



Modeling Objects with Artificial Intelligence Based Image Processing Techniques: Object Detection with Mask R-CNN

¹Ömer Faruk EREKEN, ^{*2}Çiğdem TARHAN

¹ Department of Management Information Systems, Dokuz Eylül University, Türkiye, omerfarukereken@gmail.com 

^{*2} Corresponding Author, Department of Management Information Systems & BİMER, Dokuz Eylül University, Türkiye, cigdem.tarhan@deu.edu.tr 

Abstract

Object detection and classification on digital images is an area of great importance in the digitalizing world. After deep learning methods started being implemented for object detection, classification and segmentation a rapid development has been observed in the field. Mask R-CNN is one of the most successful methods in the field and can be used for detection and segmentation purposes for many different objects. Our study focuses on the use of Mask R-CNN for weapon detection, specifically handguns. Today, there are many cameras in public areas and detecting weapons through these cameras before a forensic incident can provide great advantages. Our model achieved a mean average precision (mAP) of 0.78 in the detection of handguns on test data. Our findings demonstrate the potential of deep learning in security by detecting threats in images and live videos.

Keywords: Mask R-CNN; Deep Learning; Handgun Detection; Object Detection; Instance Segmentation

1. INTRODUCTION

In the digital world that we are living in today, digital images hold a very important position. Defining these images in a way that computers can understand has become increasingly valuable and recent advancements in deep learning, especially Convolutional Neural Networks (CNN), have made this process easier.

Making digital images understandable by machines is part of 'image processing'. If we set aside conceptual discussions, we can define image processing as a set of methods for analyzing, enhancing, and editing digital images [1].

While image processing consists of a broad area, our study specifically focuses on image classification, object detection and image segmentation, which are part of this field. Image classification is the process of classifying the images into what they represent. For example, if an image represents a class or not.

Object detection is the process of determining if there is a specific object on an image and the position of it. Image segmentation takes this one step further and does a pixel level classification and highlights the objects on the image. Image segmentation divides into two types; semantic segmentation which does an object category level segmentation and instance segmentation where a segmentation for every single object is conducted [2].

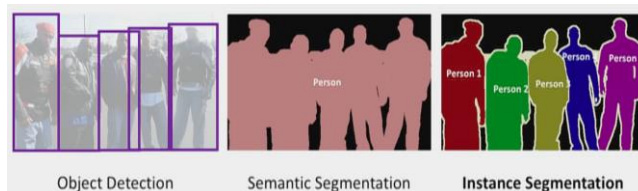


Figure 1. Detection and Segmentation [3].

The history of object detection can be divided into two main stages, with the critical point being the introduction of deep learning to the field. After the entrance of deep learning, everything changed drastically. In the first stage, we observed object detection algorithms that relied on handcrafted features. However, as the field became stagnant with handcrafted algorithms in the 2010s, a revolutionary approach was about to emerge. In 2014, R. Girshick et al. proposed their study on Regions with CNN features (R-CNN), which marked the beginning of a new era in object detection [4].

R-CNN is made from three components, category-independent region proposal generator (utilizing selective search), CNN which is extracting fixed-length feature vectors from every region, class-specific linear Support Vector Machine (SVM) [5]. R-CNN was a great breakthrough for its time, but it had its downsides like; a complex multi-stage pipeline, training being expensive, object detection and the whole process being too slow [6].

Aware of the weaknesses of R-CNN, R. Girshick proposed Fast R-CNN, which solved the drawbacks of R-CNN and enhanced its speed and accuracy. Fast R-CNN is a single-stage method where an image and a set of object proposals are taken as input. First, the whole image is passed into a CNN and a feature map is achieved. After that, every object proposal is processed with a Region of Interest pooling layer and a fixed-length feature vector is extracted from the feature map from the first stage. At the end, these feature vectors are sent to a SoftMax probability classifier and a bounding box regressor [6].

