Evaluating R-CNN and YOLO V8 for Megalithic Monument Detection in Satellite Images

Daniel Marçal, Ariele Câmara [1], João Oliveira [2] and Ana de Almeida [1-3]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal
[2] Instituto de Telecomunicações, Lisbon, Portugal
[3] CISUC - Centre for Informatics and Systems of the University of Coimbra; Coimbra, Portugal

acaer@iscte-iul.pt

Abstract. Over recent years, archaeologists have started to use object detection methods in satellite images to search for potential archaeological sites. Within image object recognition, due to its ability to recognize objects with great accuracy, convolutional neural networks (CNN) are becoming increasingly popular. This study compares the performance of existing deep-learning algorithms for the detection of small megalithic monuments in satellite imagery, namely RCNN (Region-based Convolutional Neural Networks) and YOLO (You Only Look Once). Using a satellite image dataset and after adequate preprocessing, results showed that this is a feasible approach for archaeological image prospection, with RCNN achieving a remarkable precision of 93% in detecting these small monuments.

Keywords: object detection, satellite images, CNN, megalithic monuments, archaeology

1 Introduction

Object detection, a pivotal task in computer vision, has emerged as a crucial method for archaeologists to recognize specific monuments, thereby facilitating the prospection and the study of ancient societies. In the domain of satellite imagery, this task becomes especially challenging due to a myriad of factors [13], ranging from inherent spatial resolution constraints, where significant yet relatively small constructions like dolmens might be represented by a mere handful of pixels (around 15 pixels in this case), to issues of rotation invariance, with monuments appearing in any possible orientation. Additionally, accuracy is influenced by intraclass variations, where the visualization of the same object type, such as a megalithic monument, can vary based on environmental conditions, vegetation, shadow length, and soil type [6]. Beyond these technical challenges, the intensive task of data labelling is yet another challenging task. Nevertheless, deep learning presents a promising solution for object detection, and machine learning-based methods are becoming increasingly common, albeit in recognising easily detectable monuments [9].

Acknowledging the importance and intricacies of this problem, we delve into a performance analysis of different object detection pipelines to understand their capability

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

for identifying small megalithic monuments in satellite images, aiming to provide a tool for helping archaeologists to recognise monuments that have been traditionally difficult to detect. For this purpose, with the help of an expert, we collected a customized dataset featuring high-resolution images containing known dolmen sites [3]. The new annotated dataset intends to benchmark the speed and accuracy of the pipelines developed, striving for optimal detection and localization of small heritage monuments in satellite images. The information regarding the locations of the analyzed monuments comprising our dataset is available on Zenodo [5].

To optimize the dolmens' detection and classification process using satellite imagery, we evaluate and compare two recent renowned algorithms, RCNN (Region-based Convolutional Neural Networks) and YOLO (You Only Look Once), benchmarking their performance metrics, including running time and accuracy.

This paper is organized into five distinct sections: after this introduction, a brief review of related works is presented. Next, we describe an exploration of object detection methodologies, followed by a discussion of our results. The work finishes with the presentation of conclusions and the discussion of probable implications.

2 Case Study Area Characterization

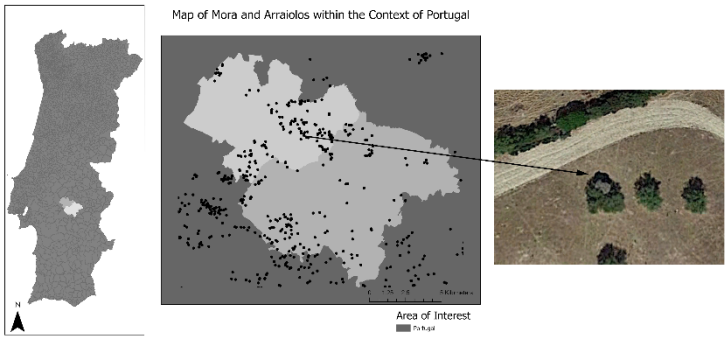

Figure 1: Map of Portugal highlighting the regions of Mora and Arraiolos in detail, as well as the dolmens in these regions. On the right, a representation of dolmen Lapeira 1 is shown: an aerial view from Google Earth.

