GEOMETRIC BASED CONVOLUTIONAL NEURAL NETWORKS FOR EFFECTIVE INDOOR SCENE LAYOUT ESTIMATION

Juntao Feng, Xuejin Chen

University of Science and Technology of China, Hefei Anhui, China.

ABSTRACT

Layout estimation from a single RGB image is a fundamental and indispensable problem for indoor scene understanding, which models the inner space as a 3D cuboid, including floor, ceiling and walls and their boundaries. However, it is significantly challenging to extract layout structure with large clutter and occlusions. In this paper, we propose a geometric based networks for a single RGB image which encodes depth and normal information from image itself. We have demonstrated that using geometric information jointly works better than using only RGB images for indoor scene layout estimation with fully convolutional neural networks(FCNN). Then an optimization framework takes full advantages of spacial labelling results and layout boundary relations from networks to generate final layout estimates. The proposed method has proven to achieve competitive accuracy of layout estimation on two commonly used benchmark datasets.

Index Terms— Scene understanding, layout estimation, geometric embedding.

1. INTRODUCTION

The main purpose of indoor scene layout estimation is to extract semantic boundaries among walls, ceiling and floor, and to obtain different planes which provide strong spacial expression of the scene, for cluttered indoor scenes from a single RGB image, as shown in Fig. 1.

(a) Input image (b) Layout estimation (c) Layout visulization

Fig. 1. Examples of indoor scene images. (a) Input images. (b) Layout estimation with semantic labels. (c) Layout visulization superimposed on the input image.

Thanks to XYZ agency for funding.

Layout estimation is a fundamental and indispensable problem for indoor scene understanding, and it plays a critical role in a diverse range of hot applications, such as scene reconstruction, robot navigation, virtual reality, and so on. Moreover, on the Large-scale Scene Understanding Challenge(LSUN), indoor scene layout estimation has become one of the popular tasks towards scene-centric challenges. However, the most important clues such as room corners and layout boundaries are often occluded by a large amount of clutter occuring everywhere in daily life(see Fig.3 (a)), which make the layout task tough to estimate. Moreover, illumination variations existed in indoor scene(see Fig.3 (b-c)), may increase the difficulty of visual understanding. Last but not the least, the views of indoor scene images are in a wide range(see Fig.3 (d-f)), and such factor can cause appearance diversity for an indoor scene image.

(a) (b) (c)

(d) (e) (f)

Fig. 2. Examples of indoor scene images. (a) Much clutter. (b) Too bright. (c) Too dark. (d-f) Different views.

In recent decades, many researchers have made massive attempts to estimate spacial layout from a single image automatically. One popular framework using a 3D cuboid to approximately express indoor scene layout was introduced by [1] in 2009. The author generated several layout candidates by using vanish point detection. Then it gathered mid-level features like the line membership features and the geometric context features for each candidate and ranked them by using a structured SVM. Unfortunately, the process of layout

candidiate generation is highly sensitive and fragile towards a large amount of clutter. Based on this milestone work, several strategies focused on layout hypothesis generation and ranking were considered. Wang et al.[2] introduced latent variables to model indoor clttuer. Moreover, in [3][4], Schwing et al. modeled cllutterd indoor scenes with higher-order potentials and jointly generated layout and objects with box shapes. Some researchers introduced low-level information as geometric restrictions as supplementary in indoor scene problem. Lee et al.[5]. proposed several physically valid structure hypotheses by geometric reasoning and verified to find the best fitting model to line segments. Ramalingam et al.[6] employed Manhattan Junction grouping to select best layout. As mentioned above, traditional methods need to design image features manually, which greatly increase the complexity and are weekly capable of adapting to dealing with all complex indoor scenes. Besides, there exists a large time consuming for layout candidate generation, which is stongly against our intention to rely on computers to understand the room layout as fastly as possible.

Since the development of the 3D-data capture devices like kinects and RGBD depth cameras, more and more 3D-based methods towards indoor scene field have been studied. Moreover, some researches using 3D-based methods for indoor layout estimation have been studied. Geiger et al.[7] proposed a high-order graphical model and jointly reasoned about the layout, objects and superpixels from RGBD image. Ren et al.[8] proposed a cloud of oriented gradient descriptor and Manhattan voxel that links the 2D appearance and 3D pose of object categories to better capture the 3D room layout geometry and object detection. Guo et al.[9] interpreted layout and 3D model jointly from a RGBD image by aligning to 3D model dataset. 3D information provide additional geometric cues based on planes and objects, which merge big planes and obejects together and weaken the effects of light and complex textures, and thus lead to robust semantic understanding. With 3D-based methods, not only can layout estimation work turn out to be more robust, but also more problems we can solve even 3D modeling problem for whole indoor scenes.

