Automatic Rifle & Sniper Detection using Yolo-NAS and Yolov7 Pose

Surbhit Shukla, C. S. Raghuvanshi and Hari Om Sharan
Faculty of Engineering & Technology, Rama University, Uttar Pradesh, Kanpur
drcsraghuvanshi@gmail.com

Abstract:
Automatic Rifle and Sniper detection can help reduce or eliminate threats in both public and private areas, such as a rooftop, a public gathering, a conflict zone, and a VVIP protection area. In the literature, deep learning-based detectors like Yolo-NAS and Yolov7 Pose approaches have been proposed to sound an alarm if a rifle or sniper is spotted in a live video or picture. However, such detectors, like CCTV & UAV and UAV, only take into account how the weapon appears in live footage. In order to increase overall performance, we suggest combining the detector with the Multiple or Individual posture information in this study. The findings indicate a development over the first Rifle and Sniper detector. Techniques named Smart Rifle and Sniper detection systems (SRSDS) for rider posture estimation and correction using Yolo-NAS and Yolov7 Pose. However, those detectors are solely based on the weapon appearance on the image. In this research work, we apply combine the detector with the individual's pose information in order to improve overall performance. To this end, a model that integrates grayscale images from the output of the Rifle and Sniper detector and heat map-like images that represent pose is proposed. The results show an improvement over the original Rifle and Sniper detector. The proposed network provides a maximum improvement of a 31.5% in AP of the proposed combinational model over the baseline Rifle and Sniper detector.

Keywords: - Rifle and Sniper, Pose Estimation, Deep Learning, Sensors, Yolo-NAS and Yolov7 Pose.
DOI Number:10.48047/NQ.2022.20.21.NQ99153 **Neuroquantology 2022; 20(21):1462-1472**

1. Introduction:

In the past few decades, advancements have been made in the areas of CCTV, UAV, and hidden camera setups for the purpose of video surveillance. These days, public, private, and highly protected areas like train stations, airports, museums, banks, or government, institutional buildings conflict zone, and a VVIP protection zone all have their own video surveillance systems and unmanned aerial vehicle (UAV) systems that are deployed and operational. Examples of these types of areas include train stations, airports, and museums. These systems are extremely useful for live investigations as well as assisting security personnel in crowd management by monitoring multiple locations at the same time. The main disadvantage of these solutions is the requirement for continuous monitoring by a human operator. The growing number of areas controlled by video surveillance cameras, as well as human factors such as fatigue or loss of attention over time, render these systems inefficient [1, 2]. Related research in this area indicates that early detection of security threats or risks is critical in order to mitigate the

damage as much as possible [3]. Examples of these types of threats, which have become all too common in recent years, include situations involving firearms, such as Rifle and Sniper attacks, mass shootings, Sniper fire incidents on school grounds [4], or terrorist attacks [5]. The development of intelligent systems capable of detecting threats or risk situations involving firearms as soon as possible can provide significant security benefits. Recently, as a result of the momentum generated by the introduction of deep learning methodologies, remarkable results in visual tasks such as image classification or object detection and segmentation have been obtained. When it comes to the detection of firearms, although the results obtained with these novel methods are promising, there are still significant limitations when they are applied in new scenarios that are different from those that are used for training [6]. In particular, there is an unacceptable number of false positives. [Citation needed] if there is no one there, there is no one there. This paper proposes using human pose as complementary information to improve the performance of current deep learning-based rifle and sniper detectors. The human pose, defined as the relative position of the human body's various joints and limbs, is quite common in shootings. On the other hand, the images obtained by CCTV and UAV cameras are typically of poor quality because of low resolution, noise, or inadequate lighting conditions. Other factors, such as the object's distance from the camera, its small visual size1, or occlusion, either total or partial, can also prevent it from being detected [7]. In this article, we make the hypothesis that contextual information about the body can help improve the robustness of detection. The following are the contributions that this paper makes:

(1) The development of an innovative method for detecting hand- hold snipers and rifle firearms.

(2) The performance evaluation of the proposed method in comparison with other well-known appearance-based detection methods such as YOLOv3 and other recent alternatives that consider human pose information.

