Towards 3D Object Detection with Bimodal Deep Boltzmann Machines over RGBD Imagery

Wei Liu, Rongrong Ji, Shaozi Li
Dep. of Cognitive Science, School of Info. Science and Eng., Xiamen University, China
Fujian Key Lab for Brain-Like Intelligent Systems

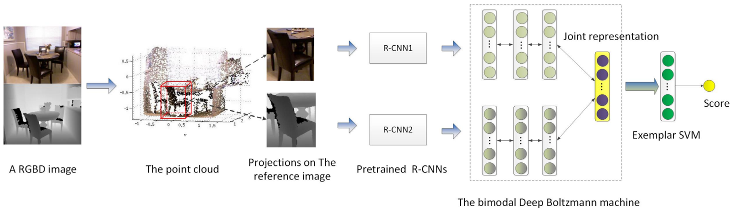

Figure 1: The framework of this paper.

Coming with the popularities of depth sensors like Kinect, nowadays have witnessed an explosive growth of RGB-Depth (RGBD) data to be processed and analyzed, with extensive applications in robotic navigation, pilotless automobile, gaming and entertainments etc. In the core of such applications lies the problem of RGBD scene parsing, i.e., inferring labels of individual verxels to parse their semantic structure.

In this paper, we focus on object detection in RGBD point clouds, which retains as an open problem in the state-of-the-art semantic parsing algorithms of 3D point clouds. 3D object detection in RGBD scenes is a very challenging task due to the deficiency of training data. To the best of our knowledge, the existing labels available for RGBD images are mostly hundreds to thousands, for instance, NYU [5], RMRC [6], and SUN3D [9], which is of scales less comparing to endeavors on the image domain like ImageNet [2], LabelMe [4], and Tiny Images [8]. Therefore, one natural thought is to "transfer" the labels obtained from the 2D domain into the RGBD case to benefit the detector training. However, it is not doable to directly borrow the labels and data structure from RGB and Depth domains, directly and respectively. Can we learn feature representations from both RGB data and Depth modalities to benefit the parsing of RGBD data? Such a cross-modality learning, if not impossible, can open a gate to the feature representation design and detector learning for RGBD. Ngiam et al. [3] have proposed a deep learning based multi-modality learning scheme that has shown to outperform the features learned from single modality. Srivastava and Salkhutdinov [7] proposed a DBM model for learning a generative model of data that consists of multiple and diverse input modalities. The model works by learning a probability over the space of visible units, in which states of latent variables are leveraged as joint representation of multi-modality input.

In this paper, we conquer this challenge by resorting to a feature-level learning crossing both RGB and Depth modalities. To this end, a bimodal deep learning framework is proposed to learn robust detectors in RGBD domain, as shown in the framework in Figure 1. Our innovation is twofold: For the bimodal feature learning, we utilize deep Boltzmann Machine(DBM) to learn features over RGB data and Depth data. For the robust detection, we train Exemplar-SVMs using fused representations of the learned DBM, it ensures the flexibility and generality by training instance-specific metrics and classifiers

Given a RGBD image of a scene, the detection task is to find instances of real-world objects such as *chair* and *table*, which are represented as 3D cuboids. Figure 1 presents an overview of the proposed framework. Our framework takes a RGBD image from Kinect with the gravity direction as input. Most objects are assumed to be aligned on gravity direction so there is only rotation around gravity axis. To support 3D sliding window, the 3D space is divided into cubic cells of size 0.1 meter. For online detection, given a RGBD image, we first generate a point cloud of the scene, based on cameral parameters [6]. Next, we exhaustively slide a 3D bounding box in the point cloud to get scores for all Exemplar-SVMs. Then, the 3D bounding box is projected into 2D bounding boxs on RGB channel and Depth channel, with the reference RGBD image. After that, raw features and the fused representation are sequentially extracted by R-CNNs and the proposed bimodal DBM respectively. Then, the joint representation is used to get scores for all Exemplar-SVMs. Finally, non-maximum suppression is performed on all detection boxes in 3D.

The works that are most similar to our work are [1] and [6]. There are, however, some fundamental differences. First, in this work, we collect both 2D and Computer Graphics(CG) CAD models training data from Internet. Second, we focus on among two diverse modalities: RGB and Depth. Third, to cope with the challenge of deficiency of training data, pretrained R-CNNs are used to extract raw feature from both RGB and Depth channels.

Experimental results on RMRC show that our bimodal DBM is able to learn useful unified representation for the task of object detection with RGBD images.

[1] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Unsupervised feature learning for rgb-d based object recognition. In *Experimental Robotics*, pages 387–402, 2013.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[3] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[4] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*.

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images, 2012.

[6] S. Song and J. Xiao. Sliding shapes for 3d object detection in rgb-d images. In *ECCV*, 2014.

[7] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[8] Antonio Torralba, Robert Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[9] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632. IEEE, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.