The archaeological data used in this study focus on the Mora and Arraiolos regions in southern Alentejo, Portugal (Figure 1). Situated within the Ancient Massif, known for its granitic, schistose, quartzites, and other metamorphic rocks, these areas harbour significant clusters of known megalithic monuments, totalling 272 structures classified as dolmens according to the Portuguese *Portal do Arqueólogo*[14]. Dating back to the Neolithic/Chalcolithic period (4000/5000 BC), these dolmens primarily served as funerary sites, typically constructed from granite or schist, with diameters ranging from 2 to 5 meters [3]. While some are visible above ground, others are buried or integrated into modern constructions, making recognition challenging. To our knowledge, there has been no prior research on object recognition of these monuments in the analyzed region. This study stands out as it goes beyond simply detecting easily identifiable figures in satellite images.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

3 Literature Review

Remote sensing has been utilized as an indispensable tool for archaeologists for decades, and recent advancements in Deep Learning (DL) present new opportunities to enhance archaeological research methodologies, notably in object recognition within satellite images [4]. Despite growing interest in machine learning for site identification, its adoption in archaeology remains limited due to its complex nature, demanding computational expertise. Among the prevailing neural network architectures in contemporary technology, Artificial Neural Networks (ANNs) stand out. Notably, CNNs have garnered significant attention and are the most rigorously explored among the different techniques. CNN architectures can generally be categorized into two groups based on their approach to object detection: one-stage detectors, such as You Look Only Once (YOLO), and two-stage detectors like Faster Region-based CNN. In the domain of archaeology, the primary use of Remote Sensing Images (RSI) revolves around detecting distinct ground structures, examples of which include burial mounds, tells, rectangular enclosures, charcoal burning platforms, and qanats [4].

In recent years, studies have increasingly utilized CNNs to recognize archaeological structures in RSIs. Among these architectures, the RCNN has been described as an ideal solution for high-precision object detection tasks. For instance, Caspari & Crespo (2019) employed a CNN to detect Early Iron Age tombs within the Eastern Central Asian steppes using optical satellite imagery. The authors' findings underlined the superior performance of CNNs in RSIs analysis, achieving an impressive accuracy of 0.99 (F1-score) in images without the presence of tombs, contrasting with a 0.91 score in their presence [8]. This study exemplifies the prowess of CNNs in precisely detecting archaeological landmarks even from satellite imagery. Another popular approach, YOLO, has also been utilized, demonstrating faster detection rates and reduced false positive detections. For example, Canedo et al. (2023) used YOLO to detect burial mounds, achieving a positive detection rate of 72.53% [7]. This method contrasts with traditional CNN approaches and offers a faster alternative. It's crucial to note that determining the best model across different training sets may vary, and other algorithms could outperform in distinct tests. In comparison with our approach in this paper, Caçador (2020) analyzed the same monuments using a different dataset and methodology, focusing on the hyperspectral signature of how dolmens appear in satellite images. This analysis revealed the challenge of discerning these signatures due to the similarity between the surrounding terrain and the monument itself. Despite utilizing panchromatic and multispectral images, identification was feasible, albeit with a high false positive rate of 87.2% [2].

While advancements have been made, challenges like false positives and limited data persist in archaeological object detection. For instance, image enhancement techniques, including rotation, flipping, and augmentation, have been employed to improve object detection in challenging environments and expand datasets for analysis [8]. Additionally, techniques such as Location-Based Ranking and Bagging have been utilized to mitigate false positives. Despite these improvements, automated results still require refinement to consistently match human expertise, highlighting ongoing challenges and the need for further advancements in the field. Current efforts are focused on detecting niches or those structures deemed 'easy' to recognize [9], but future research may explore analyzing smaller or less identifiable monuments as a potential area of interest.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

4 Methodology

After collecting all the dolmen locations within the area to be covered, the next step was to obtain their satellite images. For this study, all the images were gathered from Google Earth Pro and corresponded to the same monuments analyzed in a previous work conducted by Câmara (2017), where the author performed a photo interpretation analysis [3]. These images were then extracted and saved in 8k resolution, providing higher-quality images. Our coverage features 62 dolmens visible from a software perspective. We collected five images for each dolmen, changing the monument's position within each image and, therefore, the surrounding background also changed, giving a total dataset of 310 different images. It is important to note that the low quantity of images of monuments for visualization derives from the fact that these are millennia-old structures, many of which are either destroyed or not visible in satellite images.

