SCIENCE CHINA
Information Sciences

• LETTER •

February 2019, Vol. 62 029102:1–029102:3
https://doi.org/10.1007/s11432-017-9473-5

Effectively modeling piecewise planar urban scenes based on structure priors and CNN

Wei WANG[1*], Wei GAO[2,3] & Zhanyi HU[2,3]

[1]*School of Network Engineering, Zhoukou Normal University, Zhoukou* 466001*, China;*
[2]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 30 November 2017/Revised 15 February 2018/Accepted 14 May 2018/Published online 21 December 2018

Citation Wang W, Gao W, Hu Z Y. Effectively modeling piecewise planar urban scenes based on structure priors and CNN. Sci China Inf Sci, 2019, 62(2): 029102, https://doi.org/10.1007/s11432-017-9473-5

Dear editor,

Piecewise planar stereo methods can approximately reconstruct the complete structures of a scene by overcoming challenging difficulties (e.g., poorly textured regions) that pixel-level stereos appear powerless. In general, these methods have three basic steps: (1) over-segmenting the image into several regions (superpixels) without overlapping; (2) generating candidate planes from initial 3D points; (3) assigning the optimal plane for each superpixel using a global method. However, such methods can be unreliable and inefficient because of three reasons: (1) inaccurate image over-segmentation (generating superpixels based only on low-level image features); (2) incomplete candidate planes (failing to generate complete candidate planes from sparse 3D points); (3) unreliable regularization terms (forcing two neighboring superpixels with similar appearances to be assigned the same plane).

To solve these problems, in this study, a novel plane assignment cost is first constructed by incorporating structure priors and high-level image features obtained by Convolutional Neural Network (CNN). Then, the scene structures are reconstructed in a progressive manner that jointly optimizes image regions and their associated planes, followed by a global plane assignment optimization under a Markov random field (MRF) framework.

Methodology. Given the current image I_r and its left and right neighboring images $\{N_i\}\,(i=1,2)$, we define the following cost of assigning a plane H_s to a superpixel $s \in I_r$.

$$E\left(s,H_s\right)=E_1\left(s,H_s\right)+\gamma\sum_{t\in M(s)}E_2\left(H_s,H_t\right), \quad (1)$$

where $E_1\left(s,H_s\right)$ and $E_2\left(H_s,H_t\right)$ denote the data and regularization terms, respectively, and $M\left(s\right)$ denotes the set of reliable superpixels (i.e., they have been assigned reliable planes); the constant γ is the weight of the regularization term.

In (1), $E_1\left(s,H_s\right)$ is defined by incorporating low-level and high-level image features based on the weight ρ:

$$E_1\left(s,H_s\right)=E_{\text{pho}}\left(s,H_s\right)+\rho E_{\text{cnn}}\left(s,H_s\right). \quad (2)$$

In (2), $E_{\text{pho}}\left(s,H_s\right)$ is defined based on low-level image features and 3D point visibility constraints.

$$E_{\text{pho}}\left(s,H_s\right)=\frac{1}{2\left|s\right|}\sum_{i=1}^{2}\sum_{p\in s}C_s\left(H_s^p,N_i\right), \quad (3)$$

where $|s|$ denotes the total number of pixels in superpixel s and $C_s\left(H_s^p,N_i\right)$ is defined as

$$C_s(H_s^p,N_i)=\begin{cases}L(H_s^p,N_i), & D(H_s(p))=\text{NULL},\\ \lambda_{\text{occ}}, & d(H_s(p))>D(H_s(p)),\\ \lambda_{\text{err}}, & d(H_s(p))\leqslant D(H_s(p)),\end{cases} \quad (4)$$

where $L\left(H_s^p,N_i\right)=\min\left(\left\|I_r\left(p\right)-N_i\left(H_s\left(p\right)\right)\right\|,\delta\right)$ denotes the absolute difference of the normalized

* Corresponding author (email: wangwei@zknu.cn)

color (i.e., the value is between 0 and 1), $H_s(p) \in N_i$ denotes the corresponding point in the image N_i induced by the plane H_s with respect to the pixel $p \in s$, the parameter δ is a truncation threshold to address the robustness concern related to occlusion regions, the constants λ_{occ} and λ_{err} are the occlusion penalty and free-space violation penalty, respectively, and $d(x)$ and $D(x)$ denote the estimated depth value from the current plane and reliable depth value from the initial 3D points, respectively.

