Semantic Segmentation of Large-Scale Urban 3D Data with Low Annotation Cost

Jinglu Wang Shiwei Li Jingbo Liu Honghui Zhang Tian Fang
Siyu Zhu Runze Zhang Shengnan Cai and Long Quan
The Hong Kong University of Science and Technology
{jwangae, shiwei, jingbo, honghui,fangtian,szhu, rzhangaj, scaiad, quan}@cse.ust.hk

Abstract

We present a novel method for the semantic segmentation of 3D data in large-scale urban environment. Our method significantly reduces the intensive labeling cost in previous works by automatically generating training data from the input data. The automatic generation of training data begins with the initialization of training data with weak priors in the urban environment, followed by a filtering scheme to remove mislabeled training samples. We formulate the filtering as a binary labeling optimization problem over a conditional random filed that we call object graph, simultaneously integrating spatial smoothness preference and label consistency between 2D and 3D. Toward the final parsing, with the automatically generated training data, a Conditional Radom Field (CRF) based parsing method that integrates the coordination of image appearance and 3D geometry is adopted to perform the parsing of large-scale urban scenes.

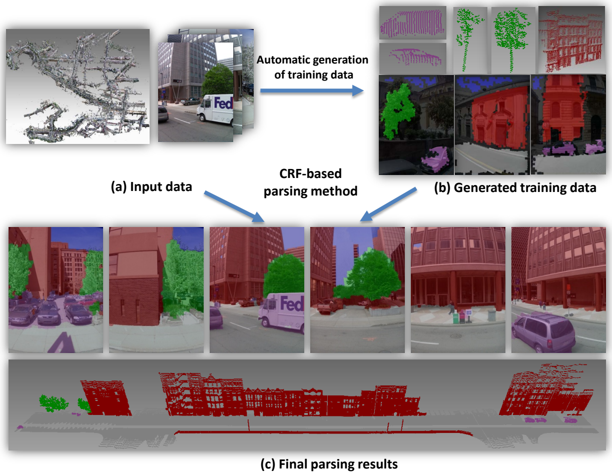
Automatic generation of training data
(a) Input data
CRF-based parsing method
(b) Generated training data
(c) Final parsing results

Figure 1: Overview of the proposed method

1. Introduction

With the maturity and popularity of automatic image capturing devices (*e.g.*, cameras mounted on cars or UAVs), high-resolution urban image data (both street-view and aerial-view) become easier to obtain than ever before. The increasing volume of urban image data has raised the scale of urban 3D reconstruction to an unprecedented city-scale level.

Extracting the semantic information (*i.e.*, parsing) from reconstructed 3D data (either point cloud or triangular mesh) becomes the upcoming goal because of the strong demand for scene understanding, content-based object retrieval or large-scale surveying. Although parsing of urban images has been studied in recent years [4, 14, 15, 7, 6] with encouraging results demonstrated, these methods require a large amount of training data that can account for the vast visual and structural variance of urban environments. Unfortunately, such training data is mostly obtained by tedious and time-consuming manual labeling in the previous approaches, which inevitably becomes an obstacle to applying these traditional parsing methods to the large-scale parsing of urban scenes. Even though there exist some databases with annotations for the urban street scene, such as the CBCL StreetScenes database [2], they are still limited in scale and variance of data sources.

To reduce the cost of manually annotating training data for the parsing of large scale urban scenes, we present a large scale parsing system that can automatically generate training data from input data. Given the coordination of the input images and 3D data, the automatic proposal of training data is achieved by fully utilizing the prevailing knowledge of urban environment. Intuitively, some simple priors can be easily used to distinguish instances of different categories in the urban environment. These priors may not be valid for every instance of the categories, but valid for most of them, and thus we call them *weak priors*. The weak priors are treated as "weak classifiers" and are combined to recognize instances of different categories from the input data. The recognized instances are likely to be misclassified since the weak priors are solely from simple observations. In the next step, a filtering scheme to remove the mislabeled training samples in the initial training data is introduced by formulating it as a binary labeling problem over a CRF. The unary confidence for the initial labeling is estimated by a cross-validation inspired algorithm. The interaction term imposes the geometric spatial smoothness and label consistency which characterizes the correspondences between images and 3D data and is encoded in a carefully designed joint 2D-3D object graph.

Finally, with the automatically generated training data, we use a CRF-based joint 2D-3D method to simultaneosly segment the 3D data clouds and images into five most common categories. In street environment, the categories include *building*, *car*, *tree*, *ground* and *sky* (in images) as the previous works [16, 7] did. In aerial view, the categories include *road*, *building*, *tree*, *grass*, *water* and *others*.