To test the algorithm's response to an image that does not contain any dolmen, two more images, very similar in brackground but devoid of dolmens were added to the test set. The data was split as follows: the test set comprises seventeen images, fifteen of which depict three different dolmens in different locations, and two images that, to the best of our knowledge, do not contain any dolmen. The remaining 285 images were used to train the algorithms. The random selection ensures a representative data distribution in training and validation subsets. In the subsequent step, the images underwent a preprocessing stage, including image labeling, enhancement, and augmentation, facilitated by Roboflow [12], augmenting the dataset to 855 images. Following a review of the state of the art [11], we determined that the optimal approach for augmentation in this setting was to introduce random values for various types of augmentation until identifying the most effective set of modifications for improved results. Non-colour-based and color-based augmentations such as cropping, rotation, hue, saturation, brightness, and exposure were employed. Contrast enhancement was applied using Roboflow's histogram and adaptive histogram equalization. These adjustments accentuate local details, making darker or lighter regions of the image more discernible. Such enhancement is especially valuable in images with a high vegetation index.

After pre-processing the images, we transitioned to the modeling phase to train the selected algorithms. Our choices included YOLO version 8, the most recent version at the time, and Fast R-CNN. YOLO, known for its efficiency in rapidly detecting objects within images, and Fast R-CNN, renowned for its high precision in object detection, were deemed suitable candidates for our archaeological structure detection project. In our experimentation, we systematically explored nine different architectures, leveraging auto-tuning in each experiment to fine-tune the parameters and hyperparameters for optimal performance on the training data. Using Fast R-CNN models, we opted to use ResNet-101 and ResNet-50 as backbone networks, without pre-training, exploring three different network structures: Feature Pyramid Network (FPN), Dilated Convolutional Network (DC), and Convolutional Network (C). We conducted experiments using two types of training schedules, namely 1x and 3x. These models were trained using the custom dataset created on Roboflow, which was then converted to COCO format, facilitating training with frameworks like Detectron2. Google Colab was used for training, and we chose to set 5000 iterations to minimize the total loss and approach the optimal learning rate for each algorithm. The training time for each Faster R-CNN algorithm was approximately 40 minutes, while YOLO required only half this time. This process consumed 10 GB of RAM and 8 GB of GPU memory in Colab. While the backbone networks were not pre-trained, using COCO-formatted data enabled us to

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

leverage pre-trained weights for specific architectures to expedite convergence during training.

5 Results

For evaluation, the most common metrics were used. Precision (P) is calculated as the ratio of true positive (TP) predictions to the sum of true positives and false positives (FP): $P = \frac{TP}{TP+FP}$. F1-score, the harmonic mean of precision and recall(R), provides a balanced measure calculated by $F1 = \frac{2*P*R}{P+R}$, where R measures the model's ability to capture positive instances and is the ratio of true positive predictions to the sum of true positives and false negatives: $R = \frac{TP}{TP+FN}$. These metrics help assess the model's ability to detect objects while minimizing the false positives correctly, and the F1-score is specifically emphasized for its relevance in binary classification scenarios [1]. Table 1 presents the average precision and F1-score metrics that have been obtained for each of the trained models for the test set. Notably, the FasterRCNN model, utilizing a ResNet-50 backbone network with a DC network for structure and a 1x training schedule, achieved the best results for this test set, achieving a precision of 0.93 and an F1-score of 0.78. The choice of backbone architecture substantially impacts the model's ability to learn complex patterns, and these metrics provide insights into the strengths and weaknesses of each configuration.

Table 1. Performance results for all the architectures tested in terms of average precision and F1-score values.

Model	Average Precision	F1 Score
R_50_FPN_3x	0.67	0.51
R_50_FPN_1x	0.69	0.64
R_50_DC_3x	0.70	0.65
R_50_DC_1x	**0.93**	**0.78**
R_50_C_3x	0.61	0.57
R_50_C_1x	0.60	0.57
R_101_FPN_3x	0.71	0.64
R_101_DC_3x	0.74	0.71
R_101_C_3x	0.63	0.59
YoloV8	0.79	0.71

The training of this particular type of algorithm involved monitoring various metrics to identify signs of overfitting or underfitting during the defined epochs. Figure 2 illustrates the losses and accuracy metrics throughout the training process. Particularly, the classification loss, which typically measures the divergence between predicted class probabilities and actual class labels, was scrutinized. The objective during training was to minimize this metric across epochs, aiming for optimal model performance. Figure 2A shows a notable decrease in classification loss that becomes more stable after 4000 epochs, indicating convergence towards an optimal solution. This observation suggests

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

that the training process effectively learns the underlying patterns in the data, enabling the model to make accurate predictions. The loss stableness implies reduced sensitivity to minor fluctuations in the training data, indicative of a well-generalised model [10].