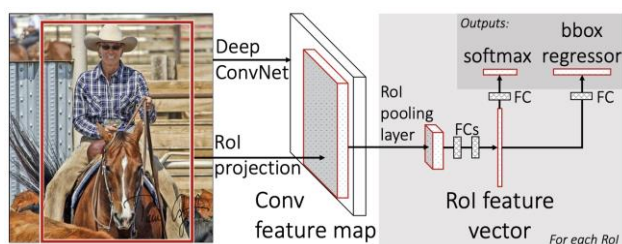


Figure 2. Fast R-CNN Architecture [6].

R-CNN and Fast R-CNN use selective search as a region proposal algorithm. Even though ROIs were introduced by Fast R-CNN, the computation of region proposals remained a bottleneck for object detection. To solve this Ren S. et.al. introduced the Region Proposal Network (RPN) with Faster R-CNN. In their study, they discovered that the convolutional feature maps employed by region-based detectors could also be utilized to generate region proposals. Briefly, RPN generates region proposals, which are subsequently passed to the ROI module. The cost-effectiveness of RPN was a significant advancement in object detection algorithms [7].

And the final algorithm we will mention, which we are using in this study, is Mask R-CNN. Mask R-CNN is a network that integrates instance segmentation capabilities into Faster R-CNN. Mask R-CNN has small differences compared to Faster R-CNN, but these small differences matter a lot. The first critical difference is the use of a new layer called ROI Align instead of ROI pooling, which enables pixel-to-pixel alignment between the network's input and outputs. Secondly, the mask and class prediction are separated [8].

Mask R-CNN can be used for a wide range of instance segmentation and object detection tasks. This versatility can be found through a simple academic search. Our focus in this study is weapon detection, specifically handguns for security purposes. With the increasing use of security cameras in public areas, the ability to detect weapons before they are used for criminal activities can be highly beneficial.

In this study we are using Mask R-CNN, which is the latest and most improved version of the R-CNN versions, thus our focus will be on this algorithm and its evolution. However, we do find value in briefly mentioning the You Only Look Once (YOLO) algorithm, which is one of the competing and popular algorithms in object detection and instance segmentation.

YOLO was introduced in 2016 as a fast (when compared to other algorithms), single step object detection model. The

study aimed to build a model which works like a human glance, in other words to look at an image and directly detect objects. Other object detection systems use multiple components and complex pipelines which make the process slow and hard to optimize. In contrast, YOLO is rather simple, using a single convolutional network which predicts bounding boxes and class probabilities, and then filters the detections based on the model's confidence [9]. Over time, YOLO received a lot of interest and made big progress in accuracy and capabilities, including instance segmentation. Many different variants have been published and the model continues to be improved by researchers [10]. YOLOv8 model was shown to outperform Mask R-CNN, with higher precision and recall in less time [11].

2. LITERATURE REVIEW

There have been many studies using different methods to detect weapons in images. One of them is referenced in our dataset source, which focuses on an automatic handgun detection alarm system using CNN. The results of the study indicate that Faster R-CNN has shown the most promising outcomes. In this study, the training set is initially constructed using the outcomes from a VGG-16 based classifier to minimize the number of false positives. The study reports a recall rate of 100% and a precision rate of 84.21% [12].

Another similar study applied three CNN based models; YOLOv3, RetinaNet and Faster R-CNN to detect handguns in images. To reduce the number of false positives, they incorporated pose information on how handguns are held, which proved to be effective for one model. While YOLOv3 achieved the best precision and F1 scores without the added pose information, Faster R-CNN received lower results compared to RetinaNet and YOLOv3. With the inclusion of pose-related information, Faster R-CNN and RetinaNet showed decreased performance, whereas YOLOv3 displayed improvement. Overall, the highest rates were achieved with a 97.23% recall rate for RetinaNet without pose information and a 96.23% precision rate for YOLOv3 with pose information [13].

One study utilizes Mask R-CNN in their threat detection system, designed to detect suspicious objects in the images captured through cameras. They primarily employ CNN for classification on live camera images and subsequently use Mask R-CNN for instance segmentation on the cloud side. Regarding Mask R-CNN, they have provided only the overall average classification accuracy for classification purposes, which stands at 93.09% [14].

