

The Problem of Fragmented Occlusion in Object Detection

Julian Pegoraro¹, Roman Pflugfelder^{1,2}

¹ AIT Austrian Institute of Technology, ^{1,2} TU Wien

{julian.pegoraro|roman.pflugfelder}@ait.ac.at, roman.pflugfelder@tuwien.ac.at

Abstract. *Object detection in natural environments is still a very challenging task, even though deep learning has brought a tremendous improvement in performance over the last years. A fundamental problem of object detection based on deep learning is that neither the training data nor the suggested models are intended for the challenge of fragmented occlusion. Fragmented occlusion is much more challenging than ordinary partial occlusion and occurs frequently in natural environments such as forests. A motivating example of fragmented occlusion is object detection through foliage which is an essential requirement in green border surveillance. This paper presents an analysis of state-of-the-art detectors with imagery of green borders and proposes to train Mask R-CNN on new training data which captures explicitly the problem of fragmented occlusion. The results show clear improvements of Mask R-CNN with this new training strategy (also against other detectors) for data showing slight fragmented occlusion.*

1. Introduction

Automated surveillance at green borders has become a hot topic for European border guards. Border guards today face several challenges in protecting EU borders. One well known occasion in public is illegal migration which had its peak in 2015.

Border surveillance today limited to 2D imaging sensors consists of color and thermal cameras, mounted on poles or used as handheld cameras by the border guards. Innovating these technical systems by adding further capabilities of automatic inference such as the automatic detection of persons, vehicles, animals and suspicious objects in general will need to apply object detectors to such imagery.

However, video of green borders especially at EU borders show significant differences to typical imagery of video surveillance such as indoor video or



Figure 1: The problem of fragmented occlusion in object detection. Top Left: no occlusion (level L_0). Top Right: slight occ. (L_1). Bottom Left: moderate occ. (L_2). Bottom Right: heavy occ. (L_3) occlusion.

video taken in man-made outdoor scenes. For example, green borders are scenes showing dense forest, hills, harsh weather and climate conditions. Such scenes draw challenges to automated surveillance and raise several interesting research questions.

This paper considers a challenge for state-of-the-art object detection in green border surveillance which is the problem of through foliage detection. To the best of our knowledge, none of the current approaches for object detection allow the detection of objects through foliage. This problem raises an interesting scientific question, namely how to detect objects with fragmented occlusion? This problem is also different to the problem of partial occlusion in

object detection. Fragmented occlusion occurs by viewing objects behind tree and bush leaves. Contrary to partial occlusion, fragmented occlusion gives no clear view on minimal recognisable parts of the object [10] which is used to detect the object [7].

We show in this work that the state-of-the-art in object detection fails on fragmented occlusion even for the moderate case. For this, we created a new dataset (Figure 1) capturing people behind trees. We labelled nearly 40,000 images in three representative videos. This data raises new challenges on the labelling and evaluation which we only partially answer in this paper. For example, bounding boxes are the standard in current evaluation of detectors but such labels are hard to find in data that contains fragmented occlusion. As the state-of-the-art detectors deliver bounding boxes, fragmented occlusion poses new questions on the evaluation methodology.

Furthermore, we augmented Microsoft COCO¹ training data by occluding the ground truth masks similarly as leaves occlude people behind bushes and trees. We then show results on training Mask R-CNN [4] on this new data showing improvement of Mask R-CNN trained on the original data with slight fragmented occlusion.

2. Related Work

State-of-the-art object detection is based on deep learning. Two-stage detectors work by finding an intermediate step bounding box proposals [3, 2] on the feature maps of the backbone CNN. A region proposal network further improves efficiency [9, 4]. One-stage detectors regress the bounding boxes directly [8, 6] which is computationally efficient on GPUs but this approach is inherently less accurate as it assumes a coarsely discretised search space. Although these methods show usually excellent performance for fully visible objects, they break down in the case of fragmented occlusion. Fragmented occlusion has not been considered for object detection so far, however there is literature about this topic in the field of motion analysis [1].

3. Methodology

We created a dataset recorded in a forest consisting of three videos with a total of 18,360 frames and 33,933 bounding boxes which were manually defined by human annotators. These bounding boxes are di-



Figure 2: A training image from Microsoft COCO (<http://images.cocodataset.org/train2017/000000001700.jpg>). Top Left: the image. Top Right: Segmentation mask of the image. Bottom Left: image overlaid with artificial trees. Bottom Right: Mask of the overlaid image.

vided into four different occlusion levels including the unoccluded case (Figure 1).

Then, we extended the Microsoft COCO dataset by adding artificial trees as foreground to the images of objects (Figure 2). We chose this dataset, because it contains pixel-wise segmentation masks in the ground truth as well as a large number of different categories including the human person.

The underlying basic idea of our approach is to add artificial fragmented occlusion to Microsoft COCO and train Mask R-CNN on this new data. By this we can adapt the original distribution of data to the case of fragmentally occluded objects. Since we are only interested in humans, we apply this augmentation only to images containing humans and use only these images for training. The trees used for the augmentation are generated from real images we have obtained from the test data. The method generates whole artificial trees by randomly adding branches to previously manually segmented tree trunks. In total 14 such trunks are extracted from the test dataset. The branches attached to these trunks are also randomly generated by also adding a few manually segmented leaves.

The trees are placed in front of objects by randomly selecting the x-coordinate on which they will be placed and an angle at which the tree will be rotated. The calculated foreground is applied to the image and its negative mask is multiplied by the segmentation mask of the objects in the image. The Mask R-CNN model is then trained with the aug-

¹<http://cocodataset.org>

mented images. The selected backbone model is the Inception v2 [5] network. This network is selected for its faster computation.