Recently, the deep learning methods and convolution neural network(CNN) have achived impressive progresses in various computer vision tasks, such as semantic segmentation [10][11], object detection [12][13], scene understanding [14][15], and so on. Towards layout estimation problem, several researchers have achieved to adopt deep learning methods to solve it. Mallya et al.[16], presented a Fully Convolutional Networks (FCNs)[10] framwork for learning informative edge maps from a single image, which provided as a new information to sample vanishing lines for layout candidates generation and ranking. This work is the first to train CNN to produce robust features used to replace hand-crafted features towards layout estimation problem. However, in the framwork of layout candidates generation and ranking, their algorithm still remained time-consuming. Dasgupta et al.[17] used the FCN to learn semantic surface labels including left wall, front wall, right wall, ceiling, and ground. Initial layout generation and optimization were all based on suface belif labels. Unfortunately, Their algorithm relies entirely on FCN's precision for semantic surface labels. In other words, if FCN result gains two much error, we may get totally wrong prediction. Moreover, optimization procedure does not utilize any restrictions based on edge information, just taking random combinations of edges instead, which is geometric inconformity compared with RGB images. Ren et al.[18] adopted a multi-task fully convolutional neural network (M-FCN) to jointly predict the room edges and semantic labels. Then a coarse-to-fine method was adopted which enforced several constraints such as layout contour straightness, surface smoothness and geometric constraints to generate fine layout from coarse prediction of room edges. They have considerd geometric constraints to predict high quality estimation results. They need complex geometric computation and rules to generate useful critical lines, however, such low-level features are exactly what we avoid to extract. Zhang et al.[19] used another deconvolution network which has multi-layer deconvolution and a receptive field as large as the entire image compared to FCN. As the result, they can obtain highly reliable edge maps. Then they follow the framwork of layout candidates generation using vanish line samplling with an adaptive line sample strategy for robustness and time reduction. A measurement of similarity is proposed between edge map and layout candidates for ranking. Unexpectedly, their results are less precise compared with [17][18]. One explanation may be that the latter two optimization strategies are more effective compared to vanish line sampling.

In this work, our algorithm is based on framwork of [17], which does not need to extract low features of lines and vanish points from a single RGB image. To enhance the existed FCN's ability, we use networks to estimate depth and normal information from one single RGB image. Then depth and normal information serves as geomtric embedded into FCN networks to jointly generate high quality suface maps and edge maps. Based on[17] optimization framwork, we introduce more geometric constraints from predicted edge maps to optimize surface labels to generate high quality layout estimation. Experimental results demonstrate that our method is robust and effective for layout estimation even facing a high clutter on two popular room layout benchmark datasets.

2. METHODOLOGY

2.1. System Overview

The pipeline is shown in Fig. 3. The coarse layout estmation about semantic layout surfaces for a single RGB image, are predicted by fully convolutional neural network with geometric information emmbeded, including depth and normal information. These geometric information are estimat-

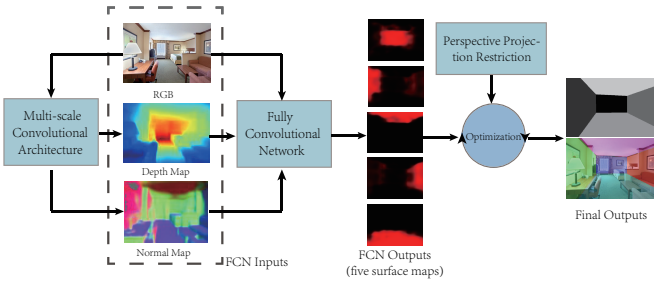

Fig. 3. An overview of our layout estimation algorithm pipeline. First we adopt a multi-scale CNN achitecture to predict geometric information from RGB image, including depth and normals. Then we encode abovementioned information into FCNN , which help to accurately estimate the layout. Optimization framwork based on perspective projection restriction is adopted to generate final precise layout estimates.

ed from source RGB image by a multi-scale convolutional architecture[20]. This will be decribied in Sec. 2.2. Then based on optimization framwork proposed in [17], which mainly uses perspective projection constraints, we can obtain final precise layout estimation results.