With the aim of preventing crimes before they happen, a study has utilized Mask R-CNN to detect guns in surveillance images. Their system takes an input image, applies preprocessing techniques like resizing, flipping etc. then they apply image sharpening with Gaussian Deblur technique and finally detect the mask with their trained model. Contrary to the popular evaluation techniques of Mask R-CNN, this study utilizes classical evaluation techniques. They achieved an F1 score of 84.69% with Mask R-CNN [15].

In a recent study called “Weapon detection system for surveillance and security” Yolo V5 is used for weapon detection and Mask R-CNN is used for instance segmentation. Before proceeding with the model training, various data augmentation and preprocessing methods are being applied. The study achieves 90.66% detection accuracy (DC) and 88.74% mean intersection over union (mIoU) [16].

All studies expressed above are focused on detecting weapons through normal camera images but there are also studies which focus on finding concealed weapons which are also very critical to detect forensic incidents. One of these studies tries to detect pistols from thermal images using deep learning. They have evaluated several deep learning algorithms for classification and segmentation. While the best result for detecting the pistols was achieved using a VGG 19-based convolutional neural network with an F1 score of 84%, for the second module which consisted of classification and bounding box detection, Yolo-V3 achieved the highest mean average precision of 95% [17].

Another area where deep learning is used for weapon detection is in X-ray images. One of the studies we examined introduced an anchor-free convolutional neural network (CNN) approach to detect weapons in X-ray baggage images. By eliminating the need for preset anchor box sizes and thus reducing computational complexity, the method demonstrates robust performance in detecting knives and handguns. By comparing different mainstream anchor-free and anchor-based methods the study has revealed that anchor-free methods YOLOx, Objects as Points and ExtremeNet have great performance in weapon detection on X-ray images [18].

All in all, there is continuous development in the use of deep learning techniques for security and surveillance. As seen in studies [12], [13], [14], [15], [16], [17] and [18], the aim is to build a system that will enhance public safety and prevent crimes through image processing techniques. The increasing number of studies suggest that there will be rapid progress in the field, and soon, using deep learning for safety will become common.

Another valuable point to mention is the new, popular Large Language Model (LLM) based tools. Even though LLMs are text-based and don't have a direct impact on object detection, as AI systems that understand user inputs, these models will undoubtedly enhance the user input/request in image processing.

3. MATERIALS AND METHODS

The first stage of our work was to create a comprehensive dataset that includes both labeled data for classification and segmentation purposes. To accomplish this, we utilized a pre-labeled dataset consisting of 3000 handgun images. These images were selected from the internet, representing at least one handgun in diverse situations [12]. The only problem about this dataset was that it only had bounding box annotations which can be used for object detection but not for instance segmentation. To solve this, we created a new dataset out of the mentioned dataset, containing 700 images

(500 training, 100 validation, 100 test), annotated in COCO-style format.



Figure 3. Dataset Image Examples

The data set used for this study contains images from various conditions, however, for a real-world project the training set would need to be expanded to contain images from all types of possible conditions with objects which might resemble a gun but are not. Since this study was for educational purpose with limited resources we focused on training a prototype model.

In the original paper of Mask R-CNN it's stated that the code is made available on GitHub [5]. This code is written in python, and it's powered by the deep learning framework Caffe2 which is now deprecated and transferred to PyTorch repository. In this study we are using Mask R-CNN's deployment through Python 3, TensorFlow and Keras which can be found on a different GitHub repository [19]. The model is based on Feature Pyramid Network (FPN) and a ResNet101 backbone.

Data is crucial for training successful deep learning models, but sometimes obtaining sufficient amount of data can be challenging. To solve this issue, scientists have built a solution known as transfer learning. In transfer learning, you can access the learned weights from previous deep learning studies and enable your model to start training on your data after gaining knowledge about other classes. In our study we used the weights of Microsoft Common Objects in Context (Coco) dataset trained model for Mask R-CNN. Microsoft Coco is a data set which contains images from 91 different objects [20].