(3) An assessment of the method's robustness in environments with poor lighting, large camera distances, and different camera orientation figs. The remainder of the article is structured as follows.

Section 2 describes previous work related to the task of detecting Rifles and Snipers. Section 3 describes the datasets used in this study. The proposed method is explained in Section 4. Section 5 summarizes the experiments performed and the results obtained. Section 6 concludes with conclusions and future work.

2. Related Work

X-ray scanning machines are widely used in public places such as airports, train stations, and museums to detect hidden weapons in luggage. An X-ray image is manually analysed by a security operator. Several approaches based on classical vision methods were proposed in this context to automate the detection process. Nercessian et al. [8] presented a detection system based on image segmentation and edge-based feature vectors in their work. Xiao et al. [9] proposed a method based on Hear-like features and Gadabouts classifiers to detect Rifles and Snipers in such images. In addition, 3D interest point descriptors for object classification in 3D baggage security computed tomography imagery have been investigated [10, 11]. While X-ray imaging-based systems are useful for locating weapons in travel bags or luggage, their scope is very limited. Furthermore, these scanning machines are quite costly. Detecting dangerous objects using RGB images captured by CCTV and UAV video surveillance cameras can be a more versatile and cost-effective option. Several works on the detection of weapons in RGB images using traditional machine learning methods have been proposed in this regard. Tiwari and Verma [12] proposed a method for detecting weapons in RGB images that relied on color segmentation and the k-means algorithm to filter out unrelated objects. The Rifle and Sniper are then located in the segmented images using

the Harris interest point detector and Fast Retina Keypoint (FREAK). Later, Halima and Hosam [13] proposed another method to detect the presence of a Rifle and Sniper in an image. SIFT features are extracted from a collection of images and clustered using the k-means algorithm in this case. Then, a histogram based on word vocabulary is implemented, and finally, a Support Vector Machine is used to determine whether the new image contains a weapon. More recent deep learning-based methods have also been applied to this task, employing various strategies. Sliding windows are an important family of them. Within the image, a large number of regions or windows of varying sizes and aspect ratios are generated (on the order of 104) and each one is classified individually by a neural network. Several studies [14, 15] used this technique to detect Rifles and Snipers in images captured by CCTV and UAV video surveillance cameras. The main disadvantage of this type of system is the lengthy processing time required to classify these windows, which makes them difficult to use in real time. Other approaches are based on region proposals, which select only actual candidates rather than all possible windows in an image. The Region-based Yolo-NAS and Yolov7 Pose family of methods [16, 17] were the first to use Yolo-NAS and Yolov7 Pose in this context. To detect hand-held arms, Verma and Dhillon [18] proposed a method based on the Yolo-NAS and Yolov7 Pose framework with a VGG-16 backbone as feature extractor trained on the IMFDB dataset [19]. Olmos et al. [20] tested and compared the sliding window and Faster RYolo-NAS and Yolov7 Pose methods for Rifle and Sniper detection. On a custom dataset of 3000 YouTube Sniper images, Faster-RYolo-NAS and Yolov7 Pose pre-trained with VGG-16 architecture produced the best results. Finally, another popular method for detecting objects is the YOLO family of methods [21, 22, 23]. Instead of multiple region proposals, a single deep neural network is applied to the entire image in these architectures. The image is divided into fixed regions, and probabilities and bounding boxes for each are predicted. Several studies have recently used YOLOv3 to detect firearms, with promising results [24, 25]. Human pose data has recently been used in Rifle and Sniper detection and threat assessment. Abruzzo et al. [26] proposed a method for identifying people, rifles, and snipers in images and then assessing the threat level of the person poses based on their body posture. However, the main limitation of this work is that the Rifle and Sniper detector used limits the detection performance (in this case YOLO). Human pose information is not taken into account in the Rifle and Sniper detection step. Basit et al. [27] proposed a classification method for human-Rifle and Sniper pairs. Human, Rifle, and Sniper are detected separately, as in previous work. Then, each detected human is paired with each detected Rifle and Sniper and, finally, a neural network is trained to classify these paired human-Rifle and Sniper bounding boxes into two classes: "carrying Rifle and Sniper" and "not carrying Rifle and Sniper". This method can be used to eliminate false Rifle and Sniper detections, but the detection performance is still limited by the Rifle and Sniper detectors used, and the human pose cannot help to reduce the number of false negatives. Salido et al. [28] recently investigated how including body pose information (skeleton key points and limbs retrieved by a pose detector) in the input images as a pre-processing step can improve Rifle and Sniper detection performance. An approach to improving a Rifle and Sniper detector through integration with the human pose was recently introduced in Velasco-Mata et al. [29]. This method employed a visual heat map representation of both the pose and the weapon location, resulting in a final greyscale image that indicates potential Rifle and Sniper regions on the image.

