RGB-D Object Detection using the Fast R-CNN

Xiaoqiang Yan
CICS, Umass Amherst
xyan@cs.umass.edu

Abstract

In this project, we study the problem of object detection for RGB-D images using the Fast R-CNN. Unlike the RGB modality, there is no powerful ImageNet pre-trained models, which play key role in achieving remarkable detection results. To solve this problem, we study two simple encoding schemes for the depth modality in order to borrow an ImageNet pre-trained model. We further investigate how to combine the two modalities. Experimental results show that using an ImageNet pre-trained model for the depth modality significantly improves the detection results. Furthermore, after combining the RGB and depth information, better performance can be achieved than simply using single of them.

1. Introduction

Object detection from a RGB-D images is an important task, which is potentially useful for manipulation robots. For the RGB images, it can achieve remarkable detection results with R-CNN [4] or Fast R-CNN [6], where ImageNet pre-trained models play crucial roles. However, it lacks such a powerful pre-trained model for the depth modality. In this project, we study different encoding schemes for the depth modality in order to borrow an ImageNet pre-trained model, where a single-channel depth image is converted to a three-channel encoded RGB image. We also investigate how to effectively combine the RGB and depth modalities in the Fast R-CNN framework [2], where they contain complementary information. Experimental results on the NYUD2 dataset [7] demonstrate that using an ImageNet pre-trained model for the depth modality significantly improves the detection results. Furthermore, after combining the RGB and depth information, better performance can be achieved than simply using single of them.

2. Related Work

The R-CNN method [3], designed for object detection in RGB images, has been successfully applied to RGB-D object detection [5]. The RGB images and depth images, cropped from region proposals, are fed into two separate CNNs, where their final features are concatenated and fed into a SVM classifier to get their object class labels and refine their bounding box coordinates. For the depth modality, an encoding scheme is used to convert a single-channel depth image to a three-channel image. By doing so, an ImageNet pre-trained model can be directly applied to the depth modality.

R-CNN [3] has achieved remarkable performance for object detection. However, it is too slow and expensive to train and test. Compared to R-CNN, Fast R-CNN [2] employs a region of interest (RoI) pooling scheme that allows to reuse the computations from the convolutional layers. The speed of Fast R-CNN is much higer than R-CNN, and the detector accuracy also increases for RGB images, so we choose Fast R-CNN [2] to be the architecture for RGB-D object detection. Although Faster R-CNN [8] is also available, the core of this project is to investigate how to utilize the depth modality in addition to the RGB images. Our findings can be directly applied to the Faster R-CNN framework and potentially get better results.

3. Technical Approach

Fine-tuning a pre-trained model allows us to achieve good feature representation even for small dataset, where we don't have enough data to train a complex and strong model. For example, by fine-tuning an ImageNet model, remarable object detection results can be achieved using RGB images [3]. So for depth images, the object detector may benefit from a pre-trained model as well. However, it's almost impossible to have a depth pre-trained model like ImageNet with millions of manually annotated images. So a simple way is to borrow an ImageNet pre-trained model. However, a depth image has only one channel, while an ImageNet pre-trained network accepts only three-channel images. Therefore, we study two ways to encode a depth image, converting it to three channels. We then investigate how to effectively combine RGB and depth modalities.

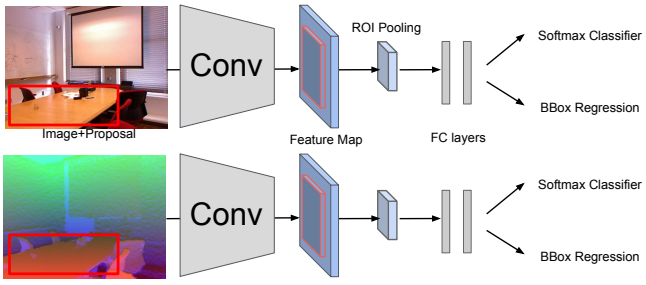

Figure 1. Object detection using the Fast R-CNN with either RGB (**top**) or depth modality (**bottom**, with the HHA encoding scheme).

