GrabCutSFM: How 3D Information Improves Unsupervised Object Segmentation

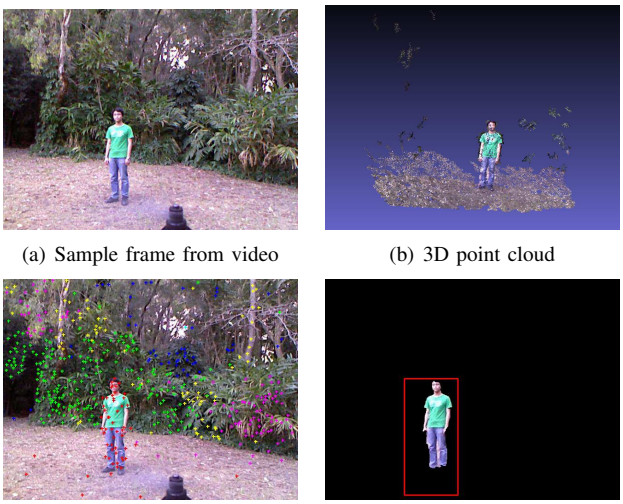
Hu He and Ben Upcroft

***Abstract*— In this paper, we present an unsupervised graph cut based object segmentation method using 3D information provided by Structure from Motion (SFM), called GrabCutSFM. Rather than focusing on the segmentation problem using a trained model or human intervention, our approach aims to achieve meaningful segmentation autonomously with direct application to vision based robotics. Generally, object (foreground) and background have certain discriminative geometric information in 3D space. By exploring the 3D information from multiple views, our proposed method can segment potential objects correctly and automatically compared to conventional unsupervised segmentation using only 2D visual cues. Experiments with real video data collected from indoor and outdoor environments verify the proposed approach.**

I. Introduction

Robust and correct object segmentation is not only useful for object tracking and obstacle avoidance in robotics, but significant for high level computer vision tasks, such as object recognition and image understanding. However, distinguishing an object from background is challenging due to ambiguous visual cues such as brightness, color or texture. Thus, traditional unsupervised image based segmentation (*e.g.*, thresholding, K-means) is prone to obtain either an over or under segmented region which cannot represent object of interest in a meaningful way that is understandable by machine or human. Meaningful objects generally hold a certain physical shape and spatial discrimination to background in 3D space, these kinds of 3D cues could provide prior knowledge to infer meaningful object segmentation. For instance, in an urban environment, the segmentation of objects such as pedestrians or cars can be quite useful for an autonomous platform. It is common that these objects always have different 3D spatial information in contrast to their environment (building, tree or road), while they might still share the similar visual appearance that cause traditional appearance based methods to fail. Therefore this suggests that informative 3D cues could provide a strong hypothesis of meaningful objects in the image and then segment them automatically.

The computer vision community has demonstrated excellent results from still images using either manual initialization from human inputs [1]–[3] or trained models [4], [5]. These methods commonly rely on external supervision to provide the hypothesis of meaningful objects, which might not always be accessible. Even though some work had explored unsupervised segmentation [6], [7], these methods only applied 2D visual cues to enforce the segmented region preserving coherent appearance which is not necessary to be a meaningful object. As addressed above, 3D cues would introduce more descriptive information. Meanwhile, 3D information is ubiquitous to the robotics and computer vision community nowadays due to bloom of cost-effective sensors and advanced algorithms (*e.g.*, RGBD camera, laser or multiple view stereo algorithm). Motivated by this, we introduce 3D cues into the traditional unsupervised object segmentation.



(a) Sample frame from video (b) 3D point cloud

(c) image with the projection of grouped 3D points (d) Segmentation result

Fig. 1. (a)A sample frame from our experimental dataset. (b)3D point cloud generated from image sequences using SFM. (c)Projections of 3D grouped points on image plane using camera pose from SFM. (d)Final segmentation result from our proposed method.

Furthermore, some works [8]–[11] also combine 3D information to achieve segmentation. However, these methods either a require large amount of training data or human intervention to initialize the model of meaningful objects. In contrast, we aim to demonstrate the usefulness of the 3D information for providing meaningful hypotheses of object segmentation automatically in certain scenario (*i.e.*, unsupervised object segmentation).

