Exploring the Benefits of Depth Information in Object Pixel Masking (Student Abstract)

Anish Kachinthaya,[1] **Yi Ding,**[2] **Tobias Hollerer**[2]
[1]Research Mentorship Program, University of California, Santa Barbara
[2]Four Eyes Lab, Department of Computer Science, University of California, Santa Barbara
anishk23733@gmail.com

Abstract

In this paper, we look at how depth data can benefit existing object masking methods applied in occluded scenes. Masking the pixel locations of objects within scenes helps computers get a spatial awareness of where objects are within images. The current state-of-the-art algorithm for masking objects in images is Mask R-CNN, which builds on the Faster R-CNN network to mask object pixels rather than just detecting their bounding boxes. This paper examines the weaknesses Mask R-CNN has in masking people when they are occluded in a frame. It then looks at how depth data gathered from an RGB-D sensor can be used. We provide a case study to show how simply applying thresholding methods on the depth information can aid in distinguishing occluded persons. The intention of our research is to examine how features from depth data can benefit object pixel masking methods in an explainable manner, especially in complex scenes with multiple objects.

Introduction

Computer vision algorithms have made significant improvements in detecting and locating multiple people within frames. Uses such as clothing parsing, action recognition, and video annotation all offer benefits for surveillance (Liu et al. 2005), analyzing sports games (Thomas et al. 2017), and many other applications. The state-of-the-art deep learning algorithm for performing pixel masking is Mask R-CNN (He et al. 2017), which simultaneously detects bounding boxes of objects within images and generates masks that are placed on the image in the right location of the bounding boxes.

While Mask R-CNN works accurately when detecting multiple people in an image, it still is prone to fail in cases that involve a person obstructing another or multiple others. Depth data offers potential insight into scenes that regular RGB streams may not provide (Ren, Meng, and Yuan 2011). RGB-D sensors provide two modalities of output: a 2D RGB image and depth information with the distances from the camera for each pixel. This additional dimension of depth can help distinguish between objects at different distances from the sensor, especially in application of detection that requires understanding the edges of people while they may be overlapping in a frame (Fujimura and Nanda 2009).

Accurate detection of occluded objects and people is a long-standing challenge for computer vision. This paper intends to show how depth information can help distinguish between people in an explainable manner through thresholding. We provide a case study of possible Mask R-CNN failure and lay a foundation for how depth information might benefit future algorithms.

Methods

Using Depth Data for Masking

Our algorithm first extracts the bounding boxes of people within images using Mask R-CNN's model pre-trained on MS COCO (He et al. 2017). The depth matrix for the bounding box is extracted and plotted on a histogram. Making the assumption that the pixels of the objects within the bounding box are all at a similar distance from the sensor, each peak represents an object in the bounding box, often two peaks for the object and the background. We then use Otsu's method (Otsu 1979) to find the depth value that separates each of these peaks through testing various thresholds and minimizing the variance between the data separated by the thresholds. The pixels within the bounding box that are lower than a certain percentile under the threshold depth value are separately labeled and masked. We randomly sample 25% of the data for tuning and find that the 85th percentile yields the best masking performance.

Data

101 RGB-D images are taken using the Intel Realsense Depth Camera D415 to have a common set of data for comparison. This dataset included a number of cases where one person was being partially occluded, with multiple people in each scene. The pre-trained Mask R-CNN model is used on the sample data to extract the pixel mask of the people within the image. For comparison with the depth thresholder, we evaluate the intersection of union between the pixels masked by both of these methods to determine how the depth-based thresholding method differs in comparison to Mask R-CNN.

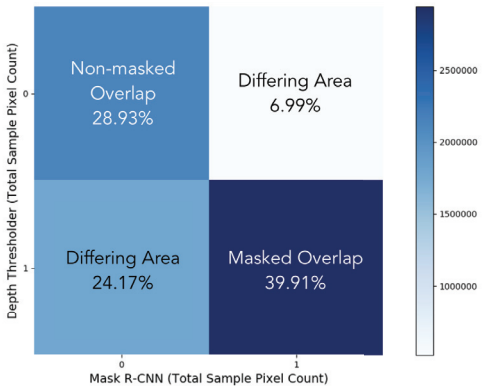

Figure 1: The confusion matrix showing the overlap between the pixels masked by our depth thresholder algorithm and Mask R-CNN. The top left area is pixels not masked by both algorithms, top right is pixels masked by Mask R-CNN but not the depth thresholder, bottom left is pixels masked by the depth thresholder but not Mask R-CNN, and the bottom right is pixels masked by both.

Results and Discussion

Comparing the masking accuracy of our depth thresholding method, the confusion matrix in Figure 1 indicates that our depth-based method overlaps with Mask R-CNN for 68.83% of the pixels in the sample set. However, it also masks many pixels that are not masked by Mask R-CNN, indicated in the lower left value of the confusion matrix: 24.17% of the pixels. Most of this difference occurs in the lower body of people, which is shown in the depth thresholding image in Figure 2, as it has difficulty distinguishing between the ground and the person when they are at a similar distance range. This causes the algorithm to mask many pixels in the lower body that Mask R-CNN does not mask. There are also outliers in depth values at different regions of the person's body, such as the hair, which cause the depth method to not mask pixels where Mask R-CNN does, as shown in the top right value of the confusion matrix. These results show that depth thresholding can mask pixels to a slightly less accurate extent than a deep learning based model, demonstrating that depth information can provide features to improve existing deep learning models.

We also found that Mask R-CNN fails to mask occluded people in certain cases. In Figure 2, it detects the bounding boxes of the three people present in the image but fails to mask the occluded person. With poor lighting, Mask R-CNN can fail to accurately mask people and objects. Depth thresholding, which is independent of lighting conditions, is able to mask the occluded person in the case of this image. This indicates that depth data can provide insights in scenes that regular RGB data may not be able to. This shows that Mask R-CNN and other computer vision algorithms can benefit from the features extracted through depth-thresholding, leading to better and more accurate object and person detection algorithms in their various applications.

Figure 2: An example image where Mask R-CNN fails to mask the person occluded in the background while the depth thresholder does.

Conclusion and Future Work

We have shown a few cases where depth thresholding can benefit object masking performance. By simply using thresholding on depth alone, we can achieve good performance for masking regions of interest. While depth data offers limited information in scene understanding without RGB information, the fusion of this information shows great promise in the task of occluded object masking and understanding. In the future, we plan to explore how to effectively fuse this depth information in an explainable manner, to better account for various situations such as occlusion and to increase the versatility of the depth-based masking algorithm.

Acknowledgements

We would like to thank all members of the Four Eyes Lab, Vrishab Krishna, Kartik Narang, and Venugopal Chillal for aiding in the generation of ideas and the discussion of results.

References

Fujimura, K., and Nanda, H. 2009. Visual tracking using depth data. US Patent 7,590,262.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Liu, X.; Tu, P. H.; Rittscher, J.; Perera, A.; and Krahnstoever, N. 2005. Detecting and counting people in surveillance applications. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, 306–311. IEEE.

Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9(1):62–66.

Ren, Z.; Meng, J.; and Yuan, J. 2011. Depth camera based hand gesture recognition and its applications in human-computer-interaction. In *2011 8th International Conference on Information, Communications & Signal Processing*, 1–5. IEEE.

Thomas, G.; Gade, R.; Moeslund, T. B.; Carr, P.; and Hilton, A. 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding* 159:3–18.