2.2. Geometric Based CNN for Coarse Layout Estmation

2.3. Geometric Information

In previous work, [17] [18] achieved to use fully convolutional neural network(FCNN) or multi-task fully convolutional neural network(MFCNN) to predict coarse semantic layout surfaces and layout edges. However, due to much clutter, complex textures and illumination variations existed, semantic surfaces like walls are visually seperated in to pieces, making whole srufaces difficult to aggregate together. Fig. 4 shows layout estimation results using FCNN architecture for prediction. Under the comparison between the predicted results and the ground truth, we can see that due to the environmental effect of indoor scenes, layout results estimated from FCNN are not reliable. When there exits clutter right on the boundaries, such critical clues are partially or entirely excluded, and thus we can not tell location of each plane precisely. Moreover, an entire plane may be predicted separately into pieces due to clutter lay in the plane.

Meanwhile, geometric information such as normals and depth, can serve as clues tending to merge big planes together, with interference factors like clutter, textures and illumination eliminated to varying degrees. These mid-level geometric information can be used to adapt and imrove performance for layout estimation compared to using RGB only. Since we restrict the input to single RGB image for the most general case, we

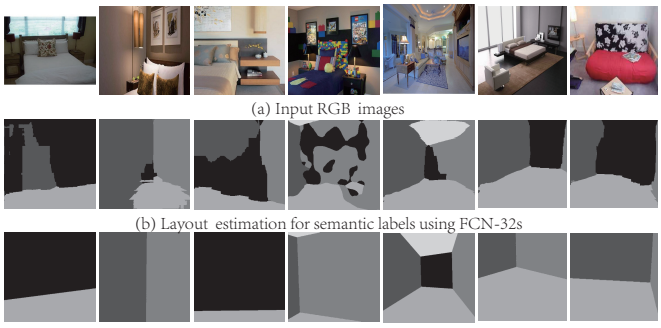

Fig. 4. Layout estimation results using FCNN architecture for prediction in [17] [18].

2.4. Modified Optimization

We adopt a popular model that several researchers[1][17][18] have been used to parameterize indoor layout based on "Manhattan World" assumption. Indoor scene layout can be modeled as

$$L = (l1, l2, l3, l4, v) \quad (1)$$

where l_i stands for i^{th} vanishing line and v stands for the specific vanishing point. The whole scene is equivalent to be labeled to five semantic surfaces, coresponding to (front, left, right, ceiling, ground), as described in Fig. **??**. Based ob Eq. (1), each surface can be reconstructed with vanishing lines, extension lines between vanishing point and Intersection point, and image boundaries. Due to the camera pose, not five surfaces are always visible, and such layout can still be modeled by Eq. (1). Different examples are given in Fig. **??**.

3. RESULTS

4. CONCLUSION

5. REFERENCES

6. REFERENCES

[1] Varsha Hedau, Derek Hoiem, and David Forsyth, "Recovering the spatial layout of cluttered rooms," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1849–1856.

[2] Huayan Wang, Stephen Gould, and Daphne Roller, "Discriminative learning with latent variables for cluttered indoor scene understanding," *Communications of the ACM*, vol. 56, no. 4, pp. 92–99, 2013.

[3] Alexander G Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun, "Efficient structured prediction for 3d indoor scene understanding," in *Computer Vision and*

Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2815–2822.

[4] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun, "Box in the box: Joint 3d layout and object reasoning from single images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 353–360.

[5] David C Lee, Martial Hebert, and Takeo Kanade, "Geometric reasoning for single image structure recovery," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2136–2143.

[6] Srikumar Ramalingam, Jaishanker K Pillai, Arpit Jain, and Yuichi Taguchi, "Manhattan junction catalogue for spatial reasoning of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3065–3072.

[7] Andreas Geiger and Chaohui Wang, "Joint 3d object and layout inference from a single rgb-d image," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 183–195.

[8] Zhile Ren and Erik 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*, 2016, pp. 1525–1533.

[9] Ruiqi Guo, Chuhang Zou, and Derek Hoiem, "Predicting complete 3d models of indoor scenes," *arXiv preprint arXiv:1504.02437*, 2015.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[12] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[14] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[16] Arun Mallya and Svetlana Lazebnik, "Learning informative edge maps for indoor scene layout prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 936–944.

[17] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese, "Delay: Robust spatial layout estimation for cluttered indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 616–624.

[18] Yuzhuo Ren, Chen Chen, Shangwen Li, and C-C Jay Kuo, "A coarse-to-fine indoor layout estimation (cfile) method," *arXiv preprint arXiv:1607.00598*, 2016.

[19] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu, "Learning to predict high-quality edge maps for room layout estimation," *IEEE Transactions on Multimedia*, 2016.

[20] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.