3D Object Detection in RGBD Image Using View Independent Feature

Jiang Liu[1], Jianxun Li[2]

1. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240
E-mail: jiangliux@sjtu.edu.cn

2. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240
E-mail: lijx@sjtu.edu.cn

Abstract: These instructions give you basic guidelines for preparing papers for Chinese Control and Decision Conference . Note that "Abstract" and "Key Words" are bold.

Key Words: Paper, Instruction, Chinese Control and Decision Conference

1 INTRODUCTION

3D object detection has received considerable attention recently for the wide range of applications in autonomous driving and personal robotics.[] Traditional detection method only localize objects in 2D bounding boxes, missing the information of 3D location, orientation and 3D extent. In contrast, 3D object detection provide accurate 3D information to understand real world, thus plays an important role in morden computer vision.

It is difficult to solve this task in RGB image due to its natrual ambiguity of missing depth. The most common approach is to discretize the viewing sphere into bins and train a 2D detector for each viewpoint [1]. However, only weak 3D information can be obtained in these methods. Besides, object-centerd methods establishes spatial connections betwee views by mapping them directly to the surface of 3D model. Though these types of models seems attractive for the continuous viewpoint represantations, their detection performance has typically been inferior to 2D deformable models. More recently, [2] extend 2D deformable part-based models(DPMs)[3] to 3D space by means of a deformable 3D cuboid.

With the the availability of inexpensive RGB-D sensors, such as Microsoft Kinect, Apple PrimeSense, Intel RealSense, and Google Project Tango, larger and more ambitious RGBD datasets are created which enabled major breakthroughs for highlevel scene understanding[4, 5]. SUN RGB-D dataset[6], which contains 10,335 RGB-D images with dense annotations in 3D, has become a defacto standard for scene understanding.

It is important to exploit the depth information to advance 3D object detection. Sliding Shapes[7] was proposed to slide a 3D detection window in 3D space, detect objects by matching to CAD models in sliding locations. While CAD models can potentially provide abundant information, the number of models for all categories is limited, and thus this method focused only on a small number of categories.

This work is supported by National Nature Science Foundation under Grant ********.

Ren[8] proposed the Clouds of Oriented Gradients(COG) feature that links the 2D appearance and 3D pose of object categories. COG accurately describes the 3D appearance of objects with complex 3D geometry.

Our contribution are three folds: first, ; second, ; third, ;

2 Related Work

2.1 3D Feature

The histogram of oriented gradient (HOG) descriptor [9] is widely used in object detection. Basiclly, it counts occurrences of gradient orientation in localized portions of an image, and its performance is good enough for 2D detection. However, since gradient orientations are determined by 3D object orientation and perspective projection, HOG descriptors that are naively extracted in 2D image coordinates is not suitable for 3D object description.

To handle this issue, many work has been done. some used extra CAD model to describe the edge from various viewpoints[10].Other assumed that object in sepecific views are near-planar[2]. previous extensions of the HOG descriptor to 3D space require full mesh model[11]. The cloud of oriented gradient(COG) feature, without all these assumption, accurately model the 3D apperence of objects. Like HOG , the calculation of COG can be divided into 3 step: Gradient computation, 3D orientation bin construction and normalization. In the first step, gradient of the image is computed by applying filters [-1,0,1], $[-1,0,1]^{\mathrm{T}}$ to RGB channels. Gradient in position (x, y) is obtained by setting maximum responses across color channels to (dx, dy), with corresponding magnitude $\sqrt{(dx^2 + dy^2)}$. Then, cuboid contains the object is devided it into $6 \times 6 \times 6$ voxels, and 9 orientation bins are constructed for each voxel to model the distribution of local 3D gradient orientation. Perspective projection is used to find corresponding 2D bin boundaries. For each point lies in a voxel, its projected gradient magnitude is sumed into corresponding 2D orientation bin, which is back-projected into 3D bin. Last, bilinearly interpolation is performed between neighboring bins. Histogram $\phi^c_{il}$for voxel l in cuboid i is normalized by setting $\phi^c_{il} \leftarrow \phi^c_{il}/\sqrt{\|\phi^c_{il}\|^2 + \epsilon}$ for a small $\epsilon > 0$, so the

dimension of COG feature is a fixed size of $6^3 \times 9 = 1944$.