In this paper, we propose an unsupervised object segmentation method to segment objects of interest in real video data captured both indoors and outdoors (Fig.1(a)), without requiring labeled training data or human intervention. In video sequences, 3D information describing scene and camera poses can be recovered using SFM [12]. Thus, we begin by reconstructing 3D point clouds using video sequence from

The authors are with School of Electrical Engineering and Computer Science, Queensland University of Technology, Gardens Point, Queensland, Australia. {h2.he,ben.upcroft}@qut.edu.au

a monocular camera (Fig.1(b)). Further, we employ K-means to group the point clouds into several clusters using spatial discrimination in 3D space. Each cluster is then projected back to the image plane using a corresponding camera pose estimated from SFM (Fig.1(c)). In addition, a bounding box denoting the hypothesis of a potential meaningful object is generated based on projected clusters on the 2D image plane, and then a state-of-the-art graph based segmentation algorithm, GrabCut [2], is applied to achieve final object segmentation (Fig.1(d)). Unlike the original GrabCut algorithm, the bounding box is initialized automatically by unsupervised learning on reconstructed point clouds in our method. We refer to the proposed method as GrabCutSFM. Fig.1 illustrates an example output of the GrabCutSFM method.

The rest of paper is organized as follows. In Section II, we discuss some related work. Section III addresses the proposed method, GrabCutSFM. Some results and conclusion are given in Section IV and V, respectively.

II. Related Work

Segmentation problems have been studied in computer vision community and other areas for decades. Yet it is still a challenging problem due to large uncertainty and ambiguity between object and background [13]. Lots of algorithms have been proposed to resolve segmentation problem in different scenarios, such as mean-shift [14], normalized cut [6], level sets [15] and graph cut based methods [1].

In this paper, we define the segmentation problem as pixel-wise labelling problem, *i.e.*, label pixels as object or background. In literature, this labelling based segmentation can be summarized as three major categories, *i.e.*, supervised segmentation, semi-supervised segmentation and unsupervised segmentation.

Supervised segmentation is a field of research analogous to classification. In order to segment the object and background, large amount and representative training data are required to achieve a discriminative model. Some works [4], [5], [16] employ 2D cues, such as textures, color, and shape, to train classifiers to discriminate pixels between object and background. Recently, due to the popularity of 3D information in robotics and computer vision community, some researchers [9], [11] combine 3D cues with 2D appearances for urban scene semantic segmentation. Likewise, sufficient training data are required to achieve good segmentation. This paper aims to achieve useful segmentation results without training data.

With respect to supervised segmentation, semi-supervised segmentation focuses on segmentation with some necessary human inputs (*interactive segmentation*) [1]–[3] or robot inputs (*active segmentation*) [17], [18] to achieve object segmentation. These methods will model the potential objects using limited external inputs and then employ graph cut based optimisation procedure to infer segmentation. In particular, GrabCut [2] is one of the advanced methods to achieve good object segmentation. Our proposed method will extend the current GrabCut method to achieve segmentation for video data without human inputs, whereas the original GrabCut is proposed on still image segmentation with human inputs.

Another stream is unsupervised segmentation which only interprets the image as several regions with coherent attributes, *e.g.*, strong contrast on the edge and uniform color on the surface. As pointed in [7], the aim of this unsupervised segmentation is to obtain perceptually important groupings or regions, which often reflect global aspects of the image. Also, Jianbo *et al* [6] address that this unsupervised segmentation is not aiming to segment a complete meaningful object. Thus inherent characteristics of the unsupervised segmentation would limit to some applications, like obstacle avoidance, manipulation, or human interpretation, we might need the potential meaningful object to be segmented instead of just several regions.

Inspired by the reviewed methods, we combine unsupervised method and interactive method to achieve potential meaningful object segmentation automatically. Specifically, we apply unsupervised method on 3D space to acquire some object hypothesises which we then use to initialize 2D interactive segmentation. In contrast to the previous work [19], object hypothesis is automatically generated from 3D information rather than provided by human in [19].

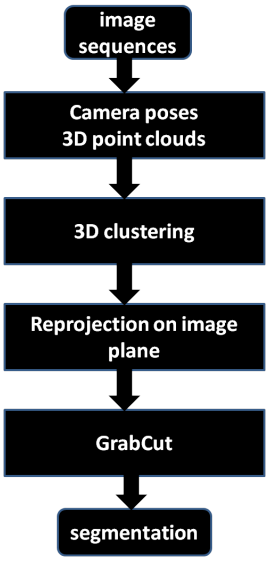


Fig. 2. Overview of the proposed GrabCutSFM method.

