Detecting and Grouping Identical Objects for Region Proposal and Classification

Wim Abbeloos[1][*], Sergio Caccamo[2][*], Esra Ataer-Cansizoglu[3], Yuichi Taguchi[3],
Chen Feng[3], and Teng-Yok Lee[3]

[1]KU Leuven   [2]KTH Royal Institute of Technology   [3]Mitsubishi Electric Research Labs (MERL)

Abstract

Often multiple instances of an object occur in the same scene, for example in a warehouse. Unsupervised multi-instance object discovery algorithms are able to detect and identify such objects. We use such an algorithm to provide object proposals to a convolutional neural network (CNN) based classifier. This results in fewer regions to evaluate, compared to traditional region proposal algorithms. Additionally, it enables using the joint probability of multiple instances of an object, resulting in improved classification accuracy. The proposed technique can also split a single class into multiple sub-classes corresponding to the different object types, enabling hierarchical classification.

1. Introduction

Recent years have seen tremendous progress in object detection and classification based on convolutional neural networks (CNNs). One of the standard pipelines is using an object proposal method and then classifying each region using a CNN [2], as shown in Figure 1 (top). While general object proposal methods work well for general scenes, some scenes such as industrial and warehouse environments have more structure and the objects show less variations in appearance.

In this work, we exploit such structure by using a multi-instance object discovery algorithm [1] that is able to discover, localize, and identify object instances that occur in a scene multiple times. The algorithm uses an RGB-D image as the input and searches for patterns of local features that occur in multiple objects. The output of the algorithm is clusters of the features, each corresponding to one instance of an object. The clusters corresponding to multiple instances of the same object are associated with each other. As shown in Figure 1 (bottom), bounding boxes of the discovered objects are used as the input to a classification network. Our object proposal method also provides the identity of the object, allowing to accumulate the probabilities and further improve the classification accuracy.

------

[*]Denotes joint first authorship
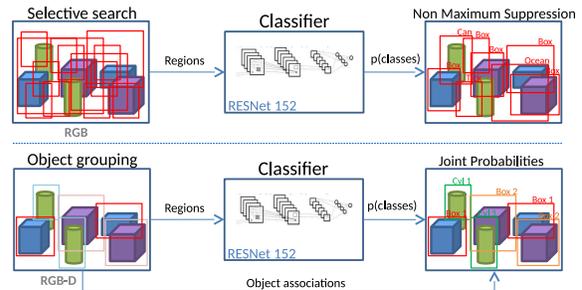
Figure 1. A typical object detection pipeline utilizing a CNN (top), and the proposed pipeline using an alternative object discovery method for region proposal (bottom).

2. Experiments and Results

A dataset consisting of ten scenes with six cereal boxes on a table was captured using an ASUS XTion RGB-D camera. There were three different object instances, two instances of each type (Figure 2). We made sure the same faces of the objects of the same type were visible so that they could be discovered, identified, and localized properly.

For our method the bounding boxes were obtained by expanding each feature cluster by 60 percent to ensure they encapsulate the entire object. The resulting regions were cropped from the images and resized to have 640 pixels on the longest side and a Gaussian filter was applied to reduce interpolation artifacts. These images were then classified using ResNet-152 [3], pre-trained on ImageNet (1000 classes). This resulted in a vector of 1000 class probabilities.

For the baseline method the bounding boxes returned by selective search [4] were resized and classified using the same ResNet-152. An extra non-maximum suppression step was used with a maximum overlap threshold of 0.5. A detection was considered correct if the intersection over union (IOU) was at least either 0.25 or 0.50. The ground truth poses of the boxes were manually annotated.

The sum of the probabilities of the categories "packet" and "comic" was used as the probability of the object being a cereal box. To evaluate the classification results, the

1

Figure 2. Detection and classification results using our method (top) and the baseline method (bottom). For our method the bounding box colors indicate object association. We show both the individual (white) and joint (in color) probabilities. For the baseline method we only show the regions with probability larger than 0.4 for better visualization.

Figure 3. The precision-recall curves for our method with (blue) and without (green) use of the joint probability. The orange and red curves are the baseline method with minimal IOU set to 0.25 and 0.50 respectively.

precision and recall were calculated as

$$ P = \frac{\sum_{i=1}^{C} tp_i}{\sum_{i=1}^{C} (tp_i + fp_i)}, \quad R = \frac{\sum_{i=1}^{C} tp_i}{\sum_{i=1}^{C} (tp_i + fn_i)}, \quad (1) $$

with tp_i, fp_i, and fn_i the true positives, false positives, and false negatives, respectively. These were summed over all $C = 1000$ classes. Note that a single incorrect classification might be counted as a false positive multiple times.

Figure 3 shows the precision-recall curves. For our method we compared the result where we considered the objects as independent (without computing the joint probability), and that where we computed the joint probability of all instances of the same object. The joint probability was computed by multiplying the class probability vectors of these instances and re-normalizing them, which increased the probabilities of correct classifications and improved the precision-recall performance.

Figure 2 visually shows some of the scenes with the detected and classified objects. As can be seen we have a much smaller number of detected regions, while the baseline method tends to generate many false positives. The average number of proposed regions by our method was six, while that of the baseline method was 94. After non-maximum suppression the average number of remaining regions was 15.

3. Discussion

Our object detection framework proposes object regions based on discovering multiple instances of identical, feature rich objects. While these are strong assumptions, several real-world robotic scenarios satisfy these conditions, which would benefit from our method. Currently our evaluation was performed using a limited number of simplified scenes. We plan to extend the evaluation using a larger number of more complex scenes.

References

[1] W. Abbeloos, S. Caccamo, E. Ataer-Cansizoglu, Y. Taguchi, and Y. Domae. 3D object discovery and modeling using single RGB-D images containing multiple object instances. Submitted to IROS, 2017. 1

[2] R. Girshick. Fast R-CNN. In *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, pages 1440–1448, 2015. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[4] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int'l J. Computer Vision*, 104(2):154–171, 2013. 1