3. Materials

This section describes the datasets used to evaluate the proposed method's performance. Images were collected from various sources, including public Rifle and Sniper datasets,

YouTube clips, and even synthetic images obtained from video games, in order to consider different contexts and image features.

3.1. Datasets for Rifle and Sniper

The proposed method is intended for use in CCTV and UAV surveillance systems across a wide range of scenarios. Unfortunately, most public Rifle and Sniper datasets contain weapon profile images that take up the entire image and have a uniform background, which differs significantly from the type of images captured by surveillance cameras. Surveillance scenarios are typically distinguished by a large distance between the subjects being recorded and the camera, as well as low image resolution or quality and poor lighting conditions. Salazar Gonzalez et al. [30] recently published a new dataset that combines CCTV and UAV images from a real video surveillance system with synthetic images generated using the Unity game engine. However, the CCTV and UAV images in this dataset are unnaturally posed for Rifle and Sniper attacks or mass shootings. On the other hand, some public datasets or parts of them that are realistic enough can be found. The first dataset used in our study is made up of 665 640x480 images extracted from videos in the Snipers Movies Database [14]. In these clips a man is holding a Rifle and Sniper in a few shooting poses in an roof area. CCTV and UAV scenarios are well represented by camera distance, image resolution, and lighting conditions. Figure 1 shows two examples of images from this dataset

Figure 1: Sample images from Sniper Movies Database

In addition, 300 512x512 images were obtained from the publicly available Monash Snipers Dataset [31] for testing purposes. These images depict various CCTV and UAV scenarios with people holding Rifles and Snipers in various body poses. Figure 2 shows two examples of images from this realistic dataset.

Figure 2: Sample images from Môn ash Guns Dataset

YouTube videos (3.2) another good place to look for videos of people carrying or holding weapons and/or shooting is YouTube. As in the previous case, finding real CCTV and UAV footage of Rifle and Snipers is difficult. Nonetheless, there are videos of shooting practice sessions that are appropriate for our needs. This dataset contains 952 1920x1080 images extracted from 12 YouTube clips. There are various camera positions, background scenarios, shooting poses, and lighting conditions in these videos. Figure 3 depicts two images from this dataset.

Figure 3: Sample images from YouTube dataset

3.2. Video game images

Computer-generated video game images Games can also be used to generate new data for this task. It is possible to recreate representative situations or scenarios using specific video games and then extract videos or images. In this case, a synthetic dataset was created on a PC platform using the popular shooter video game Watch Dogs 2. In-game videos can be recorded from various camera locations, distances, or angles using the novel NVIDIA Ansell feature2. In this manner, four video sequences were captured, each of which performed a full camera rotation around the main character with two different heights in various shooting animations. Finally, 650 3840x2160 images were obtained from these video sequences. Figure 4 depicts two examples of images from this dataset. 3.4. Data enhancement and dataset splitting a large and representative dataset is required to achieve good performance in deep learning and Yolo-NAS and Yolov7 Pose-based novel object detectors.

Figure 4 Image of terrorist having Sniper/ Rifle from UAV

4. Methodology

The various steps involved in the proposed method are detailed in this section, beginning with the input image and ending with the final Rifle and Sniper detections. 4.1. Human pose estimation the first step is to gather the human pose information from the input image. This was accomplished using the YoloV7 Pose framework and COMBINATION OF Yolo-NAS and Yolov7 Pose (see Figure 6) [34]. YoloV7 Pose is an open-source multi-person pose estimator that can predict 2D keypoints and keypoint associations with high accuracy and low inference time. This step predicts a set of 25 2D keypoints for each person in the image, as well as predicted confidence for each of them. These key points include the human body position information (neck, shoulders, elbows, wrists, etc.) required to define each person's pose. 4.2. Extraction of the hand region The