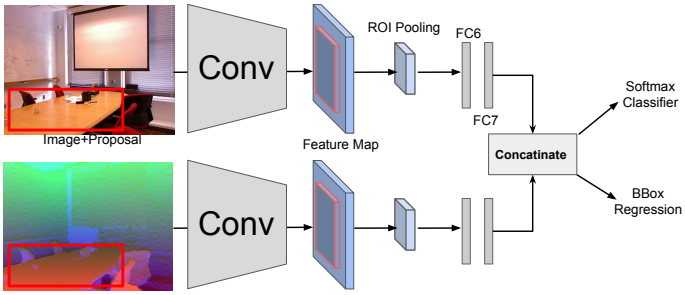

Figure 2. RGB-D object detection using joint training.

3.1. Depth Image Encoding

If we plan to leverage ImageNet pre-trained model to train a new detector for depth images, we should split depth modality into 3 channels which is similar to RGB modality. In our project, we use two approaches.

3.1.1 Channel Duplication

In this way, the depth image will be fed into a CNN to make predictions. To utilize an ImageNet pre-trained VGG16, we duplicate the single-channel depth image three times to match the three-channel filters in VGG16 and scale the input so as to have a similiar distribution with RGB images. This is the same as the method to average the weights across the three RGB channels. We expect this synthetic pre-trained model performs better than learning from scratch.

3.1.2 HHA Encoding [5]

In this approach, the depth image is encoded by three channels, including horizontal disparity, height above ground, and angle the pixel's local surface normal makes with the inferred gravity direction, which is aka HHA. The algorithm of computing the latter two channels is proposed in [5]. Also, every channel is linearly scaled to the 0 to 255 range. The HHA represents internal features in images (depth, surface normal, and height), so we can extract more features from only depth images.

3.2. RGB-D Modality Combination

After fine-tuning both RGB and depth images on VGG16 pre-trained model, we get two object detectors. How to combine these two models to get a better detector for RGB-D images?

3.2.1 Average Final Scores

The first way is training two independent models all the way and then simply get an average of two class scores before SoftMax and use the average score to do classification. Same for the bounding box regression branch. Surprisingly, we get significant performance boost than each single modality with this simple fusion strategy.

3.2.2 Joint Training

Joint training is another way to allow interactions between RGB and depth streams. A concatenation layer is added after the fc7 layers of both RGB and depth modalities. Then we use the concatenated features to do classification and regression. The architecture is shown in Figure 2. In this method, we train a new model for RGB-D images, which extract both color features and depth features. By using two modalities and fusing their features before making the final decision, the detection result is expected to be better than the simple score average scheme.

4. Experiments

We conduct experiment on the NYUD2 dataset [7]. We use the standard split [5], where 795 images are used for training and 654 for testing. What we focus on is the major furniture categories in the dataset, like table, bed, sofa(listed in Table 1). The region proposals are generated by combining two leading approaches from [4] and [1]. We present experimental results using different pre-trained models, different depth features and different combination methods for RGB and depth modalities. Following previous approaches, we report mean average precision (mAP) scores for each category.

All the detection models, initialized with ImageNet pre-trained VGG16, are trained for 20K iterations using a fixed learning rate of 0.001 with the SGD optimizer. It takes about 5 hours on a NVIDIA TitanX GPU for each model.

