Master Seminar Report for
Recent Trends in 3D Computer Vision

Learning with Side Information through Modality Hallucination

Judy Hoffman, Saurabh Gupta, Trevor Darrell. CVPR 2016

Nan Yang
Supervisor: Benjamin Busam
Computer Aided Medical Procedures
Technische Universität München
15.12.2016

Table of Contents

1 Introduction

Object detection is a classical and important task in computer vision. In some cases, as shown in Figure 1, using only RGB images will provide ambiguities. The reason causing this specific case might be that the object in the red box is made up of one leg of a desk and a bag which is like the shape of a chair. In fact, there are already works focus on using different image modalities simultaneously to achieve better performance than using single modality[2,3,4]. Then one might come up with the idea that always using RGB-d image pairs to detect objects. However, even though depth cameras are more common than before, they are still not pervasive. So achieving good performance in object detection with RGB images as input is still worth researching and valuable. In this paper Hoffman *et al.* propose a novel method which take RGB-d images paris at training time but only RGB images at test time for object detection task. By doing this, the convolutional neuron network learns to hallucinate mid-level convolutional features from a RGB image. A hallucination network is used at training time to transfer features which can be commonly extracted from depth images to RGB images. This method outperforms the standard network with RGB only images as input.

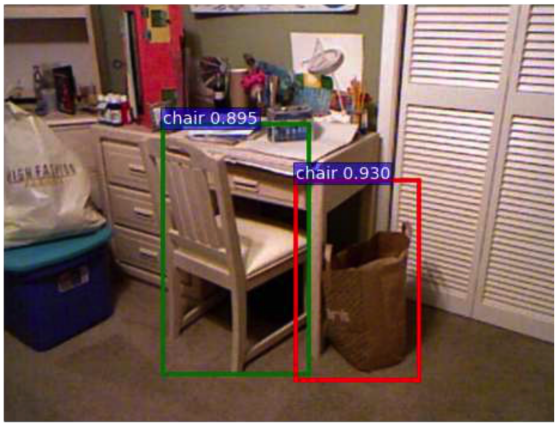


Fig. 1: RGB images could provide ambiguities which lead to a failure: the detection in the red box is the result of RBD object detection.

2 Related Work

The main method of this paper is to use side information, which is depth information here, at training time to transfer information through a new representation, hallucination network, to model of test time.There are 4 main related works of

this paper: RGB-d Detection, Transfer Learning, Learning with side information and Network transfer through distillation.

RGB-d Detection. Object detection aims at 2 purposes; determine where is the objects and to which category each object belongs to. Fast R-CNN[2] is a classical framework for object detection, as shown in Figure 2

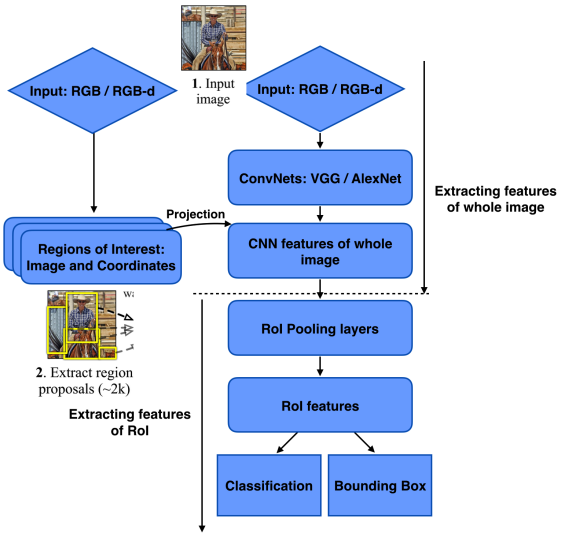


Fig. 2: Fast R-CNN algorithms procedure.

Depth information can provide complementary features of images. This fact has been made use of by previous works taking RGB-d image pairs as input to achieve higher mean average precision rate than RGB only models. Using raw depth information is not efficient enough to learn a precise detector. As a result, many methods introduce new depth representations[5,6,7,8,3]. Recently some methods adding an additional depth convolutional neuron networks are presented [2,1,4]. This work is inspired by these approaches.