hand regions for each detected person are inferred and extracted in the second step using the collected pose information. The positions of the elbows and wrists, as well as the distances and directions between them, are used to generate a set of red bounding boxes around all of the hand regions in the input image (see Figure 4). The YoloV7 Pose confidence score for each keypoint is used to filter out incorrect or inaccurate detections, and an intersection over union (IoU) threshold between the predicted bounding boxes is checked to avoid overlapping areas (e.g., a Rifle and Sniper held with both hands is considered as a single region, since both bounding boxes are overlapping)

4.1. Hand region classification for this stage

A convolutional neural network was trained to classify previously generated hand regions as Rifle and Sniper or non-Rifle and Sniper areas based on the presence or absence of a Rifle and Sniper within the region (see Figure 6). The selected network was Darknet-53, the backbone feature extractor used in the YOLO-NAS object detector. This hand region classifier will now be referred to as HRC. The dataset used to train the hand region classifier included 6177 images derived from hand areas extracted from the 3000 training images described in Section 3. These regions were labelled automatically by comparing the hand areas to the actual Rifle and Sniper locations. Instead of using the IoU score, we used the overlap measure proposed by Velasco-Mata et al. [29]: intersection over minimum area (IoMin), as shown in Equation 1. Ground truth Rifle and Sniper locations are typically smaller than hand bounding boxes. Because bounding boxes of different sizes are not penalized, this metric allows for a more accurate overlap measurement in this scenario. [29] Contains additional information.

If the regions overlap by 0.5 IoMin, the hand area is designated as a Rifle and Sniper area. If there is no overlap or the treshold is less than 0.5 IoMin, the hand area is labelled as no Rifle and Sniper. Each hand region was also resized to a fixed size of 256x256. The model was trained in 60 epochs with a batch size of four and the Adam optimization algorithm as the loss function.

4.2. Pose combination method

Consideration was given to a further modification of the Yolo-NAS with Yolov7 Pose method described in the previous section. In this instance, the network is modified to combine the hand region image with YoloV7 Pose and Yolo-NAS and Yolov7 Pose -obtained human pose information. This was done to aid the region classifier by exploiting the correlation between the individual's pose information and the region's classification. Pose data is used to generate binary images of fixed size 512x512 for each detected person in the input image, including the drawing of key points and connections. In addition, a normalization procedure is used to remove variable factors such as camera distance and absolute position from the image in order to focus solely on the relative position between the key points.

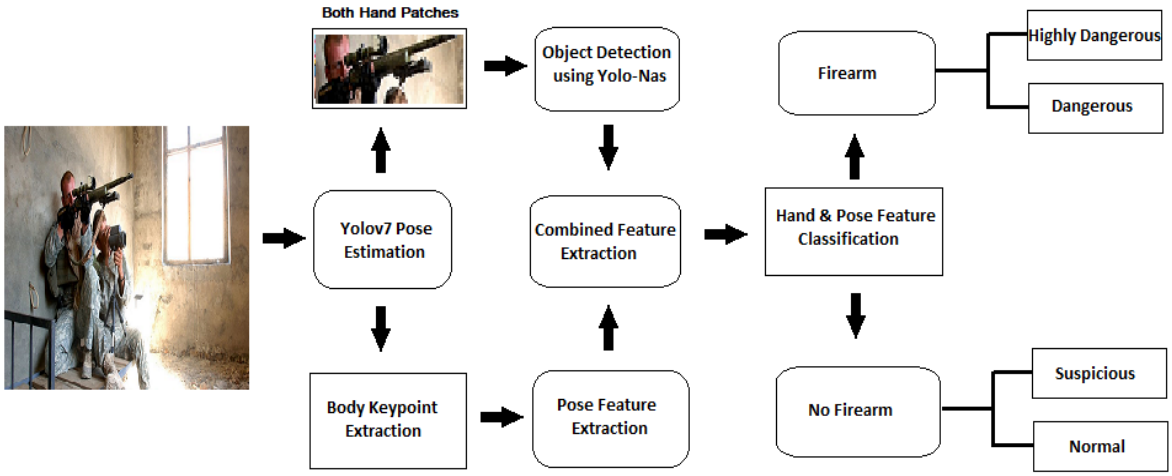

Figure 5: Sniper & Rifle Detection using pose estimation & Yolo-NAS and Yolov7 Pose