III. Method Overview

Figure 2 shows an overview of the proposed method in this paper. It starts from estimating camera poses from multiple views using SFM. Common to all systems that computing point clouds is the requirement for high quality camera pose estimation. We use SFM to acquire camera poses from image sequences using the method proposed in [20]. The reconstructed point clouds are then clustered using K-means based on 3D spatial information and then projected onto the correspondent image through the computed camera poses. These 3D point projections provide possible object

candidatures for GrabCut initialization. Finally, the object is segmented by applying GrabCut. Note that the proposed method currently assumes that the meaningful object is static, which is reasonable for most real environment, either indoor [18] or outdoor environments [9]. More details of this method are described in the following sections.

A. Camera Pose Estimation and Point Cloud Reconstruction

Camera poses and 3D reconstruction from a video sequence has long been an active research topic in computer vision. As the primary focus of this work is to investigate automatic object segmentation using 3D information, we only briefly outline camera pose estimation for each view and 3D point cloud reconstruction of the video sequence. The basic method used in this paper is summarized as follows. Firstly, we calibrate the camera using our modified Bouguet's Calibration Toolbox [21] to get camera focal length, principle point and distortion parameters. Then SIFT features are extracted from each view and tracked over views using epipolar constraints. Finally, camera poses and sparse structures are estimated using camera resection and triangulation, followed by bundle adjustment optimization to refine the solution [20].

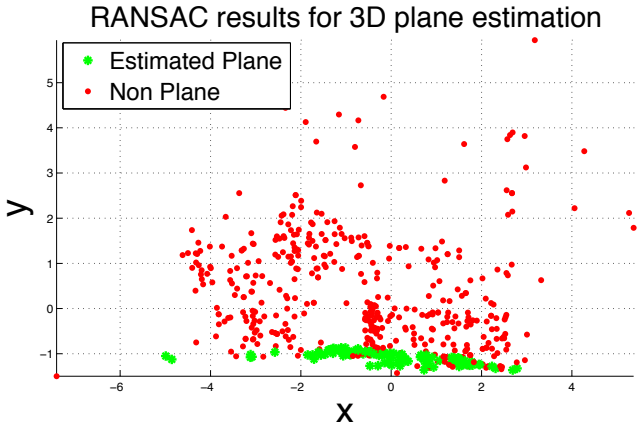


Fig. 3. RANSAC results for 3D plane estimation. Green dots are estimated points belonging to plane while the red ones are outliers (*i.e.*, non-plane). Note that 3D points are visualized in $x-y$ plane.

B. 3D clustering

K-means clustering [23] is a method aiming to partition observations $\mathbf{x}$ with d dimensions into k clusters in which each observation belongs to the cluster with the nearest mean μ. Given the reconstructed 3D point cloud $\mathbf{X}$, K-means aims to minimize the objective function in Eq.1.

$$\underset{\mathbf{C}}{argmin} \sum_{i=1}^{k} \sum_{\mathbf{X}_j \in \mathbf{C}_i} \|\mathbf{X}_j - \mu_i\|^2 \qquad (1)$$

where μ_i is the centroid of 3D points cluster $\mathbf{C}_i$.

Due to the dataset on which our algorithm employs, the points from support surface are also reconstructed. We found a large amount points from the surface which cause K-means to converge to local minimum. Thus points from

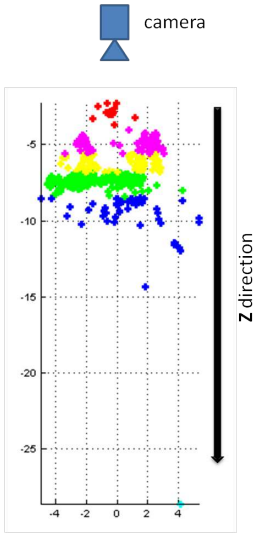


Fig. 4. Clusters of reconstructed 3D point cloud after rejecting points belonging to a plane. Different color represents potential object.