The evaluation of Mask R-CNN is different than standard deep learning algorithms. In Mask R-CNN we have object classification and segmentation predictions to evaluate. As stated in [13], popular object detection competitions have used mean average precision (mAP) as the primary evaluation metric for the models. We can briefly say that mAP is the mean of estimated area under the precision-recall curve. mAP value is used in multiclass detection problems where the Average precision (AP) value is averaged for all the classes. AP is an approximation of the area under precision-recall curve and it's obtained from the equation (1) by interpolating the curve values. $P(\tilde{r})$ in (2) represents the precision where the recall is \tilde{r} .

$$AP = \sum_{n=0} (r_{(n+1)} - r_n) P_{interp}(r_{(n+1)}) \quad (1)$$

in which:

$$P_{interp}(r_{(n+1)}) = \max_{\tilde{r}: \tilde{r} \geq r_{n+1}} P(\tilde{r}) \quad (2)$$

To determine if a prediction is True Positive (correct), False Negative (undetected) or False Positive (incorrect)

confidence score and Intersection Over Union (IoU) values are used. Details for the evaluation of a Mask R-CNN model can be found in [21], [22] and [13]. We used the functions from [19] to calculate the mAP.

In order to enhance our handgun detection system, we used the python OpenCV library to conduct real-time predictions on streaming videos. This approach enabled generations of predictions directly from the video captured by our laptop camera. If this system gets implemented on security cameras it can provide efficient and prompt analysis to detect handguns.

4. FINDINGS

Our model trained on the coco-style formatted dataset gave us 0.81 mAP on training data, 0.78 mAP on validation and test data in 25 epochs. The instance segmentation was satisfying. You can find some examples of our predictions on the test data below.

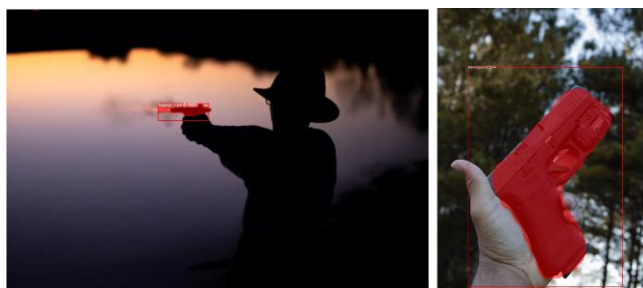


Figure 4. Model Prediction Examples

The examples above are from the predictions made on images. Then we ran tests on live video. The results were also satisfying as seen on the screenshots below.

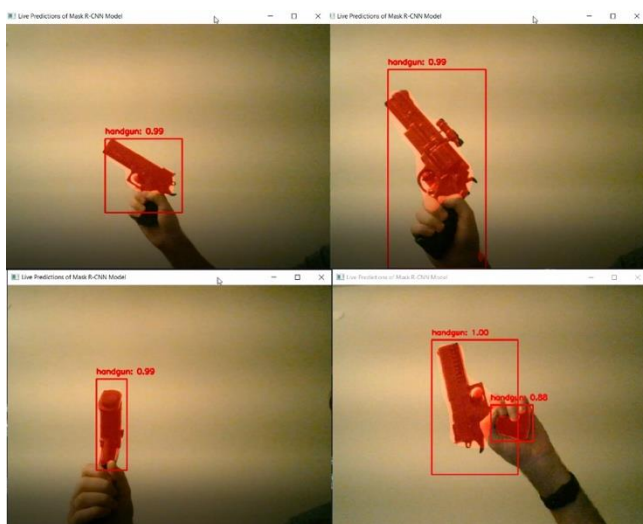


Figure 5. Predictions on Real-Time Video (With Toy Guns)

5. DISCUSSION AND CONCLUSIONS

It is obvious that deep learning is going to be used even more extensively in security issues in the future. A successfully trained AI system can detect security problems in seconds, which is usually not feasible for a person monitoring dozens of cameras, often becoming fatigued after hours of

surveillance. The aim of Management Information Systems (MIS) is to help people and managers make decisions using the correct technologies. Using Mask R-CNN for security purposes is a great example of helping people through technology.

It's important to note that, although AI seems to bring great value for security, it may raise concerns regarding human rights. Continuous surveillance by AI will need strict regulations to prevent it from being controlled or used by wrong hands for improper purposes.