4.3. Bounding box prediction

The last step of the proposed method consists of generating the Rifle and Sniper predictions in the image. Each hand region of each detected person is passed through the YOLO-NAS to obtain a class label (Rifle and Sniper vs no-Rifle and Sniper). Then, the bounding boxes of the regions classified as Rifle and Sniper are included in the output list of predicted Rifle and Snipers (see Figure 6).

Figure 6: image of terrorist having rifle

5. Results

This section describes the outcomes of the tests conducted to assess the performance of the proposed method. Precision, Recall, and Average Precision are commonly used to evaluate object detection models [35]. These metrics are actually based on True Positives (TP), False Positives (FP), and False Negatives (FN) (FN). The overlap between the ground truth bounding boxes and those predicted by the detector is taken into account when calculating these values. In the same manner as in the automatic labelling process for the training of the hand region classifier (Section 4.3), the IoMin is used to calculate the overlap between the predicted bounding boxes and the

ground truth data due to the size difference between them. The proposed pose-combined approach (HRC+P) was evaluated against three distinct Rifle and Sniper detectors:

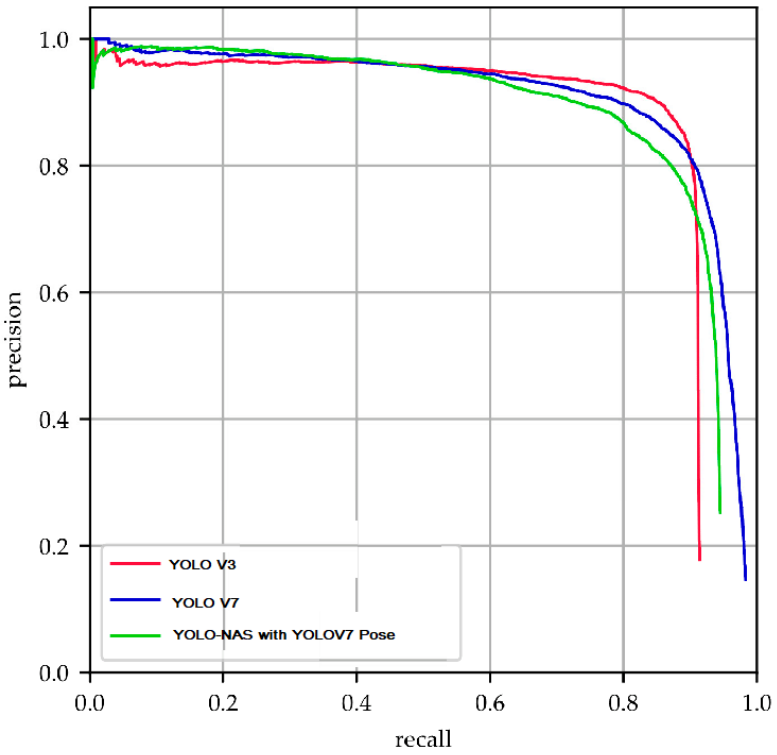

Figure 7: Precision-Recall curves obtained for the original images.

Methods	Precision (%)	Recall (%)	F1 (%)	AP (%)
YOLOV3	88	78	82.7	85.84
YOLOV7	87	81	83.89	86.12
YOLO-NAS with YOLOV7 Pose	93	88	90.43	90.89

- **YOLOv3 [23]:** YOLOv3 is one of the fastest and most accurate deep learning-based object detectors. The Darknet-53 CNN backbone is used as feature extractor, which provides an interesting baseline for comparison.
- **YOLOV7 [28]:** The newest YOLO algorithm surpasses all previous object detection models and YOLO versions in both speed and accuracy. It requires several times cheaper hardware than other neural networks and can be trained much faster on small datasets without any pre-trained weights.
- **YOLO NAS with YOLOV7 Pose:** To check the effect of including the 2D human pose information in the hand region classifier, the hand region processing branch without pose combination is taken for comparison. All methods were trained and tested using the datasets described in Section 3.