potential object could be clustered into the incorrect centroid. However, points from a surface (*e.g.*, ground or table) share the similar normal direction complied with plane constraints. RANSAC [24] was employed to detect the points belonging to a plane and then remove them from reconstructed point cloud. Additionally, we found our method is not sensitive to k by varying k from 5 to 15. With further investigation, we notice that the k is affected by the relative location between objects and background. In this paper, we set k as 6. Furthermore, we noticed that 3D information along z-axis (camera viewing direction) is more discriminative comparing to x-axis and y-axis in the reconstructed point cloud, which implies that depth information is a strong cue in 3D space. Fig.3 shows the detected points from plane (in green) and Fig.4 illustrates that clusters after rejecting points of plane.

C. Automatic initialization

In order to employ GrabCut framework for object segmentation, an initial object hypothesis is required to model both object and background attributes. Instead of conducting initialization manually, 3D clusters will be projected back to image plane using the corresponding camera pose to generate possible object hypothesis. Eq.2 describes the relationship between 3D points $\mathbf{X}$ and projections $\mathbf{x}$ on image plane.

$$\mathbf{x} = \mathbf{P}\mathbf{X} \qquad (2)$$

where $\mathbf{P}$ is camera matrix which encapsulates camera intrinsic parameters as well as camera rotation and translation information.

Fig.1(c) shows the projections of clustered 3D points on image plane. Due to inaccurate camera poses, we might obtain some incorrect projections on image plane. Since potential object of interest always occupy a certain region in 3D world rather than spread out the whole background, we apply RANSAC again on projected 2D points to remove

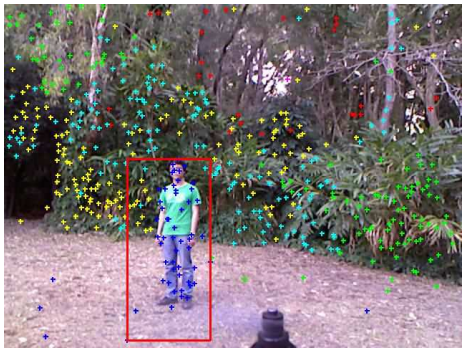


Fig. 5. 2D projections of clustered 3D points on the image plane. Color is associated with corresponding 3D cluster. Bounding box in red is estimated based on projected points. (This figure is best viewed in color).

some outliers which have large variance. Finally we can compute the bounding box which contains the inliers (*i.e.*, potential object) from RANSAC, which is shown in Fig.5. In this paper, the bounding box is simply computed based on the maximum and minimum coordinates of inliers in the image space.

D. Segmentation Model

In this paper, we mainly consider the scenario where the objects of interest are closed to camera compared to the background. In addition, we assume there is only one object of interest in the field of view or multiple objects closed to each other with discriminative distance to the background.

Given the clusters of point clouds learnt by K-means, we project all these clusters onto image using camera projection matrix. Bounding box of the potential object is obtained by the projections of the cluster whose centre is closest to the camera.

GrabCut models object and background using Gaussian Mixture Models (GMMs) learnt from pixels inside and outside of estimated bounding box. Then graph-cut algorithm will be applied to infer segmentation results. More details can be found in [2].

IV. Experimental Results

A. Experiment setup

For the sake of simplicity and cost, we collect data using a monocular camera. In this paper, we present experiments on three video sequences (one for outdoor and two for indoor). Specifically, the outdoor dataset was collected by a moving camera mounted on a quadrotor platform flying around a stationary person, which contains 410 frames at a 640×480 resolution. The two indoor datasets include a robot (NAO) and a box where both were taken from a hand held camera and contain around 200 frames with 640×480 resolution. Sample frames of the data are shown in Fig.6(a) and 7(a). The goal of the proposed method, GrabCutSFM, is to demonstrate correct and automatic segmentation of the potential meaningful object (*i.e.*, person, robot an box) in the image.

In terms of processing speed, we process prerecorded videos with a 3.2GHz i5 Core CPU and achieve segmentation at 3$\sim$4 frames per second. Since frames are segmented independently after 3D clustering, therefore parallelization could be employed to achieve close to real time processing using GPU.

B. Segmentation results

To validate the proposed unsupervised segmentation method, we conducted two kinds of comparisons to demonstrate that GrabCutSFM is not only promising to achieve correct and meaningful object segmentation without any human inputs, but still achieve comparable segmentation results with respect to the interactive segmentation method, GrabCut.

