ATLANTIS PRESS

Advances in Engineering Research, volume 113

International Forum on Mechanical, Control and Automation (IFMCA 2016)

Automatic segmentation of plant point cloud from Multi-view stereo

Jingwei Guo [1,a], Dawei Li[2,b*], Lihong Xu[1,c*]

[1]College of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China

[2]College of Electronics and Information Engineering, Donghua University, Shanghai, 201620, China

[a]1811486604@163.com, [b]daweili@dhu.edu.cn, [c]xulhk@163.com

Keywords: point cloud segmentation, dense CRF, Random Forest classifier, Multi-view stereo reconstruction, Adaptive Normalized Cross-Correlation

Abstract. In this paper, a method for automatic segmentation of plant point cloud is proposed. We get the quasi-dense point cloud of plant from Multi-view stereo reconstruction based on surface expansion. The Adaptive Normalized Cross-Correlation algorithm is used as matching cost to match points of interest in two images, which is robust to radiometric factors and can reduce the fattening effect of boundaries. An efficient segmentation framework is proposed to segment plant from background. After oversegmenting the input point cloud, we extract the 3D feature for each segment and calculate conditional label probabilities using a Random Forest classifier. The out of the classifier is to initialize the unary potentials of a dense CRF whose optimization yields the final labeling. A highly efficient approximate inference algorithm based on mean field approximation is applied to the dense CRF models, in which the pairwise edge potentials are defined by Gaussian kernel. Experimental results show that our segmentation framework based on dense CRF can separate plant from background effectively.

Introduction

In the greenhouse crop cultivation, plant point cloud segmentation from background has significant implication. Plant structural attributes such as height, crown diameter, canopy based height, basal area can be derived. The three-dimensional structure of plants can show the real-time continuous information in the whole process of crop growth by the most direct way, also it is the most intuitive feature for plant phenotype analysis [1]. We get 3D structure information of plants using the Multi-view stereo method because it is relatively cheap, convenient and can obtain complete 3D structure information. The Multi-view stereo reconstruction based on image sequence is used more widely because it can realize automatic calibration of the camera and obtain more complete 3D structure information. Multi-view stereo algorithms can be roughly categorized into three classes [2]. The first class operates by first computing a cost function on a 3D volume, and then extracting a surface from this volume. The second class is image-space method by fusing multiple depth maps. The third class is surface expansion method which builds a quasi-dense point cloud from a set of matches. In this paper, we propose a Multi-view stereo reconstruction method based on surface expansion which iteratively expands a sparse set of initial matches into a quasi-dense point cloud representing the surfaces of the scene. In order to overcome the effects of complex natural light environments and to obtain more accurate plant point cloud, the Adaptive Normalized Cross-Correlation algorithm [3] is used as matching cost to match points of interest in two images, which is robust to radiometric factors such as illumination direction, illuminant color and imaging device changes. The ANCC method also can reduce the fattening effect that object boundaries are not reconstructed correctly, which the Zero-mean Normalized Cross-Correlation (ZNCC) suffers from.

We segment the plant from background after getting point cloud from above Multi-view stereo reconstruction method. The segmentation of 3D point cloud can be roughly categorized into five

* Corresponding authors.

classes [4]: edge based methods, region based methods, attributes based methods, model based growing methods, and graph based methods. Silberman et al. [5] trains a neural network classifier and applies a Conditional Random Field (CRF) model which does not incorporate specific class label relations. Valentin et al. [6] builds a triangulated meshed representation of the scene and classifies the mesh faces with a JointBoost classifier followed by a CRF. Also, their CRF does not take individual label relations into account. Wolf et al. [7] trains a Random Forest classifier and formulate a Markov Random Field, which only performs spatial smoothing on the classification result. Our work has a similar framework compared to Hermans al. [8]. However, we use 3D features instead of 2D features. After oversegmenting the input point cloud, we extract the 3D feature for each segment and calculate conditional label probabilities using a Random Forest classifier. The out of the classifier is to initialize the unary potentials of a dense CRF whose optimization yields the final labeling. The last step smoothes the labeling out to correct ambiguous classification results due to noisy local patch information. A highly efficient approximate inference algorithm based on mean field approximation is applied to the dense CRF models, in which the pairwise edge potentials are defined by Gaussian kernel [9].

Multi-view Stereo Reconstruction

The proposed approach consists of two steps: structure from motion (SFM: from images to sparse points) and dense matching (from sparse points to quasi-dense points). The SFM step is to get 3D sparse feature points location of the scene and camera poses (location and orientation). In the dense matching step, we use the ANCC algorithm as matching cost to match points of interest for a robust and accurate correspondence measure which is robust to radiometric factors and can reduce the fattening effect.