6. Conclusions

The 2D human pose is utilized quite frequently in activities such as the recognition of actions or gestures. However, the majority of the proposed methods for the detection of threats or dangerous objects like firearms are based

solely on the outward appearance of the objects, without taking into account the human pose or any additional information. The purpose of this study is to present a revolutionary way that combines in the same architecture the visual appearance of the Rifle and Sniper with the information regarding the 2D human position. There are some instances in which the object cannot be viewed accurately due to factors such as the distance between the camera and the subject, inadequate lighting, or either partial or complete occlusion. In these circumstances, the human body pose contributes to the detection of the presence of rifle and sniper fire, which otherwise would not be identifiable in the absence of this supplemental data. On the other hand, because the pose information is used to classify only the hand regions of the people detected, it is possible to remove false positives that may appear in other locations of the image. This is because the pose information is used to classify only the hand regions of the people detected. The evaluations carried out with the various datasets demonstrate that the proposed method which makes use of the pose combination obtains superior results in every circumstance. It is also fascinating to consider the fact that the metrics for the smaller versions of the photographs are even higher than those for the full-size versions of the images. The automatic and real-time detection of rifles and snipers in video surveillance images captured by CCTV and UAVs is still an open problem, and there is room for improvement in this area. The authors have expressed their hope that the work that has been presented can serve as a source of motivation for new approaches that are based on 2D human pose information, with the end goal of improving the overall detector performance in applications of this kind. Last but not least, it is important to keep in mind that in real-world scenarios, common handheld objects like cell phones, keys, or wallets may be a significant source of false positives or misclassifications. In future work this aspect will be addressed with more specific methods.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Human participants and/or animals

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding:

This study does not contain any funding

References

[1] Livak, K. J., & Schmittgen, T. D. (2001, December). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2CT Method Methods, 25(4), 402–408. https://doi.org/10.1006/meth.2001.1262

[2] Place, B. J., & Field, J. A. (2012, June 15). Identification of Novel Fluorochemicals in Aqueous Film-Forming Foams Used by the US Military Environmental Science & Technology, 46(13), 7120–7127. https://doi.org/10.1021/es301465n

[3] Chang, C. C., & Lin, C. J. (2011, April). LIBSVM. ACM Transactions on Intelligent Systems and Technology, 2(3), 1–27. https://doi.org/10.1145/1961189.1961199

[4]Papastergiou, M. (2009, January). Digital game-based learning in high school Computer science education: impact on educational effectiveness and student motivation Computers & Education, 52(1), 1–12. https://doi.org/10.1016/j.compedu.2008.06.004

[5] Hardin, G. (1968, December 13). The Tragedy of the Commons, 162 (3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

[6] FitzhughLander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzhughUsesnke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky,

J., LeVine, McEwan, P., Morgan, M. J. (2001, February 15). Initial sequencing and analysis of the human genome Nature, 409 (6822), 860–921. https://doi.org/10.1038/35057062

[7] Hochreiter, S., & Schmidhuber, J. (1997, November 1). Long-Short-Term Memory Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Ren, S., He, K., Girshick, R., & Sun, J. (2017, June 1). Faster R CNN: Towards Real-Time Object Detection with Region Proposal Networks IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031

[9] Konapala, G., Kao, S. C., Painter, S. L., & Lu, D. (2020, September 21). Machine learning-assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15(10), 104022. https://doi.org/10.1088/1748-9326/aba927

[10] Johnson, S. A. (2018, February 26). Discussion of Change Needed Following the Shooting at Marjory Stoneman Douglas High School in Parkland Florida, A Brief Response. Journal of Forensic Sciences & Criminal Investigation, 7(5). https://doi.org/10.19080/jfsci.2018.07.555725

[11] Roosevelt, T. S. (1995, November 15). Preventing Rifle and Sniper Violence. Annals of Internal Medicine, 123(10), 813. https://doi.org/10.7326/0003-4819-123-10-199511150-00033

[12] Romero, D., & Salamea, C. (2019, July 24). Convolutional Models for the Detection of Rifle and Snipers in Surveillance Videos. Applied Sciences, 9(15), 2965. https://doi.org/10.3390/app9152965