In order to illustrate how GrabCutSFM outperforms the conventional 2D unsupervised image segmentation, we employed K-means on image space and clustered the pixels into two region (*i.e.*, object and background) using color information only. Qualitative comparisons are shown in Fig.6 and Fig.7, respectively. The first column is the sample of our dataset, the second column shows the projections of clustered 3D points on image plane as well as estimated bounding box to cover potential interesting object. The third column shows the segmentation results from 2D unsupervised segmentation, whereas the last column illustrates the segmentation results generated by our method.

It was shown that conventional 2D unsupervised segmentation suffered providing the actual meaningful object segmentation, which might not be useful for object based applications, such as manipulation, tracking and obstacle avoidance. Whereas our proposed method of using 3D cues can provide useful and correct object segmentation automatically, which can be served as the pre-process step of many high level applications, like object classification or recognition.

Fig.8 shows that our automatically generated bounding box is quite close to the one chosen by human, thus GrabCutSFM and GrabCut provide near identical segmentation. However, through the entire video, insufficient projections of 3D points on the boundary of the object due to self-occlusion would cause the estimated bounding box to be smaller than the actual object size, therefore occasionally missing tiny regions near the object boundary.

For a quantitative comparison, we manually segmented the person from the scene for every tenth frame (resulting in 41 ground truth frames) and computed the $F1$ score *w.r.t* ground truth. The $F1$ score is defined as:

$$F1 = \frac{2Precison \times Recall}{Precison + Recall} \tag{3}$$

where *Precison* is the fraction of our segmentation overlapping with the ground truth and *Recall* is the fraction of the ground truth overlapping with our segmentation.

Fig.9 shows that $F1$ score is about 0.87 for our GrabCutSFM which significantly outperforms conventional unsupervised segmentation whose $F1$ score is less than 0.1,