In summary, the contributions of our approach are three folds. 1) The utilization of weak priors in urban images and 3D data automates the generation of training data, significantly reducing the intensive manual labeling in previous works. To our best knowledge, this is the very first exploration of this idea for scene parsing. 2) The novel joint 2D-3D object graph significantly purifies the automatically generated training samples. 3) We demonstrate the potential of fully automatic large-scale parsing of urban scene with comparative performance to that achieved by using manually labeled training data.

Related work For the parsing of urban images, different methods have been proposed [14, 16, 7, 12], which usually formulate the parsing problem with graphical models, such as the CRF. The semantic segmentation of 3D urban data is well studied in the previous works [5, 11]. In [11], the authors introduced a probabilistic, two-stage classification framework for the semantic segmentation of urban maps as provided by a mobile robot, using both appearance information from color images and geometric information from scan data. Without exception, the training data in all these methods is obtained by manually labeling.

2. Our Approach

2.1. Automatic Generation of Training Data

The automatic generation of training data includes two successive steps: 1) labeling initialization in the input data which includes both 3D data and images; 2) filtering of mislabeled training samples. For the labeling initialization, objects of different categories are first segmented in the point cloud or mesh, and recognized with weak priors about these categories in 3D space. Then, we transfer the labeling of the recognized objects from 3D space to image space to initialize the training data for the urban image parsing.

Labeling initialization. The labeling initialization starts with extracting objects of different categories from the 3D data with a hierarchical primitive-fitting algorithm [1]. Objects of different categories are extracted and recognized sequentially.

Based on some weak priors about each category, the object model for the category specifies several discriminative properties for recognizing objects of the category, which includes properties of the following several aspects: (1) *Covered area on the dominant plane*. (2) *Average height above the ground*. (3) *Shape estimation*. We use the ratio of eigenvalues of covariance matrix for all points in an object to distinguish between three basic shapes of objects: line, surface and scatter cloud, similar defined in [10]. (4) *Ratio of points whose dominant normal direction are vertical and horizontal*. This attribute is to estimate the verticalness of objects.

By leveraging pre-defined priors, we can initially identify the category that each object belongs to.

With the initialized labeling of the categories in the 3D data and the registration parameters to images, the labeling initialization for these categories in image space is carried out by transferring the initialized labels of 3D points to image space.

Filtering of mislabeled training samples. As the automatically initialized training samples are generated with only weak priors, some of them are probably mislabeled. To remove the mislabeled training samples, we propose a filtering scheme based on the flexible CRF formulation [13, 9]. The confidence for the initial labeling served as the unary potential in CRF-based formulation is estimated jointly with the appearance information from 2D images and geometric information from 3D data. Furthermore, we integrate the spatial smoothness and label consistency between images and 3D data as well. All these cues are integrated into a CRF model that we call object graph to robustly identify and remove the mislabeled training samples.

The confidence estimation process for each category follows the standard Leave-one-out cross-validation of multiple rounds with random data partitions. In each round of the cross-validation, the initial training data is randomly partitioned into two sets, training set and testing set. A binary classifier is trained by the data from training set. Suppose the binary classifier performs better than random guess, if the initial label of a sample from testing set agrees with that predicted by the trained classifier, then the probability that the initial label is correct is large than 1/2. As the testing in each round of the cross-validation is based on random data partition and thus can be treated as independent testing, the more times a sample's initial label agrees with the predicted label, the more likely its initial label is correct.

In the following, we use $\mathbf{P}(y, k)$ to denote the probability that a sample is classified as a positive sample k times during the N iterations, where $y \in \{-1, +1\}$ denotes the true label of the sample. Suppose the classification accuracy of the trained classifiers during the N iterations is q, then we have:

$$\mathbf{P}(k|y=-1) = C_N^k (1-q)^k q^{N-k} \tag{1}$$

$$\mathbf{P}(k|y=+1) = C_N^k q^k (1-q)^{N-k} \tag{2}$$

For a sample classified as a positive sample k times during the N iterations, the probability that its initial label is correct is:

$$\begin{aligned} \mathbf{P}(y=+1|k) &= \frac{\mathbf{P}(y=+1,k)}{\sum_{y\in\{-1,+1\}} \mathbf{P}(y,k)} \\ &= \frac{q^k(1-q)^{N-k}}{q^k(1-q)^{N-k} + \frac{\mathbf{P}(y=-1)}{\mathbf{P}(y=+1)}(1-q)^k q^{N-k}} \end{aligned} \tag{3}$$