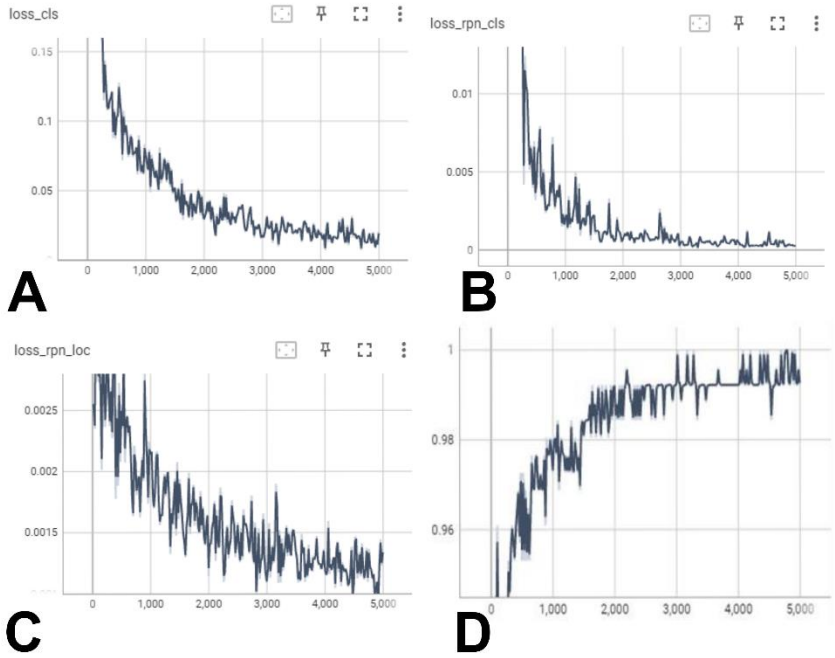

Figure 2: Figure 2 A-D depicts the plots of the train results in terms of loss and accuracy: (A) Classification Loss; (B) Classification loss in the Region Proposal Network; (C) Location loss in the Region Proposal Network; and (D) Classification Accuracy.

In the Region Proposal Network (RPN) of the faster RCNN algorithm (R_50_DC_1x), minimizing the classification loss allows the prediction of high objectness scores for anchors overlapping significantly with ground-truth object bounding boxes and low scores for anchors far distant from any object. This ensures that the model focuses on relevant regions likely to contain objects while ignoring background or irrelevant areas, which is particularly relevant given that each dolmen was tested against five different backgrounds [10]. Figure 2.B) illustrates that after approximately 2000 epochs, the model becomes more stable. This stableness suggests that the model has reached a point of diminishing returns regarding classification improvement. While this stabilisation is a positive sign, indicating the likely convergence of the RPN to a satisfactory level of classification accuracy, it's crucial to acknowledge that further training beyond this point may yield insignificant additional benefits and may even risk overfitting. Minimizing the localization loss is crucial because it ensures that the algorithm effectively learns to accurately predict the correct bounding box coordinates for the positive region proposals. Through improved training, the regression of the predicted bounding box coordinates for each positive anchor aligns more closely with the ground-truth bounding box. The successful training of the Faster R-CNN algorithm, exhibiting minimal losses and high accuracy, underscores the feasibility of implementing object detection models for small megalithic monuments in a rocky terrain through

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

remote imagery, making this a feasible approach for automating and enhancing archaeological prospection work.

