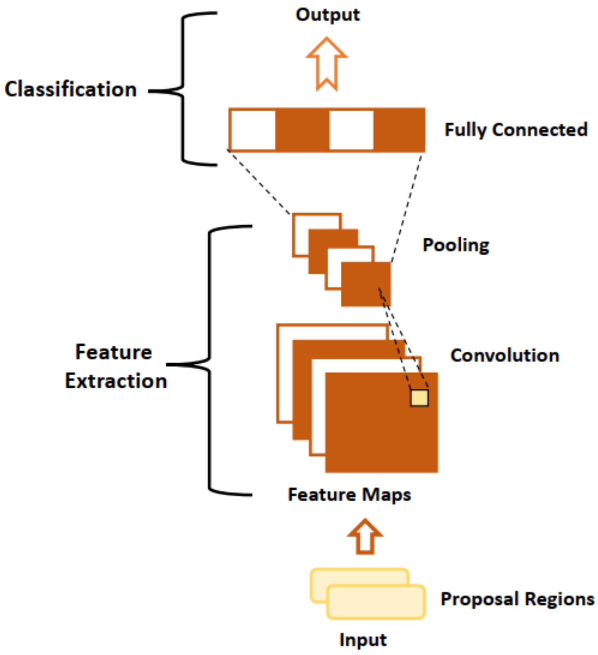

Figure 7. CNN architecture

object detector and OpenCV_CSRT_Tracker as the tracking algorithm for the tracking method.

V. Experimental Results

Our image dataset was collected from Internet resources such as blogs, reports, news portals, etc. The difference between our image collection and other image collections is that our images are taken by drone cameras, helicopter cameras and security cameras installed in local urban environment. We have collected 340 images in total, which were divided into 220 images for the training process and 120 images for the testing process. The extracted regions from images there are two categories: Suspicious Entity (SE) and Not-suspicious Entity (NE).

After finishing 5000 epoch times training with our dataset, we got our object classifier model and tested our object detection (classifier) model by applying images from different open source resources, reaching the accuracy rate of 86%, as shown in Figure 8.

We can see the remarkable results of the object detection model, where it is detecting all objects in the frame. While tracking, we may face thousands of positions, locations, shapes that means it requires more images on dataset with different positions and environments for better results, as shown in Figure 9.

We have tested our tracking method with video sequences taken by helicopter and obtained promising results. The experiments show that the R-CNN_CSRT tracker algorithm can re-detect the object once it is gone from the current frame. Figure 10 presents qualitative results for the video sequences taken by a helicopter, with frameworks below 25 fps. Even if

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-918-8

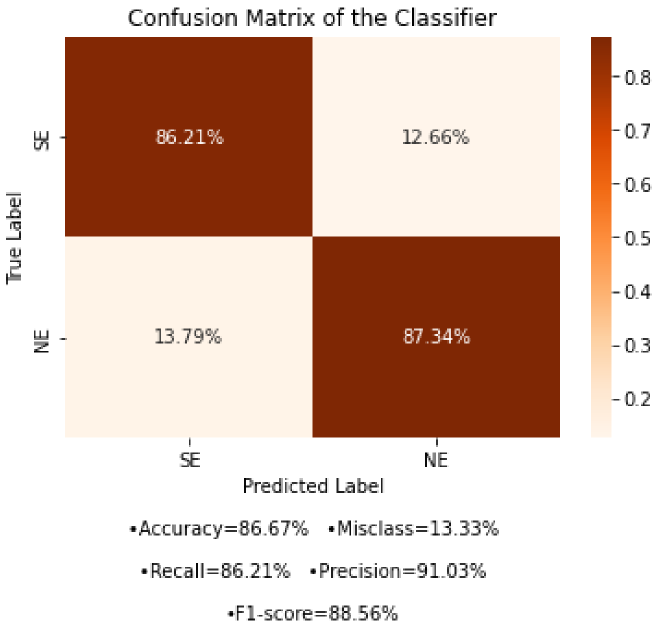

Figure 8. Statistics the trained classifier.