Structure from motion

An overview of the SFM process is presented in Figure 1. The first step is to find feature points in each image. We use the SIFT key point detector [10], because of its invariance to image transformations. Other feature detectors could also potentially be used; several detectors are compared in the work of Mikolajczyk et al. [11]. In addition to the key point locations themselves, SIFT provides a local descriptor for each key point. Next, for each pair of images, we match key point descriptors between the pair, using the approximate nearest neighbors package [12], then robustly estimate a fundamental matrix for the pair using RANSAC. During each RANSAC iteration, we compute a candidate fundamental matrix using the eight-point algorithm [13], followed by Levenberg–Marquardt algorithm. After get the fundamental matrix, we convert it to essential matrix and then get camera poses from essential matrix [13]. At last, we can get the 3D positions of features points through triangulation calculation [13].

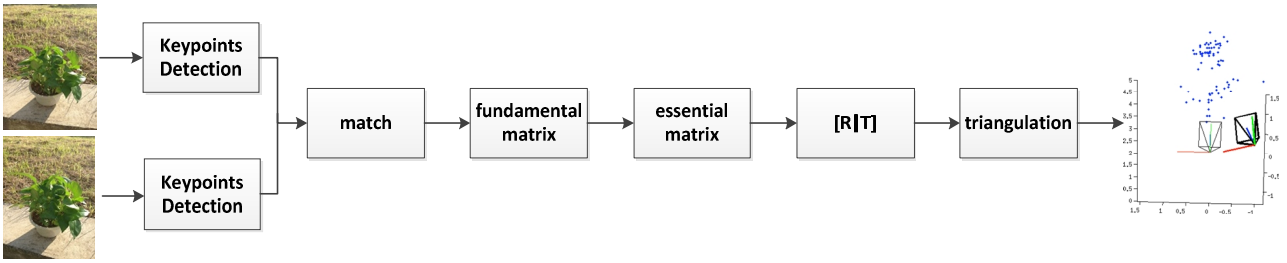


Fig.1 A simplified overview of the SFM process

After get the point cloud of features points between each pair of images, we should merge point clouds because these points aren't in the same coordinate system. Finally, we get 3D position of all the feature points in one coordinate system.

Dense matching

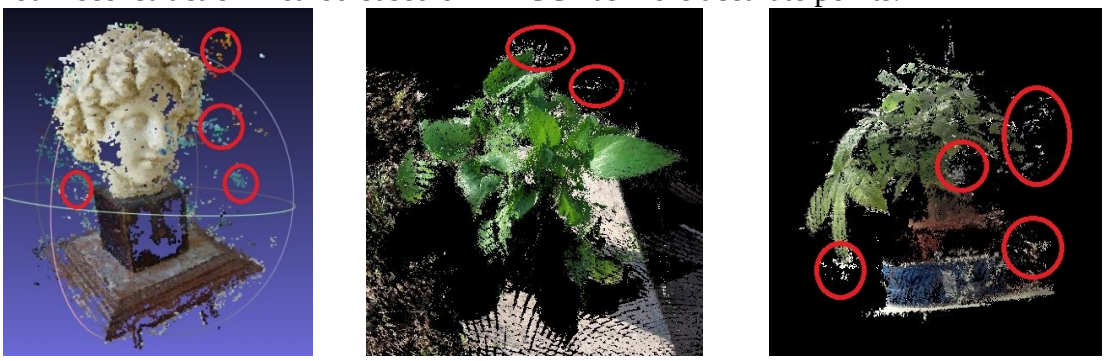
In this stage, we expand a sparse set of initial matches into a quasi-dense point cloud representing the surfaces of the scene based on the best-first expansion strategy by Lhuillier et al. [14], which is robust to initial seed match outliers and efficient in time and space. In order to overcome the effects of complex natural light environments and to obtain more accurate plant point cloud, we use the ANCC algorithm as matching cost to match points of interest instead of ZNCC. The ANCC method is robust to radiometric factors such as illumination direction, illuminant color and imaging device changes and also can reduce the fattening effect, which the Zero-mean Normalized Cross-Correlation (ZNCC) suffers from.

The similarity measure between two pixels in the left and right images is defined by the following equation.

$$ANCC(p,p') = q\sum_{x}\frac{ANCC_x(p,p')}{3} + (1-q)\sum_{k}\frac{ANCC_k(p,p')}{3} \quad (1)$$

where $x \in \{\log Chrom\_R, \log Chrom\_G, \log Chrom\_B\}, k \in \{R,G,B\}$. q is a relative weighting factor between the log-chromaticity color and the original color. $ANCC_x(p,p')$ and $ANCC_k(p,p')$ are the similarity measure in log-chromaticity and original color space [3]. In log-chromaticity color space, the nonlinear relationship that exists between corresponding pixel color values is transformed into a linear one, which can handle the various radiometric changes. The similarity measure of original color is to increased discriminability lowered by log-chromaticity color.