[13] Personal Characteristics of Military Personnel and Law Enforcement Officers that Determine Psychological Readiness for Duty with Rifle and Sniper. (2018). LEX RUSSICA (РУССКИЙ ЗАКОН). https://doi.org/10.17803/1729-5920.2018.142.9.119-128

[14] Hans, A. S. A., & Rao, S. (2021, March 31). A CNN-LSTM BASED DEEP NEURAL NETWORKS FOR FACIAL EMOTION DETECTION IN VIDEOS. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES, 7(1), 11–20. https://doi.org/10.29284/ijasis.7.1.2021.11-20

[15] Weng, Z., Li, W., & Jin, Z. (2021, January 11). Human activity prediction using saliency-aware motion enhancement and weighted LSTM network. EURASIP Journal on Image and Video Processing, 2021(1). https://doi.org/10.1186/s13640-020-00544-0

[16] Bidirectional LSTM-. (n.d.). www.baeldung.com. Retrieved May 27, 2023, from https://www.baeldung.com/cs/bidirectional-vs-unidirectional lstm#:~:text=Bidirectional%20LSTM%20(BiLSTM)%20is%20a,utilizing%20information%20from%20both%20sides.

[17] Olmos, R., Tabik, S., & Herrera, F. (2018, January). Automatic handgun detection alarm in videos using deep learning. Neurocomputing, 275, 66–72. https://doi.org/10.1016/j.neucom.2017.05.012

[18] G. K. Verma, A. Dhillon, A handheld gun detection using faster R-CNN deep learning, in: Proceedings of the 7th International Conference on Computer and Communication Technology, 2017, pp. 84–88.

[19] IMFDB: Internet Movie Firearms Database, http://www.imfdb.org/wiki/Main_Page, 2020. Accessed: 20/07/2021.

[20] R. Olmos, S. Tabik, F. Herrera, Automatic handgun detection alarm in videos using deep learning, Neurocomputing 275 (2018) 66–72.

[21] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[22] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[23] A. Farhadi, J. Redmon, Yolov3: An incremental improvement, Computer Vision and Pattern Recognition (2018).

[24] A. Warsi, M. Abdullah, M. N. Husen, M. Yahya, S. Khan, N. Jawaid, Gun detection system using yolov3, in: 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), IEEE, 2019, pp. 1– 4.

[25] R. F. de Azevedo Kanehisa, A. de Almeida Neto, Firearm detection using convolutional neural networks., in: ICAART (2), 2019, pp. 707–714.

[26] B. Abruzzo, K. Carey, C. Lowrance, E. Sturzinger, R. Arnold, C. Korpela, Cascaded neural networks for identification and posturebased threat assessment of armed people, in: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), IEEE, 2019, pp. 1–7.

[27] A. Basit, M. A. Munir, M. Ali, N. Werghi, A. Mahmood, Localizing firearm carriers by identifying human-object pairs, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 2031–2035.

[28] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, O. Deniz, Automatic handgun detection with deep learning in video surveillance images, Applied Sciences 11 (2021). URL: https: //www.mdpi.com/2076-3417/11/13/6085. doi:10.3390/app11136085.

[29] A. Velasco-Mata, J. Ruiz-Santaquiteria, N. Vallez, O. Deniz, Using human pose information for handgun detection, Neural Computing and Applications (2021).

[30] J. L. Salazar Gonz´alez, C. Zaccaro, J. A. Alvarez-Garc´ıa, L. M. Soria-Morillo, F. San- ´cho Caparrini, Real-time gun detection in cctv: An open problem, Neural Networks 132 (2020) 297 – 308.

[31] J. Lim, M. I. Al Jobayer, V. M. Baskaran, J. M. Lim, J. See, K. Wong, Deep multi-level feature pyramids: Application for non-canonical firearm detection in video surveillance, Engineering Applications of Artificial Intelligence 97 (2021) 104094.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[34] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y. A. Sheikh, OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2019) 172–186.

[35] R. Padilla, S. L. Netto, E. A. B. da Silva, A survey on performance metrics for object-detection algorithms, in: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, pp. 237–242.