2.2 3D Search

Search algorithm define the hypothesis space for potention object location in testing image. Sliding window method is widely used in 2D detection algorithm[9, 3] and the core idea is intuitive. An exhaustive search with predefined step is performed where all the location within the image is scanned to not miss any potential object location. Scale is considered by examined different size of image patch at that location. However, searching all the possible location is computationally expensive. Selective search is proposed for fast computation with high recall[12]. First, initial regions is obtained by performing fast segmentation[13]. Later, the most similar neighbouring region pair is merged into a new region and add to the set in every iteration. Finally, object location boxes are extracted from all regions in the set.

The current search methods in 3D naively generlized from 2D. Sliding shape[7] train an Exemplar-SVM on a CG model rendered at a specific 3D location relative to the virtual camera, then perform exhaust search only at the nearby location of the CG model in order to improve the speed. This 3D local search approach results a relative low recall rate. In [8], 3D space is discritized into grids. The at each possible location, certain size of bounding boxes which based on the empirical statistics of trainning bounding boxes, along with 16 candidate orientation are examined. This strategy brings a high recall rate but introduce many false positive detection.

3D selective search(3D SS) is first proposed in [14]. But it finally outputs candidate bounding boxes in 2D image. [15] rectified this method and output 3D bounding boxes. First, RANSAC is used to fit the pane of 3D point cloud to obtain an initial segmentation. For each plane that its projection covers more than 10% of the total image area, RGB-D UCM segmentation from [16](with threshold 0.2) is used for further splitting. Then starting with this oversegmentation, different segmentation regions are hierarchically grouped, with the following similarity measures:

- $s_{color}(r_i, r_j)$ is the measurment of color similarity between region r_i and r_j using RGB color histogram intersection.
- $s_{\#pixels}(r_i, r_j) = 1 - \frac{\#pixels(r_i)+\#pixels(r_j)}{\#pixels(im)}$, where $\#pixels(\cdot)$ is number of pixels in this region.
- $s_{volume}(r_i, r_j) = 1 - \frac{volume(r_i)+volume(r_j)}{volume(room)}$, where $volume(\cdot)$ is the volume of 3D bounding boxes of the points in this region.
- $s_{fill}(r_i, r_j) = 1 - \frac{volume(r_i)+volume(r_j)}{volume(r_i \cup r_j)}$ describe how well region r_i and r_j fit into each other.

The weighted sum of these four terms form the final output of similarity measuremenmt. Since both 3D and color cues are combined, this very strong baseline achieves an average recall 74.2%.

3 PIPELINE

During trainning(Sec. 3.1), we learn a struct svm for each class, each of which is trained with RGBD image without the extra information of CG model. During testing(Sec. 3.2), we used learned structured SVMs to classify candidate cuboids produced by improved 3D Selective Search method, and output 3D bounding boxes with detection scores. we combine COG with geometry features.

3.1 Trainning

3.2 Testing

4 EVALUATION

5 CONCLUSION

$$\lambda_{1,2} = 0.5 \left[c_{11} + c_{12} \sqrt{(c_{11} + c_{12})^2 + 4c_{12}c_{21}} \right] \quad (1)$$

REFERENCES

[1] C. Gu and X. Ren, "Discriminative mixture-of-templates for viewpoint classification," *Computer Vision–ECCV 2010*, pp. 408–421, 2010.

[2] S. Fidler, S. Dickinson, and R. Urtasun, "3d object detection and viewpoint estimation with a deformable 3d cuboid model," in *Advances in neural information processing systems*, pp. 611–619, 2012.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," *Computer Vision–ECCV 2012*, pp. 746–760, 2012.

[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*, pp. 141–165, Springer, 2013.

[6] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

[7] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images," in *European conference on computer vision*, pp. 634–651, Springer, 2014.

[8] Z. Ren and E. B. Sudderth, "Three-dimensional object detection and layout prediction using clouds of oriented gradients," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1525–1533, 2016.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[10] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3762–3769, 2014.

[11] N. Buch, J. Orwell, and S. A. Velastin, "3d extended histogram of oriented gradients (3dhog) for classification of road users in urban scenes," 2009.

[12] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *In-*

ternational journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.

[13] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[14] A. Kanezaki and T. Harada, "3d selective search for obtaining object candidates," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 82–87, IEEE, 2015.

[15] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 808–816, 2016.

[16] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*, pp. 345–360, Springer, 2014.