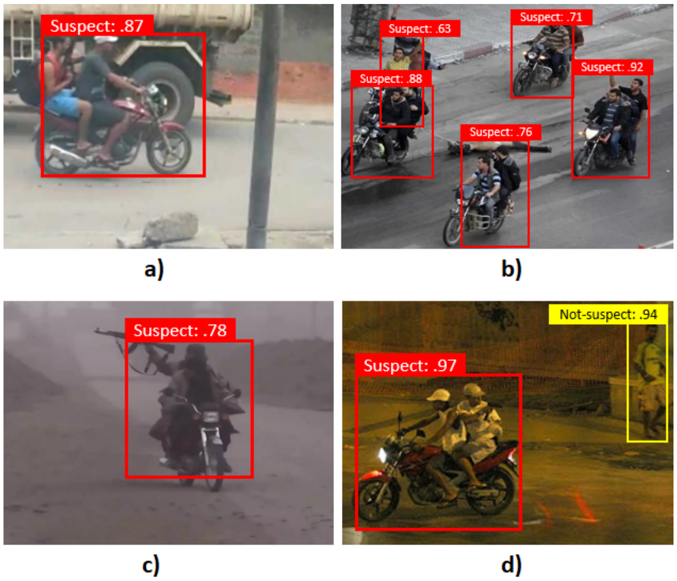

Figure 9. Object detection results of the trained R-CNN object classifier: images from our dataset.

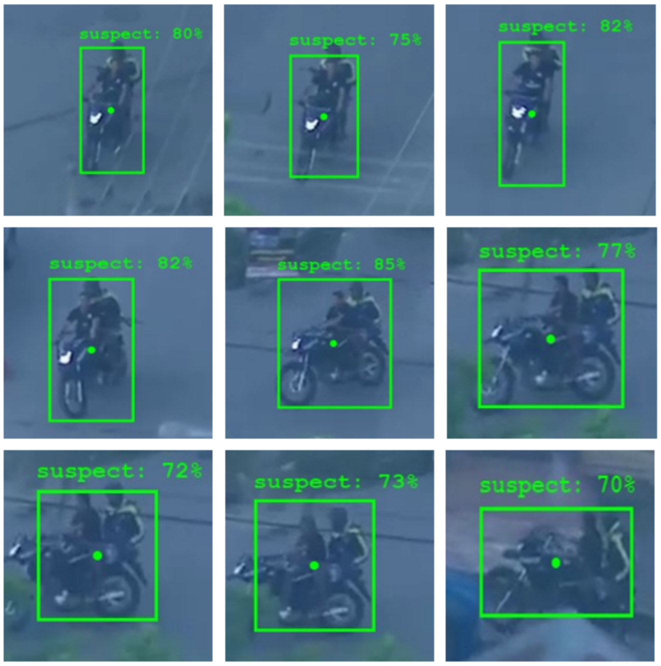

Figure 10. Qualitative results of the proposed tracking method in extracted frames.

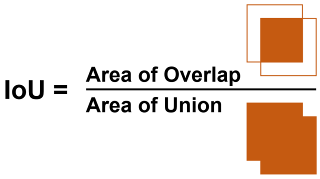

Figure 11. The calculation of the IoU is done by dividing the overlap area between the bounding boxes by the union area.

the shape or appearance of the tracking object changes, the tracker can track the object properly.

In order to measure how good our object tracker is at predicting bounding boxes, we use the Intersection over Union (IoU) metric. The IoU method calculates the ratio of the overlapping area to the joint area between the predicted bounding box and the ground truth bounding box. The IoU is an evaluation metric used to measure the accuracy of an object detector/tracker against a specific dataset. This evaluation metric is often used in object detection and tracking challenges, as in approaches with R-CNN, Faster R-CNN, YOLO and Deep SORT as in [17]. However, the actual algorithm used to generate the predictions does not matter. The intersection over the union is simply an evaluation metric. Any algorithm that provides predicted bounding boxes as output can be evaluated using IoU.

More formally, in order to apply the IoU to evaluate an (arbitrary) object detector/tracker, we need: (i) the ground truth bounding boxes (that is, the hand-labeled bounding boxes of the test suite that specify where the object is in the Image); and (ii) the predicted bounding boxes of the generated model as show Figure 11. The average accuracy of classification of our proposed method reached 74.83% when applied to video sequences.

VI. Conclusion

We presented a R-CNN_CSRT tracker to track suspicious entities in critical urban environments. With this method, we integrate the object classifier model based on deep learning with the OpenCV implementation version of the CSRT tracker of the DCF-CSR algorithm supported by the DNN OpenCV module. Our results showed that our trained object classifier model was accurate after 5000 epoch times training times, with only 220 images for training and 120 images for testing. However, when applied to video sequences with images captured by helicopters, the tracker performed below expectations. These images need to be in good resolution, with a greater amount of angulation, detailing the position and shape of the entities present.