The point clouds of from Multi-view stereo reconstruction in different matching method are illustrated in Figure 2 and 3. From the result we can see that many points aren't reconstructed correctly (marked by red circle) based on the ZNCC matching method because of the fattening effect. And our reconstruction method based on ANCC has more accurate points.



(a) bust (b) bunting (c) money tree

Fig. 2 Multi-view stereo reconstruction based on ZNCC

(a) bust (b) bunting (c) money tree

Fig. 3 Multi-view stereo reconstruction based on ANCC

Point Cloud Segmentation of Plant

To segment the plant from background, first we create an oversegmentation of the scene, clustering it into many small homogeneous patches. Second, we extract the 3D feature for each segment. Third, we calculate conditional label probabilities using a Random Forest classifier, which are used in the final step to initialize the unary potentials of a Conditional Random Field. The CRF model contains two pairwise potentials defined by Gaussian kernel, which can not only remove the noise of the classifier stage but also resolve ambiguous classification results

Oversegmentation

Like the majority of scene segmentation approaches, we calculate an oversegmentation of the input point cloud, which takes color and surface orientation into account to group adjacent points together. We use the supervoxel clustering algorithm proposed by Papon et al. [15], which is publicly available in the PointCloud Library [16]..

Feature Extraction

Based on the work of Wolf et al. [7], we calculate a feature vector x for each of the patches generated by the oversegmentation,, which captures color information as well as geometric properties of the patch. A list of all used features is given in Table 1. $l_0 \leq l_1 \leq l_2$ are the eigenvalues of the scatter matrix of the patch.

Table 1: List of all features calculated for each 3D patch

Feature
Compactness (l_0)
Planarity ($l_1 - l_0$)
Linearity ($l_2 - l_1$)
Angle with ground plane (mean and std. dev.)
Height (top, centroid, and bottom point)
Color in CIELAB space (mean and std. dev.)

Random Forest Classifier

We use a standard Random Forest classifier [17] to get a probabilistic output. RFs have the advantage that they can cope with different types of features without the need for any further preprocessing of the feature vector. We adapt the default training procedure for RFs to our application, such that we end up with a pre-defined number of trees recursively splitting up the data with respect to the evaluation of randomly chosen split functions. Leaf nodes are created at the defined final depth level of the trees or if data cannot be split up any further. These nodes store the distribution of the labels of the training data which has reached the respective node. The conditional probability of label being assigned to a patch with feature vector is then defined as the mean of all label distributions stored in the reached leaf nodes.

Dense Conditional Random Field

After the initial segmentation, there still exist some small isolated regions. Therefore we need to classify these regions to what they belong to. In image processing area, a common approach is to pose this problem as maximum a posteriori (MAP) inference in a conditional random field (CRF) defined over pixels or image patches [18-21]. The CRF potentials incorporate smoothness terms that maximize label agreement between similar pixels, and can integrate more elaborate terms that model contextual relationships between object classes.

In this paper, we use a fully connected CRF that establishes pairwise potentials on all pairs of patches in the point cloud. Consider a random field X defined over a set of variables $\{X_1, \ldots, X_N\}$. The domain of each variable is a set of labels $L = \{l_1, l_2\}$. Consider also a random field I defined over

variables $\{I_1, \ldots, I_N\}$. In our setting, I_j is the feature vector of patch j and X_j is the label assigned to patch j. The corresponding Gibbs energy of the fully connected pairwise CRF model is

$$E(x) = \sum_i y_u(x_i) + \sum_{i<j} y_p(x_i, x_j). \quad (2)$$

where i and j range from 1 to N. A conditional random field (I, X) is characterized by a Gibbs distribution $P(x \mid I) = \frac{\exp(E(x))}{Z(I)}$. The maximum a posteriori (MAP) labeling of the random field is $x^* = \arg\max_{x \in L^N} P(x \mid I)$. The unary potential is computed independently for each patch by the RFs classifier. The pairwise potentials in our model have the form

$$y_p(x_i, x_j) = \sum_{m=1}^{K} m^{(m)} w^{(m)} k^{(m)}(f_i, f_j). \quad (3)$$

where k is a Gaussian kernel $k^{(m)}(f_i, f_j) = \exp(-\frac{1}{2}(f_i - f_j)^T \Lambda^{(m)}(f_i - f_j))$, the vectors f_i and f_j are feature vectors for patch i and j in an arbitrary feature space and μ is a label compatibility function. A simple label compatibility function m is given by the Potts model $m(x_i, x_j) = [x_i \neq x_j]$. For our points labeling problem, we define two kinds of kernel functions. The first one is a smoothness kernel, which is only active in the local neighborhood of each voxel and reduces the classification noise by favoring the assignment of the same label to two close voxels with a similar surface orientation

$$k^{(1)} = \exp(-\frac{|p_i - p_j|}{2q_{p,s}^2} - - \frac{|n_i - n_j|}{2q_n^2}). \quad (4)$$

where p are the 3D patch positions and n are the respective surface normal. The second kernel function is an appearance kernel, which also allows information flow across larger distances between voxels of similar color

$$k^{(2)} = \exp(-\frac{|p_i - p_j|}{2q_{l,s}^2} - - \frac{|c_i - c_j|}{2q_c^2}). \quad (5)$$

where c are the color vectors of the corresponding patches.