4.1. Effectiveness of Depth Encoding

One benefit of depth encoding is to use an ImageNet pre-trained model. But is the usage of such a pre-trained model really helpful? We also train a Fast R-CNN model with random initialization, which is trained also for 20K iterations

	A	B	C	D	E	F
	pre-trained models					
	VGG16	VGG16	VGG16	Random	VGG16	VGG16
	RGB	Channel Dup.	HHA	HHA	Average Score	Joint Training
bathtub	25.2	24.4	30.5	23.9	43.1	**48.5**
bed	65.5	68.9	74.3	65.7	78.9	**80.2**
bookshelf	46.9	25.0	29.0	18.8	**49.0**	48.9
box	2.62	0.32	0.94	1.70	**3.20**	2.64
chair	42.3	38.0	38.9	27.0	48.7	**51.0**
counter	44.7	39.2	44.1	25.6	**56.0**	54.0
desk	13.5	6.11	13.3	6.92	**19.1**	18.6
door	26.6	8.73	11.0	9.95	26.1	**27.3**
dresser	38.8	21.5	37.7	15.1	**54.1**	50.9
garbage-bin	36.3	11.5	20.8	13.2	**43.1**	41.8
lamp	32.5	24.7	28.7	11.4	38.7	**40.0**
monitor	43.8	8.92	29.4	12.7	**53.6**	52.1
night-stand	31.8	38.7	34.8	14.0	**48.7**	43.8
pillow	30.5	19.9	36.3	25.8	42.3	**42.9**
sink	39.9	23.1	28.3	17.6	44.3	**46.7**
sofa	44.5	45.6	52.8	40.4	58.7	**61.3**
table	21.6	21.3	23.9	20.9	28.4	**30.9**
television	47.6	6.83	15.3	16.9	43.9	**45.7**
toilet	51.5	31.7	59.4	21.6	69.2	**70.5**
mean	36.1	24.4	32.1	20.5	44.1	**45.1**

Table 1. Detailed results of RGB-D object detection using Fast R-CNN on the testing set of NYUD2 dataset.

but with a fixed learning rate of 0.01 using the SGD optimizer. Table 1 shows that starting from an ImageNet pre-trained model gives significantly better results (mean mAP 32.1 in column C versus 20.5 in column D). The results of fine-tuning based on VGG16 agrees with [6] (mAP of 34.1 with longer training and model distillation). The results imply that fine-tuning from an ImageNet pre-trained VGG16 performs better than scratch.

We can also see that channel duplication encoding scheme performs worse than the HHA encoding, yielding 24.4 mAP (column B). This is reasonable since HHA contains richer information than depth only.

4.2. Effectiveness of RGB-D Modality Combination

Experimental results above show that HHA encoding scheme performs better than channcel duplication, so we train RGB-D detection system combining RGD and HHA modalities. Even with the simple score averaging, we can get significant performance boost, achieving an mAP of 44.1 (column E). By using concatenated features for proposel classification and regression, we can get better results with mAP of 45.1 (column F), which is also consistent with [6] with mAP of 49.1 (with longer training and model distillation). But we found fine-tuning all layers leads to overfitting on the training data due to limited number of training samples. Therefore, we fix first three layers of a VGG16 model.

Figure 3 show the detection results of different detection systems. Generally, by fusing RGB and HHA information, we have less false detections and thus achieve better performance. Sometimes, even on cluttered scenes (*e.g.*, the last third row), our fused detection model can almost detect almost every object instance. But there are also missing object detections (*e.g.*, in the second row), indicating there is still space to further improve the performance of RGB-D object detection.

5. Conclusion

In the project, we investigate two ways to encode depth images in order to borrow an ImageNet pre-trained model for RGB-D object detection. We also study how to fuse the RGB and depth modalities. Experimental results on the NYUD2 dataset show that better detection performance can be achieved by using an ImageNet pre-trained model than training from scratch, validating our motivation of depth encoding. In addition, by fusing the features from RGB and depth modalities and then performing classification and regression, significant performance boost can be achieved.

Figure 3. Detection results on the NYUD2 dataset. Last two rows are failure cases.

References

[1] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1841–1848, 2013. 2

[2] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 1

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic

segmentation. In *CVPR*, pages 580–587, 2014. 1

[4] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 564–571, 2013. 1, 2

[5] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014*, pages 345–360. Springer, 2014. 1, 2

[6] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. *arXiv preprint arXiv:1507.00448*, 2015. 1, 3

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2

[8] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1