In this study, we aimed to demonstrate an example of how security threats can be detected both from images and live videos. In future studies our aim is to enhance our system's ability to detect a wide range of weapons in challenging environments and generate alarm signals for security forces through real-time video analysis.

Author contributions: Ömer Faruk EREKEN: Methodology, data processing, writing-original draft preparation; Çiğdem TARHAN: Conceptualization, methodology, writing-reviewing and editing.

Conflicts of interest: The authors declare no conflicts of interest.

Ethical Statement: This article is an expanded version of the paper titled 'Modeling Objects With Artificial Intelligence Based Image Processing Techniques: Handgun Detection With MASK R-CNN' presented at the 10th International Conference on Management Information Systems (IMISC 2023) held on 18-20 October 2023.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

- [1] The Editors of Encyclopedia Britannica, "Image processing", Encyclopedia Britannica, Accessed on: Feb. 27, 2023. [Online]. Available: <https://www.britannica.com/technology/image-processing>
- [2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523-3542, 2021. [Online]. Available: <https://arxiv.org/pdf/2001.05566.pdf>
- [3] Computer Vision Foundation Videos, "Mask R-CNN," YouTube, Nov. 17, 2017. [Online]. Available: <https://www.youtube.com/watch?v=g7z4mkfRjI4>
- [4] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," 2023. [Online]. Available: <https://arxiv.org/abs/1905.05055v3>
- [5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. [Online]. Available: <https://arxiv.org/abs/1311.2524>
- [6] R. Girshick, "Fast R-CNN," 2015. [Online]. Available: <https://arxiv.org/abs/1504.08083v2>

- [7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," 2016. [Online]. Available: <https://arxiv.org/abs/1506.01497v3>
- [8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2018. [Online]. Available: <https://arxiv.org/abs/1703.06870v3>
- [9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
- [10] M. Hussain, "YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection," *Machines*, vol. 11, 2023, Art. no. 677. [Online]. Available: <https://doi.org/10.3390/machines11070677>
- [11] R. Sapkota, D. Ahmed, and M. Karkee, "Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments," *Artificial Intelligence in Agriculture*, vol. 13, pp. 84–99, 2024. [Online]. Available: <https://doi.org/10.1016/j.aiia.2024.07.001>
- [12] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66-72, 2018. [Online]. Available: doi: 10.1016/j.neucom.2017.05.012
- [13] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic handgun detection with deep learning in video surveillance images," *Applied Sciences*, vol. 11, no. 13, p. 6085, 2021. [Online]. Available: doi: 10.3390/app11136085
- [14] A. A. Ahmed and M. Echi, "Hawk-eye: An AI-powered threat detector for intelligent surveillance cameras," *IEEE Access*, vol. 9, pp. 63283-63293, 2021.
- [15] A. Goenka and K. Sitara, "Weapon Detection from Surveillance Images using Deep Learning," in 3rd International Conference for Emerging Technology (INCET), 2022. pp. 1-6. [Online]. Available: doi: 10.1109/INCET54531.2022.9824281
- [16] S. Khalid, A. Waqar, H. U. Ain Tahir, O. C. Edo, and I. T. Tenebe, "Weapon detection system for surveillance and security," in 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 2023. pp. 1-7. [Online]. Available: doi: 10.1109/ITIKD56332.2023.10099733
- [17] O. Veranyurt and C. O. Sakar, "Concealed pistol detection from thermal images with deep neural networks," *Multimed Tools Appl*, vol. 82, pp. 44259–44275, 2023. [Online]. Available: doi: 10.1007/s11042-023-15358-1
- [18] Y. Huang, X. Fu, and Y. Zeng, "Anchor-Free Weapon Detection for X-Ray Baggage Security Images," *IEEE Access*, vol. 10, pp. 97843-97855, 2022.
- [19] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," GitHub, 2017. [Online]. Available: https://github.com/matterport/Mask_RCNN
- [20] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2015. [Online]. Available: <https://arxiv.org/abs/1405.0312v3>
- [21] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 2020.
- [22] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit," *Electronics*, vol. 10, p. 279, 2021.