4. Evaluation

To evaluate whether training with the augmented dataset is useful, the model trained on the augmented data must be compared with the model not trained on this data. However, the intersection over Union (IoU) measure is not meaningful in this case.

Standard evaluation metrics such as the mean average precision (mAP) define an IoU threshold (e.g. 0.5) and check whether a ground truth object and a detected object have an IoU value above this value. If this is the case, the detected object is defined as a True Positive (TP). If an object is detected but there is no respective ground truth with an IoU above this specific threshold, the detected object is defined as a False Positive (FP). If there is ground truth but no detected object with an IoU above the threshold, the object is defined as a False Negative (FN).

These evaluation methods cannot be easily applied to ground truth showing fragmented occlusion, because of the following two observations:

IoU too small: Since the data is based on fragmented detections, a detector can only detect parts of the person. An image where this problem occurs is shown in Figure 3. The bounding box is clearly a TP, based on the fact, that fragmented objects should be detected, but due to the occlusion by the branches of the tree, the whole body cannot be recognized. This leads to an IoU of only ≈ 0.2 .

Multiple detections: Another major problem with the standard evaluation metrics is that exactly one detected bounding box and one ground truth bounding box match. However, when handling fragmented objects, human heads and/or other body parts should be detected separately if body parts are covered. This creates the problem that parts of the body (like a head) is detected as well as the whole body. Figure 4 shows some examples.

To tackle these two problems, this paper proposes a different evaluation metric. For each bounding box in the evaluation data set, we calculate the maximum region in the image where there is no overlap with another ground truth bounding box. This region is then extracted and fed into the model. If the model detects an object, we define it as TP, otherwise as FN. To assess FPs, we create an additional dataset that represents the maximum region in an image without



Figure 3: Ground truth (green) and the detection (blue) vary substantially due to the occlusion effects.

overlap with any ground truth bounding box. We extracted in total 45,340 such regions with different aspect ratios, different parts of the image and at different time instants. In addition to FPs, we can also calculate the TNs using this evaluation metrics.

Figure 5 shows these results as recall vs. precision curve (ROC). There is no significant difference between Mask R-CNN trained on Microsoft COCO and on the augmented dataset for L_0 occlusion. However, clear improvement has been achieved for L_1 and L_2 occlusion which proves the applicability of the idea to model fragmented occlusion by the masks. Nevertheless, all approaches basically do not reach the expected robustness and accuracy for moderate L_2 and heavy L_3 occlusion. One reason for this is that our current technique is not accurate enough to model fragmented occlusion. Furthermore, clear limits exist as heavy fragmented occlusion removes local spatial and structural information necessary for current approaches in object detection.

We further recognise that bounding box labelling is not the appropriate approach for labelling data showing fragmented occlusion. Especially for L_3 and L_4 occlusion, it is frequently impossible to manually define the bounding box. Such occlusion levels allow an approximate localisation of the object in the image but make the observation of the object's extent impossible. While the recall in Figure 5 is still meaningful, the precision is basically undefined. This observation has severe consequences on the labelling,

Figure 4: The problem of multiple detections. Ground truth is shown in green. Left: state-of-the-art yields two bounding boxes of the same, single person. Middle: two persons are visible. Detection yields two bounding boxes which are difficult to associate. Right: an even harder case with three persons.

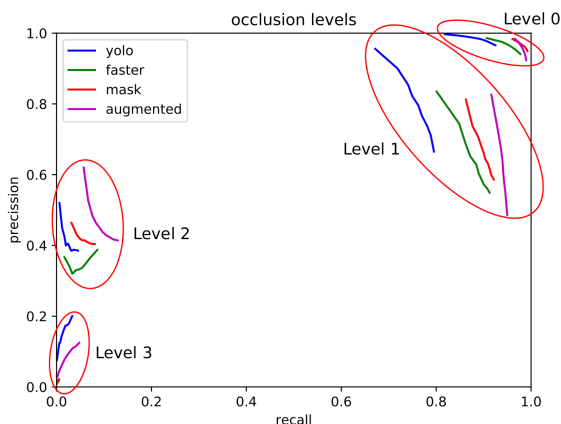


Figure 5: This ROC plot shows results of Faster R-CNN (green), YOLO (blue), Mask R-CNN (red) and our method (purple) for all occlusion levels.

but also on the evaluation and on the detector which we leave open for future research.

5. Conclusion

This paper formulates a new scientific question on object detection with fragmented occlusion which is different to partial occlusion. We show by a study that current object detectors fail in this case. We generated and labelled a new dataset showing people behind trees in a forestry environment. Such scenes frequently occur in border surveillance which has become very important in EU security policies. We try to tackle the occlusion challenge by augmenting Microsoft COCO including the pixel-wise segmentation masks to capture the occlusion problem. We show that Mask R-CNN trained on this data improves on fragmented occlusion, however, we also observe severe loss of spatial, structural information and that the bounding box itself is not the appropriate description to cope with fragmented occlusion. This

has severe implications on the detection approach itself, but also on dataset labelling and evaluation. A potential solution is left open for future work.

Acknowledgments

This research was supported by the European Union H2020 programme under grant agreement FOLDOUT-787021. We thank all our students on internship to label the new dataset.

References

- [1] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 01 1996.
- [2] R. Girshick. Fast r-cnn. In *CVPR*, pages 1440–1448, 2015.
- [3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.
- [4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *CVPR*, pages 2961–2969, 2017.
- [5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.
- [7] G. Nebehay and R. Pflugfelder. Clustering of static-adaptive correspondences for deformable object tracking. In *CVPR*, June 2015.
- [8] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.
- [9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.
- [10] S. Ullman, L. Assif, E. Fetaya, and D. Harari. Atoms of recognition in human and computer vision. *PNAS*, 113(10):2744–2749, 2016.