2.2. Associative Hierarchical CRF for Joint Optimization

With the extracted objects in 3D data, we use the associative Hierarchical CRF [8] to formulate the joint segmentation problem of images and 3D data. We define a hierarchical graph $\mathcal{G}' = \langle \mathcal{V}' + \mathcal{T}', \mathcal{E}_{\mathcal{V}'} + \mathcal{E}_{\mathcal{T}'} + \mathcal{E}_{\mathcal{N}'} \rangle$. The global node set $\mathcal{T}'$ denotes the extracted objects in the 3D data, and the nodes in $\mathcal{V}'$ denote the pixels in images. For each extracted object, we build an object graph as a part of $\mathcal{G}'$. For each pixel in images, we add the four neighborhood links, denoted by $\mathcal{E}_{\mathcal{N}'}$. $\mathcal{E}_{\mathcal{V}'}$ denotes the links between nodes in $\mathcal{V}'$ associated with the object graphs, and $\mathcal{E}_{\mathcal{T}'}$ denotes the links between nodes in $\mathcal{V}'$ and $\mathcal{T}'$. The energy function associated with $\mathcal{G}'$ is defined as:

$$\begin{aligned} E'(\mathbf{X},\mathbf{Y}) = \sum_{i\in\mathcal{T}'} \phi_i'(y_i) + \alpha \sum_{(i,j)\in\mathcal{E}_{\mathcal{T}'}} \phi_{ij}'(y_i, x_j) + \\ \beta \sum_{i\in\mathcal{V}'} \varphi_i'(x_i) + \gamma \sum_{(i,j)\in\mathcal{E}_{\mathcal{V}'}+\mathcal{E}_{\mathcal{N}'}} \varphi_{ij}'(x_i, x_j) \end{aligned} \tag{4}$$

This energy function integrates the image appearance information, geometry information and object level information obtained from the scanned point cloud or mesh together. Different cost terms are explained in the following:

Energy cost for 2D images. The random variable $x_i \in \mathbf{x}$ associated with one node in $\mathcal{V}'$ takes a value from the label set. The data term $\varphi_i'(x_i)$ is a unary potential term to encode the probability of a node taking a certain label estimated by the two random forest classifiers R_{2D} and R_{2D3D}, similar to the TextonBoost framework [13]. R_{2D} is trained with patch level appearance features: Texton and SIFT, applied to the pixels without projections of 3D points. R_{2D3D} is trained with the patch level appearance features for R_{2D} combined with the normal, height, and depth of the 3D points that project to the corresponding pixels, applied to the pixels with projections of 3D points.

Energy cost for 3D objects. Random variables $y_i \in Y$ associated with nodes in $\mathcal{T}'$ take a value from the label set $L = \{l_{good}, l_{bad}\}$ that denotes whether the initial label of the extracted object is correct or wrong. For an object with initial label L_i, the data term $\phi_i'(y_i)$ takes the following form:

$$\phi_i'(y_i) = \begin{cases} |c|(1-P_{y_i}) & y_i = l_{good} \\ |c|P_{y_i} & y_i = l_{bad} \end{cases} \tag{5}$$

Here, the estimated confidence P_{y_i} is obtained by testing with the trained binary Random Forest classifiers for the category L_i, the updated object model for the category L_i. $|c|$ is the number of points contained in the object. The smooth term $\phi_{ij}'(y_i, x_j)$ are used to encode label consistency between the 3D point cloud and 2D images, which are defined as:

$$\phi_{ij}'(y_i, x_j) = \begin{cases} (1-2P_{y_i})/0.1|c| & y_i = l_{good}, x_i \neq L_i \\ 0 & y_i = l_{bad} \text{ or } x_i = L_i \end{cases} \tag{6}$$

The weights α, β, γ in (4) can be estimated by the cross validation on a hold-out set, and we solve the optimization problem (4) with the α-expansion algorithm [3].

3. Result and Conclusion

In this paper, we propose a novel method for the parsing of images and 3D data of large-scale urban scenes, which can automatically generate training data from the given input data with weak priors. First, utilizing the weak priors about the most common categories in the urban environment, the initialized training data is generated. Then, a filtering algorithm is proposed to remove those mislabeled training samples in the initialized training data. Finally, with the generated training data, a CRF-based parsing module is proposed for the parsing of large scale street scenes, which uses both the image appearance information and geometry information. Our method are able to apply to large-scale scenes. The results are shown in Figure 2.

References

[1] M. Attene, B. Falcidieno, and M. Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *The Visual Computer*, 22(3):181–193, 2006.

[2] S. Bileschi. *StreetScenes: Towards Scene Understanding in Still Images*. PhD thesis, MIT, 2007.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[4] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. *ECCV*, 2008.

[5] B. Douillard, D. Fox, and F. Ramos. Laser and vision based outdoor object mapping. *Proceedings of the Robotics Science and Systems*, 2008.