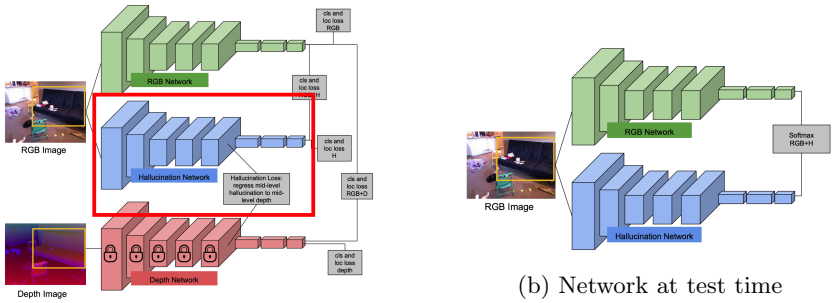
Transfer Learning, Learning with side information and Network transfer through distillation. These three related works are the theoretical foundations of transferring information from one modality to another one through additional network. **Transfer learning** is about sharing information from one task to another one. The information involved is usually, under the situation of deep learning, the weights. So transferring weights learnt from one task to a new task to reduce the training time and improve the performance is the main purpose of transfer learning. In fact, the weights learned from the old task provide a better initialization. Because of very high dimension of the loss function, better initialization means it's closer to the global minima, or at least a better local minima. The method of modality hallucination by learning an additional RGB representation is explored in[9]. **Learning using side infor-**

mation is also a good perspective to be viewed from for this work. The definition of learning using side information is fairly straight forward. It is when a learning algorithm uses additional modality information at training time to achieve better performance, or we can say a stronger model. Surface normals are features could be extracted from depth images. It informs the detector of which directions of surface normals are more common for a specific category objects, e.g. most of normal vectors of a bed are pointed upward instead of downward or leftward, etc, which means that a bed should lay on the ground instead of hanging on the wall. This approach can also be viewed from **Network transfer through distillation**. This method explains the that even though the input modalities of different tasks can also be different, distillation of information from one task to another one can also provide better performance. There are several methods to implement network distillation while in this work, transferring supervision from RGB-d image pairs by introducing joint optimization with multiple loss functions.

3 Algorithm Explanation

This approach uses different network structures, as shown in Figure 3, in training stage and test stage. For training time, there are 3 networks to be learned: RGB network, Hallucination network and Depth network. RBG network takes RGB images as input for the object detection task. Mid-level RGB features are extracted through convolutional neuron networks. Depth network is similar to RGB network. Hallucination network is a special one. It takes RGB images as input but aims to extract depth features through convolutional neuron networks from RGB images. To understand this, we need to know how convolutional neuron networks work as for this hallucination network is just a ConvNets. The weights learned through back propagation guided by a loss function are used for convolution operation on image. The result of convolution operation on images are called feature maps which conclude the whole feature space of the input image, which are the most important data for image localization and classification. So we hope that the feature maps generated by hallucination network is as identical as possible to the ones generated by depth network, but the weights learned are different as for the input modalities of both networks are different.

To train the networks, firstly train the RBG network and depth network independently using Fast R-CNN framework with corresponding image inputs. The next task is to initialize the hallucination network. The architectures of all three networks are identical so either RBG network or depth network could be used to initialize the hallucination network. In this paper, initialization with depth network is used as for experiments show that initialization with depth network can achieve highest mean average precision among initialization with depth network, initialization with RGB network and random initialization. The reason for this could be that it is depth features are supposed to be extracted, so the parameters provided by depth network present a closer start point to the global minima or better local minima than other two initializations. The next stage is



(a) Network at training time

(b) Network at test time

Fig. 3: Network structures

to optimize three networks jointly and transfer depth modality features to hallucination network. The objective guides the training of hallucination network is defined as :

$$L_{hallucinate}(l) = \left\| \sigma(A_l^{dNet}) - \sigma(A_l^{hNet}) \right\|_2^2 \tag{1}$$

where l is the layer number to which we add the hallucination loss and $\sigma(x) = 1/(1+\exp^{-x})$ is the sigmoid function. A_l^{dNet} and A_l^{hNet} are the activation output of layer l in depth network and hallucination net respectively. **Asymmetric transfer** is used to train the hallucination network. While training and updating the weights of hallucination network, the weights of depth network should not be updated. The approach to achieve this is to simply set the learning rate of layers lower than l to 0 which will result in no updates for depth network according to the principle of SGD. Sigmoid function is used inside the Euclidean loss. The author doesn't mention the reason why she does this but some guesses can be proposed. It's always not good for SGD and back propagation to have unbounded loss. Sigmoid function here provided a reasonable bound, which is from -1 to 1, to make the training more robust. And the reason why she doesn't use ReLU function is obvious; ReLU is a not strictly monotonic function.

Whole loss function is defined as:

$$\begin{aligned} L = \gamma L_{hallucinate} \\ + \alpha[L_{loc}^{dNet} + L_{loc}^{rNet} + L_{loc}^{hNet} + L_{loc}^{rdNet} + L_{loc}^{rhNet}] \\ + \beta[L_{cls}^{dNet} + L_{cls}^{rNet} + L_{cls}^{hNet} + L_{cls}^{rdNet} + L_{cls}^{rhNet}] \end{aligned} \tag{2}$$

as shown in Figure 3(a). Weights γ, α and β are used to balance the loss function. Hallucination loss is a regression loss and the hallucination network's input is RGB images but the weights are initialized with depth weights. So at the beginning, hallucination loss is much larger than classification loss and localization loss. However, if hallucination loss dominates the whole loss function, it's not suitable for this **multi-task optimization** process. To balance the loss function, α is set to 0.5 and β is set to 1.0. A heuristic method is used to determine which number should be set to γ; the contribution of the hallucination loss should

be around 10 times the size of the contribution from any of the other losses. Other methods like gradient clipping is also used to improve the robustness of the training procedure.