Using high-level image features, we first extract three image patches that appropriately contain superpixel $s \in I_r$ and the corresponding projected regions $\{s_i\}(i=1,2)$ in the images $\{N_i\}(i=1,2)$, and resize them to 224×224. Then, we simply consider these three image patches as a 3-channel image and adopt the VGG-M architecture proposed in [1] to extract the features of the 3-channel image. Finally, we directly feed the features to a fully connected linear regression layer and use the output as the plane assignment cost $E_{\text{cnn}}(s, H_s)$ based on the high-level image features.

In this study, the regularization term incorporates the angle prior and is defined as

$$E_2(H_s, H_t) = \begin{cases} C_{\text{sim}}, & H_s = H_t, \\ \mu C_{\text{sim}}, & A(H_s, H_t) \in A_p, \\ \lambda_{\text{dis}}, & \text{otherwise}, \end{cases} \quad (5)$$

where $A(H_s, H_t)$ denotes the intersection angle between plane H_s and H_t corresponding to superpixels s and t, respectively, A_p is the angel prior and set to $[30^{\text{o}}, 45^{\text{o}}, 60^{\text{o}}, 90^{\text{o}}, -60^{\text{o}}, -45^{\text{o}}, -30^{\text{o}}]$. The constants λ_{dis} and μ are the plane discontinuity penalty and the relaxation parameter, respectively.

In (5), C_{sim} measures the color dissimilarity of the superpixels and is defined as $C_{\text{sim}} = 1/\left(1 + e^{-\|c(s)-c(t)\|}\right)$, where $\|c(s) - c(t)\|$ denotes the difference between the mean colors (normalized to a range of 0 to 1) corresponding to superpixels s and t, respectively.

According to the definition of the plane assignment cost, superpixels and their associated planes are jointly optimized through the following steps.

Step 1. Pre-processing. (1) Image I_r is first oversegmented as a set of superpixels (R_0) using any of image oversegmentation methods; (2) Initial candidate planes (H_0) are generated from initial 3D points using the multi-model fitting method [2]; (3) The scene vertical direction and the ground are estimated according to the detected vanish points and the location of the current camera.

Step 2. Jointly optimizing superpixels and their associated planes. (1) For the superpixel s containing initial 3D points, we select plane H_s from H_0 as its reliable plane according to the following condition:

$$T_s = \left(E_1(s, H_s) < \overline{E}\right) \wedge \left(N(P_s, H_s) < \overline{N}\right), \quad (6)$$

where P_s denotes the 3D points that are projected in superpixel s and $N(P_s, H_s)$ denotes the average orthogonal distance between 3D points P_s and plane H_s; $\overline{E}$ and $\overline{N}$ are the average values of the minimal $E_1(s, H_s)$ values and minimal $N(P_s, H_s)$ values of all superpixels containing 3D points, respectively. (2) Let $\mathcal{H}$ and $\mathcal{R}$ denote the current reliable planes and associated superpixels (other superpixels are denoted by $\overline{\mathcal{R}}$). For each superpixel $s \in \overline{\mathcal{R}}$, we first detect the set Π of its neighboring superpixels that have been assigned reliable planes, and then rotate the plane with the axis by the vertical direction and a point at the boundary between superpixels s and $t \in \Pi$. Finally, we consider all the planes with the angle prior A_p as candidate planes of superpixel s. (3) We compute the minimal $E(s, H_s)$ value from these candidate planes, and compare it with $\overline{E}$. If $E(s, H_s) < \overline{E}$, we assign plane H_s (reliable plane) to superpixel s, and save them to $\mathcal{H}$ and $\mathcal{R}$, respectively. Otherwise, we further resegment superpixel s and save the resulting sub-superpixels to $\overline{\mathcal{R}}$. For the superpixels with $E(s, H_s)$ values larger than a predefiend threshold (i.e., $5\overline{E}$), we filter out them as unrelated regions (e.g., sky, ground) and avoid unnecessary plane assignments. (4) To enhance the reliability, we select each superpixel s from $\overline{\mathcal{R}}$ according to the plane assignment priority $\rho_s = N(s) \cdot B(s)$, where $N(s)$ is the number of neighboring superpixels with reliable planes of superpixels s and $B(s)$ is the total length of the boundaries between superpixel s and all superpixels in $N(s)$.