In Figure 2.C), the analysis includes the localization loss in the RPN. However, the presence of numerous spikes suggests that there are instances where the model encounters challenges in precisely pinpointing object locations. These spikes may be due to various factors, including complex object geometries, augmentations, or variations in image backgrounds. In Figure 2.D), the classification accuracy for this algorithm was tracked by epoch, illustrating a consistent improvement trend. This means the model's growing adeptness in accurately classifying objects during training. While the Faster R-CNN algorithm, trained with minimal losses and high accuracy, demonstrates proficiency in object detection, it's noteworthy that YOLOV8, despite a lower confidence rate in true positive results, excels in minimizing false positive detections. This reduction in false positives is particularly advantageous for our overarching goal of providing a helpful tool for archaeologists in their prospection work. This trade-off bears practical implications, as Faster R-CNN models may be preferable for precise localization, whereas YOLO models could be advantageous in scenarios prioritizing false positive reduction. Moreover, the results underscore the influence of background contexts on confidence scores, emphasizing the importance of background diversity in training datasets to enhance the adaptability of object detection models in real-world scenarios. These findings stress the necessity of comprehensive evaluation considering environmental characteristics for robust detection performance.

6 Conclusions

The paper addresses the challenge of object detection in identifying dolmens in satellite imagery. Its primary contribution includes a set of 62 annotated high-quality images of dolmens in Portugal. Given data constraints, image augmentation and enhancement were crucial in increasing the dataset from 285 to 855 images, as well as highlighting the monuments, which can often be obscured by surrounding features. However, challenges persist due to the scarcity of expert-confirmed dolmen locations, resulting in a relatively small dataset. Evaluation of results highlighted the YOLOv8 model that, although showing lower confidence in true positives presented fewer false positives. Nevertheless, the Faster R-CNN model, despite the higher number of false positives, presents the lowest confidence rate in erroneous identifications.

Future research should prioritize the collection of a broader and diversified dataset for a more comprehensive evaluation assessment. Moreover, venturing into advanced or hybrid modeling techniques could improve accuracy in detecting dolmens in satellite images.

Acknowledgments. This work was partially supported by the Fundação para a Ciência e a Tecnologia, I.P. (FCT) through the ISTAR-Iscte project UIDB/04466/2020 and UIDP/04466/2020, through the scholarship UI/BD/151495/2021.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20

References

1. Andreas C. Müller and Sarah Guido: Introduction to Machine Learning with Python A Guide for Data Scientists. (2017).
2. Caçador, D.G.C.: Automatic recognition of megalithic objects in areas of interest in satellite imagery. ISCTE (2020).
3. Câmara, A.: A fotointerpretação como recurso de prospeção arqueológica. Chaves para a identificação e interpretação de monumentos megalíticos no Alentejo: aplicação nos concelhos de Mora e Arraiolos. Universidade de Évora (2017).
4. Câmara, A. et al.: Automated methods for image detection of cultural heritage: Overviews and perspectives. Archaeol Prospect. (2022). https://doi.org/10.1002/arp.1883.
5. Câmara, A.: Data Description. (ICCS). Zenodo. (2024) https://doi.org/10.5281/zenodo.10988490
6. Câmara, A., Batista, T.: Photo interpretation and geographic information systems for dolmen identification in Portugal: The case study of Mora and Arraiolos. Presented at the (2017). https://doi.org/10.23919/cisti.2017.7975890.
7. Canedo, D. et al.: Uncovering Archaeological Sites in Airborne LiDAR Data With Data-Centric Artificial Intelligence. IEEE Access. 11, (2023). https://doi.org/10.1109/ACCESS.2023.3290305.
8. Caspari, G., Crespo, P.: Convolutional neural networks for archaeological site detection – Finding "princely" tombs. J Archaeol Sci. 110, 104998, (2019). https://doi.org/10.1016/J.JAS.2019.104998.
9. Davis, D.: Theoretical repositioning of automated remote sensing archaeology: Shifting from features to ephemeral landscapes. Journal of Computer Applications in Archaeology. 4, 1, 94–109 (2021). https://doi.org/10.5334/JCAA.72/METRICS/.
10. Goodfellow, I. et al.: RegGoodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for Deep Learning. Deep Learning, 216–261.ularization for Deep Learning. Deep Learning. (2016).
11. Guo, H. et al.: Dynamic low-light image enhancement for object detection via end-to-end training. In: Proceedings - International Conference on Pattern Recognition. (2020). https://doi.org/10.1109/ICPR48806.2021.9412802.
12. Roboflow: Introduction - Roboflow Docs, https://roboflow.com/, last accessed 2024/04/18.
13. Tahir, A. et al.: Automatic Target Detection from Satellite Imagery Using Machine Learning. Sensors. 22, 3, (2022). https://doi.org/10.3390/s22031147.
14. Archaeologist's Portal, https://arqueologia.patrimoniocultural.pt/, last accessed 2024/03/04.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_20