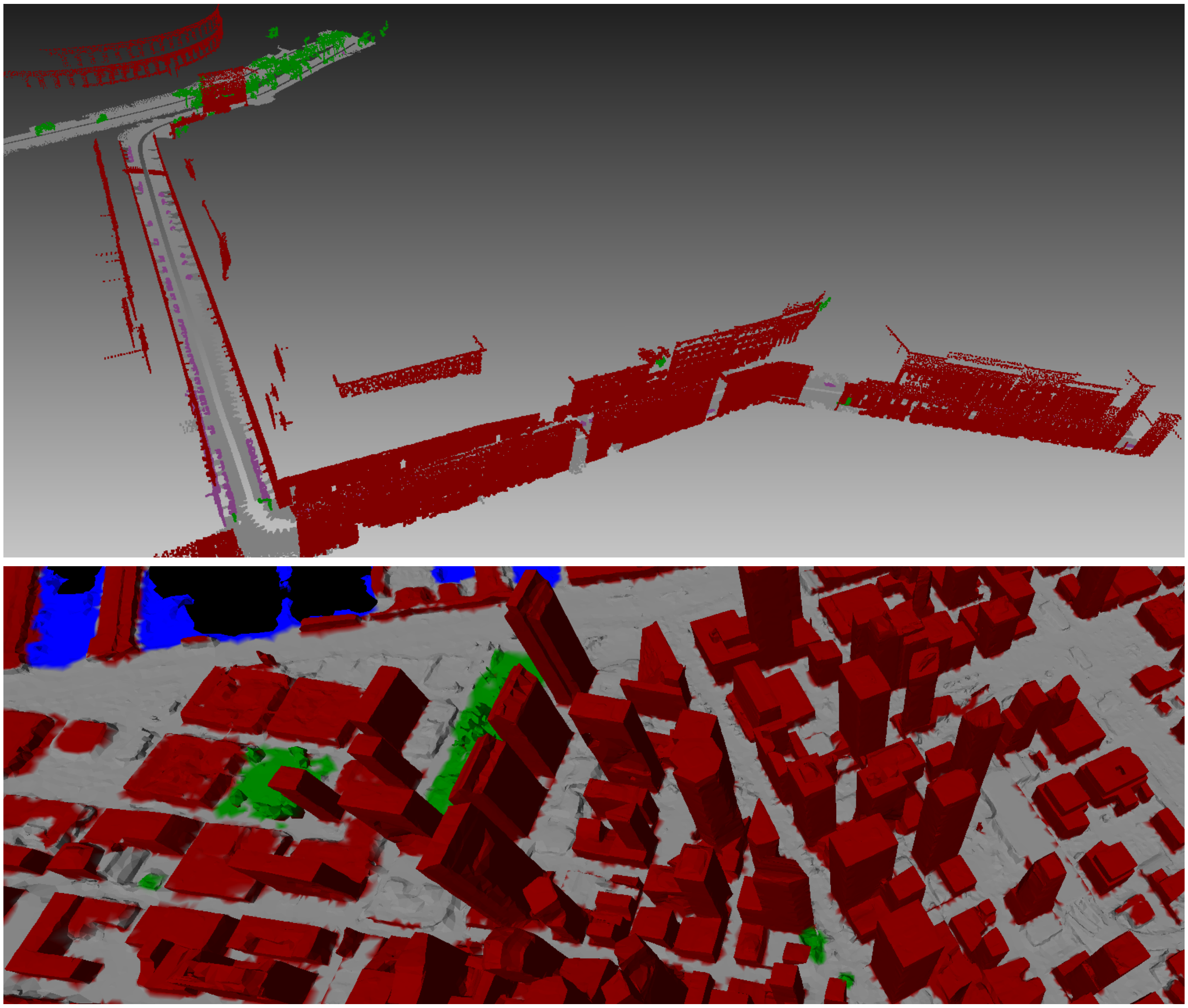

Figure 2: Examples of the parsing results obtained by our method. Red: building; green: tree; gray: road; purple: car; blue: water.

[6] G. Floros and B. Leibe. Joint 2d-3d temporally consistent segmentation of street scenes. *CVPR*, 2012.

[7] Q. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. *CVPR*, 2011.

[8] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. *International Conference on Computer Vision*, 2009.

[9] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. *International Conference on Computer Vision*, 2009.

[10] J.-F. Lalonde, N. V, D. F. Huber, and M. Hebert. Natural terrain classification using three-dimensional ladar data for ground robot mobility. *Journal of Field Robotics*, 23:839–861, 2006.

[11] I. Posner, M. Cummins, and P. Newman. Fast probabilistic labeling of city maps. *Proceedings of Robotics Science and Systems*, 2008.

[12] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr. Urban 3d semantic modelling using stereo vision. *ICRA*, 2013.

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *ECCV*, 2006.

[14] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. *International Conference on Computer Vision*, 2009.

[15] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. *ECCV*, 2010.

[16] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. *ECCV*, 2010.