In conclusion, our work had some limitations. We need a greater number of images representing the class of suspicious entities, as well as an improvement in the CNN structure, increasing its convolutional layers and the number of training cycles. In future works we will use variations of R-CNN, such as a Faster R-CNN. These changes can increase the accuracy of the proposed method, possibly making the tracker more robust and effective.

Acknowledgment

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance code 001.

References

[1] T. Samad, J. S. Bay and D. Godbole, "Network-Centric Systems for Military Operations in Urban Terrain: The Role of UAVs," in Proceedings of the IEEE, vol. 95, no. 1, pp. 92-107, Jan. 2007, doi: 10.1109/JPROC.2006.887327.

[2] E. Semsch, M. Jakob, D. Pavlicek and M. Pechoucek, "Autonomous UAV Surveillance in Complex Urban Environments," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, pp. 82-85, doi: 10.1109/WI-IAT.2009.132.

[3] H. Geng, J. Guan, H. Pan and H. Fu, "Multiple Vehicle Detection with Different Scales in Urban Surveillance Video," 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), 2018, pp. 1-4, doi: 10.1109/BigMM.2018.8499095.

[4] C. Hu , G. Qu , H. S. Shin, A. Tsourdos, "Distributed synchronous cooperative tracking algorithm for ground moving target in urban by UAVs," International Journal of Systems Science, 2020. DOI: 10.1080/00207721.2020.1844340.

[5] M. Daikoku, S. Karungaru and K. Terada, "Automatic detection of suspicious objects using surveillance cameras," The SICE Annual Conference 2013, 2013, pp. 1162-1167.

[6] L. S. Alves P. Integrated Aerial Imaging Systems in Special Operations in an Urban Environment. Thesis (Masters in Defense Engineering) – Military Institute of Engineering (IME). Rio de Janeiro, p. 84. 2020.

[7] S. Bhagat and P. B. Sujit, "UAV Target Tracking in Urban Environments Using Deep Reinforcement Learning," 2020 International Conference on Unmanned Aircraft Systems (ICUAS), 2020, pp. 694-701, doi: 10.1109/ICUAS48674.2020.9213856.

[8] S. Mane and S. Mangale, "Moving Object Detection and Tracking Using Convolutional Neural Networks," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 1809-1813, doi: 10.1109/ICCONS.2018.8662921.

[9] H. L. Ooi, G. A. Bilodeau, N. Saunier, D. A. Beaupré, "Multiple Object Tracking in Urban Traffic Scenes with a Multiclass Object Detector," In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2018. Lecture Notes in Computer Science, vol 11241. Springer, Cham. https://doi.org/10.1007/978-3-030-03801-4_63.

[10] Y. Wu, S. Wu and X. Hu, "Cooperative Path Planning of UAVs UGVs for a Persistent Surveillance Task in Urban Environments," in IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4906-4919, 15 March15, 2021, doi: 10.1109/JIOT.2020.3030240.

[11] J. Zhang and Y. Zhang, "A Method for UAV Reconnaissance and Surveillance in Complex Environments," 2020 6th International Conference on Control, Automation and Robotics (ICCAR), 2020, pp. 482-485, doi: 10.1109/ICCAR49639.2020.9107972.

[12] Y. Zhou and S. Maskell, "Detecting and Tracking Small Moving Objects in Wide Area Motion Imagery (WAMI) Using Convolutional Neural Networks (CNNs)," 2019 22th International Conference on Information Fusion (FUSION), 2019, pp. 1-8.

[13] A. Lukežič, T. Vojíř, L. Čehovin Zajc et al., "Discriminative Correlation Filter Tracker with Channel and Spatial Reliability," Internacional Journal of Computuer Vision 126, pp. 671–688 (2018). https://doi.org/10.1007/s11263-017-1061-3.

[14] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, (2014), pp. 580-587, doi: 10.1109/CVPR.2014.81.

[15] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers et al., "Selective Search for Object Recognition," Internacional Journal of Computer Vision 104, pp. 154–171 (2013). https://doi.org/10.1007/s11263-013-0620-5.

[16] k. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015 The 3rd International Conference on Learning Representations, pp. 1-8 (ICLR2015). https://arxiv.org/abs/1409.1556.

[17] A. Pramanik, S. K. Pal, J. Maiti and P. Mitra, "Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking," 2021 in IEEE Transactions on Emerging Topics in Computational Intelligence, doi: 10.1109/TETCI.2020.3041019.