Step 3. Globally optimizing plane assignments: the plane assignments obtained in Step 2 are optimized under the MRF framework [3]. The energy function is defined as

$$E(\mathcal{H}) = \sum_{s \in \mathcal{R}} \left(E_{\text{pho}}(s, H_s) + \omega \sum_{t \in \mathcal{N}(s)} E_2(H_s, H_t) \right), \quad (7)$$

where $\mathcal{R}$ and $\mathcal{H}$ are respectively the set of superpixels and their associated planes obtained in Step 2, $\mathcal{N}$ (s) is the set of all neighboring superpixels of superpixel s, and the constant ω is the weight of the regularization term.

Results and discussion. To evaluate the performance of our method, we conducted experiments

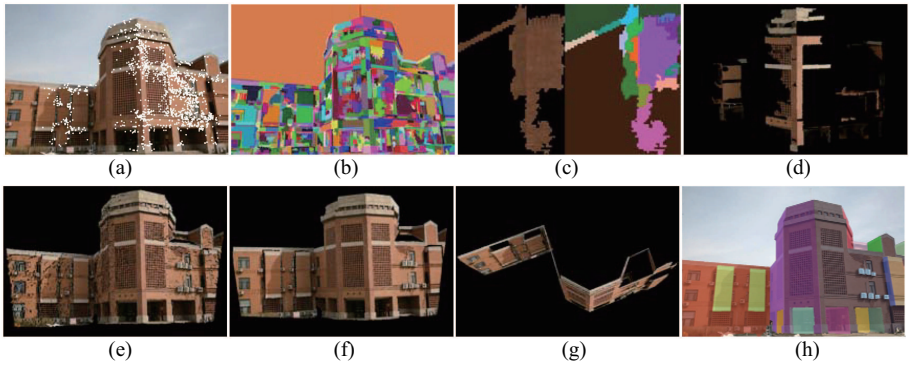

Figure 1 (Color online) Piecewise planar reconstruction. (a) Image I_r and 2D projected points from initial 3D points (white points); (b) initial superpixels; (c) sample of resegmenting superpixels; (d) initial reliable planes; (e) initial plane assignments; (f) final plane assignments; (g) top-view; (h) superpixels corresponding to final plane assignments.

on the life science building dataset[1)] (more experimental results are shown in supporting information). As indicated in Figure 1(b) and (d), at first, only a small number of reliable planes associated with the superpixels could be determined from initial sparse 3D points (Figure 1(a)) because the majority of the superpixels (i.e., inaccurate superpixels) straddled two or more planes and could not be modeled as single planes.

Based on these initial reliable planes, as indicated in Figure 1(e), our method resegmented inaccurate superpixels (Figure 1(c)) according to the plane assignment cost and simultaneously optimized their associated planes. In the meantime, unrelated regions (e.g., sky, ground) were reliably filtered out, which significantly improved the overall efficiency and visualization effects. Clearly, because the initial plane assignments are basically reliable, the global plane assignment optimization could produce superior results as indicated in Figure 1(f) and (g). Figure 1(h) displays the superpixels corresponding to the optimized planes; it can be observed that our method performs well in reconstructing the boundaries between different planes.

Moreover, our method performed quickly because of the guidance of the angle priors. It only needs about 77 s in the sample shown in Figure 1.

Conclusion. Given initial sparse 3D points of a scene, the study constructed an effective plane assignment cost based on scene structure priors and high-level image features obtained by CNN. It then jointly optimized the superpixels and their associated planes, followed by globally optimizing initial scene structures under the MRF framework. Experimental results confirm that our method can effectively and efficiently reconstruct the complete structures of the scene with high accuracy and efficiency.

Acknowledgements This work was supported by National Key Research & Development Program of China (Grant No. 2016YFB0502002), National Natural Science Foundation of China (Grant Nos. 61333015, 61772444, 61472419), Open Project Program of the National Laboratory of Pattern Recognition (Grant No. 201700004), Natural Science Foundation of Henan Province (Grant No. 162300410347), Key Scientific and Technological Project of Henan Province (Grant No. 162102310589), and College Key Research Project of Henan Province (Grant Nos. 17A520018, 17A520019).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of British Machine Vision Conference, 2014
2 Chin T J, Yu J, Suter D. Accelerated hypothesis generation for multistructure data via preference analysis. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 625–638
3 Gorelick L, Boykov Y, Veksler O, et al. Submodularization for binary pairwise energies. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recogition, 2014. 1154–1161

1) http://vision.ia.ac.cn/data/index.html.