Usually the complexity of inference in fully connected models has restricted their application. We use a highly efficient inference algorithm based on a mean field approximation to the CRF distribution [9]. This approximation yields an iterative message passing algorithm for approximate inference. As message passing in the presented model can be performed using Gaussian filtering in feature space, we utilize highly efficient approximations for high-dimensional filtering, which reduce the complexity of message passing from quadratic to linear. The result after inference of fully connected CRF is show in Figure 4. We can see that the plants are successfully segmented from the background.

(a) input point cloud (b) results after CRF

Fig. 4 Results after dense CRF inference

Conclusion

We introduced an efficient segmentation framework for plant point clouds, which combines a Random Forest classifier with a dense Conditional Random Field. We use a highly efficient approximate inference algorithm based on mean field approximation for the dense CRF models, which can not only remove the noise of the classifier stage but also resolve ambiguous classification results. Our method achieves good results for the plants in the greenhouse.

In the future, we will study other point cloud segmentation algorithms to get organs such as leaf, stem and fruit from plant to realize accurate measurement of plant organs.

Acknowledgements

This work was supported in part by the National High-Tech R&D Program of China under Grant 2013AA102305, the National Natural Science Foundation of China under Grant 61573258 and 61374094, China Postdoctoral Science Foundation under Grant 2013M540385, and in part by the U.S. National Science Foundation's BEACON Center for the Study of Evolution in Action, under cooperative agreement DBI-0939454.

References

[1] Sarlikioti V, De Visser P H B, Buck-Sorlin G H, et al. How plant architecture affects light absorption and photosynthesis in tomato: towards an ideotype for plant architecture using a functional–structural plant model[J]. Annals of Botany, 2011: mcr221.

[2] Seitz S M, Curless B, Diebel J, et al. A comparison and evaluation of multi-view stereo reconstruction algorithms[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 1: 519-528.

[3] Heo Y S, Lee K M, Lee S U. Robust stereo matching using adaptive normalized cross-correlation[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 33(4): 807-822.

[4] Nguyen A, Le B. 3d point cloud segmentation: a survey[C]//2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM). IEEE, 2013: 225-230.

[5] Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor[C]//Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011: 601-608.

[6] Valentin J P C, Sengupta S, Warrell J, et al. Mesh based semantic modelling for indoor and outdoor scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2067-2074.

[7] Wolf D, Bajones M, Prankl J, et al. Find my mug: Efficient object search with a mobile robot using semantic segmentation[J]. arXiv preprint arXiv:1404.5765, 2014.

[8] Hermans A, Floros G, Leibe B. Dense 3d semantic mapping of indoor scenes from rgb-d images[C]//2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014: 2631-2638.

[9] Krähenbühl P, Koltun V. Parameter Learning and Convergent Inference for Dense Random Fields[C]//ICML (3). 2013: 513-521.

[10] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[11] Mikolajczyk K, Tuytelaars T, Schmid C, et al. A comparison of affine region detectors[J]. International journal of computer vision, 2005, 65(1-2): 43-72.

[12] Arya S, Mount D M, Netanyahu N S, et al. An optimal algorithm for approximate nearest neighbor searching fixed dimensions[J]. Journal of the ACM (JACM), 1998, 45(6): 891-923.

[13] Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge university press, 2003.

[14] Lhuillier M, Quan L. Match propagation for image-based modeling and rendering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(8): 1140-1146.

[15] Papon J, Abramov A, Schoeler M, et al. Voxel cloud connectivity segmentation-supervoxels for point clouds[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2027-2034.

[16] Rusu R B, Cousins S. 3d is here: Point cloud library (pcl)[C]//Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011: 1-4.

[17] Decision forests for computer vision and medical image analysis[M]. Springer Science & Business Media, 2013.

[18] He X, Zemel R S, Carreira-Perpiñán M Á. Multiscale conditional random fields for image labeling[C]//Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on. IEEE, 2004, 2: II-695-II-702 Vol. 2.

[19] Kohli P, Torr P H S. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82(3): 302-324.

[20] Rabinovich A, Vedaldi A, Galleguillos C, et al. Objects in context[C]//2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-8.

[21] Shotton J, Winn J, Rother C, et al. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context[J]. International Journal of Computer Vision, 2009, 81(1): 2-23.