4 Result

As shown in Figure 4, the novel method proposed in this paper achieves best performance among all the methods on the NYUD2 dataset. The author trains the initial RGB and depth networks using the strategy proposed in[2], but use Fast R-CNN instead of RCNN as used in[2]. And then initialize the hallucination network using the depth parameter values. Finally, we jointly optimize the three channel network structure with a hallucination loss on the pool5 activations. For each architecture choice this work first compares against the corresponding RGB only Fast R-CNN model and find that the hallucination network outperforms this baseline, with 30.5 mAP vs 26.6 mAP for the AlexNet architecture and 34.0 mAP vs 29.9 mAP for the VGG-1024 architecture. Note that for this joint AlexNet method, A-RGB + A-H, averaging the APs of the joint model using each of the AlexNet RGB baseline models.

method	btub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB only [10] (A)	7.5	50.6	**36.8**	1.4	30.2	34.9	10.8	21.5	27.8	16.9	26.0	32.6	20.6	**25.1**	31.6	**36.7**	14.8	25.1	54.6	26.6
RGB ensemble (A-A)	10.5	53.7	33.6	1.6	32.0	34.8	12.2	20.8	34.5	19.6	28.6	45.7	28.5	24.4	31.4	34.7	14.5	**34.0**	**56.1**	29.0
Our Net (A-RGB, A-H)	**13.9**	**56.1**	34.4	**1.9**	**32.9**	**40.5**	**12.9**	**22.6**	**37.4**	**22.0**	**28.9**	**46.2**	**31.9**	22.9	**34.2**	34.2	**19.4**	33.2	53.6	**30.5**
RGB only [10] (V)	15.6	59.4	38.2	1.9	33.8	36.3	12.1	24.5	31.6	18.6	25.5	46.5	30.1	20.6	30.3	40.5	19.5	**37.8**	45.7	29.9
RGB ensemble (A-V)	14.8	60.4	**43.1**	**2.1**	36.4	40.7	13.3	**27.1**	35.5	20.8	29.9	**52.9**	**33.5**	26.2	33.0	44.4	19.9	36.7	50.2	32.7
Our Net (A-RGB, V-H)	**16.8**	**62.3**	41.8	**2.1**	**37.3**	**43.4**	**15.4**	24.4	**39.1**	**22.4**	**30.3**	46.6	30.9	**27.0**	**42.9**	**46.2**	**22.2**	34.1	**60.4**	**34.0**

Fig. 4: Detection (AP%) on NYUD2 test set.

To understand what is learned by hallucination network and the reason why depth features could improve the performance for object detection, Figure 5 shows one of the detection results of hallucination network (green box) and RGB network (red box). The RGB network classifies the painting on the wall as a bed. The reason for this misclassification could be that beds are usually covered by colorful blankets and have square shape. So the texture information provided by this painting on the wall misleads the RGB network. However, hallucination network can extract mid-level depth features as surface normals. So it knows that the square object hanging on the wall has no chance to be a bed.

5 Conclusion

This paper proposed a novel technique for fusing information of different modalities, i.e. depth information and RGB information, through hallucination network taking RGB images as inputs but extracting depth features. This approach outperforms the corresponding Fast R-CNN RGB detection models on the NYUD2 dataset. Future application could be fusing different sensor data of tasks e.g. SLAM to achieve better performance.

Fig. 5: Comparison between results of hallucination network (green box) and RGB network (red box).

6 References

1. S. Gupta, P. A. Arbelaez, R. B. Girshick, and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
2. S. Gupta, R. Girshick, P. Arbelaez, 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.
3. C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision (ECCV)*, 2014.
4. A. Wang, J. Lu, J. Cai, T. Cham, and G. Wang. Large-margin multi-modal deep learning for rgb-d object recognition. In *IEEE Transactions on Multimedia*, 2015.
5. A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3D object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*. 2011.
6. B. soo Kim, S.Xu, and S.Savarese. Accurate localization of 3D objects from RGB-D data using segmentation hypotheses. In *CVPR*, 2013.
7. S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *ACCV*, 2012.
8. E. S. Ye. Object detection in rgb-d indoor scenes. Master?s thesis, EECS Department, University of California, Berkeley, Jan 2013.
9. L.Spinello and K.O.Arras. Leveraging rgb-d data: Adaptive fusion and domain adaptation for object detection. In *ICRA*, 2012.