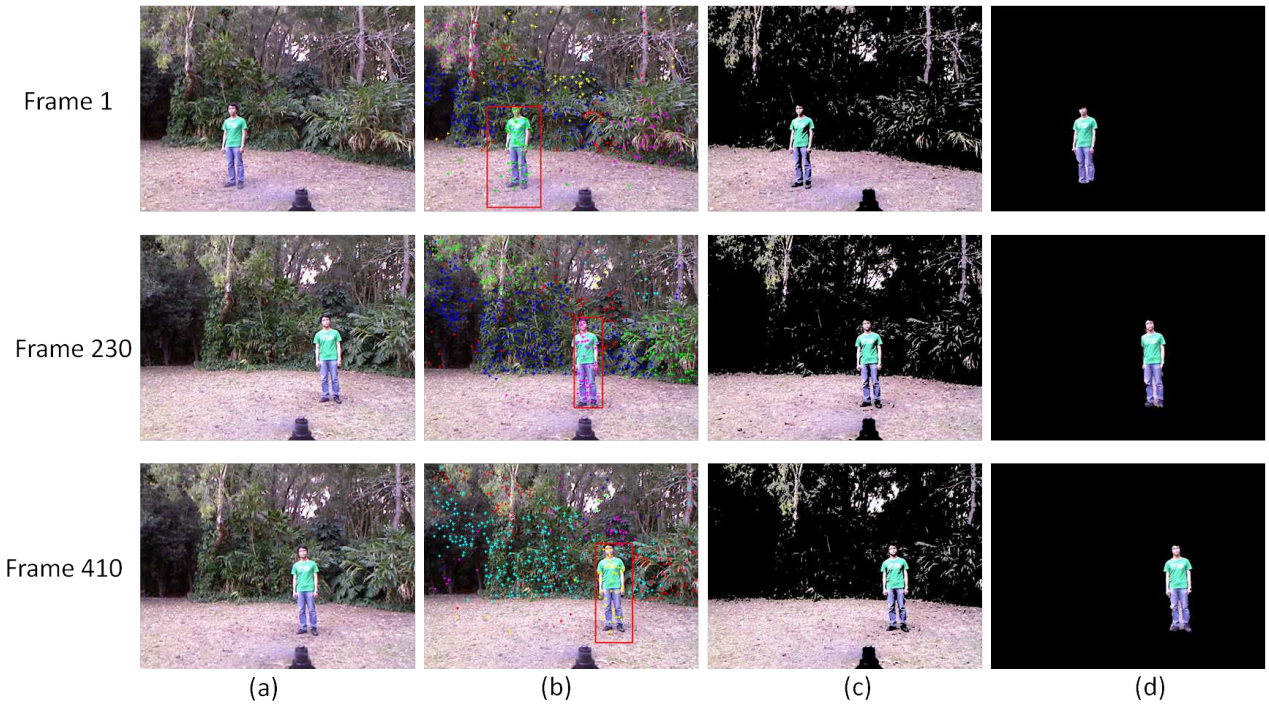


Fig. 6. Sample frames and corresponding segmentation results for outdoor data. (a) Sample frames from our video data; (b) Projections from clustered 3D points on image coordinates as well as estimated bounding box in red; (c) Segmentation using traditional unsupervised method (K-means); (d) Segmentation using GrabCutSFM. (This figure is best viewed in color).

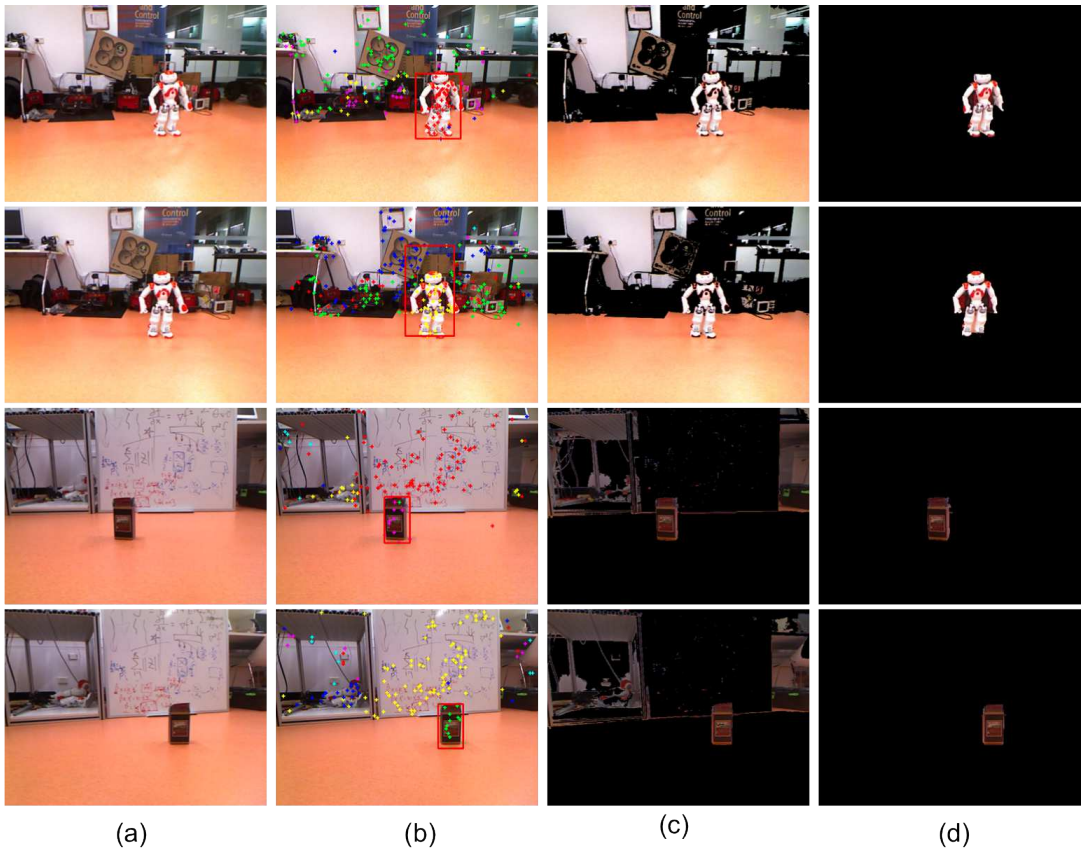


Fig. 7. Sample frames and corresponding segmentation results for indoor dataset. (a) Sample frames from our video data; (b) Projections from clustered 3D points on image coordinates as well as estimated bounding box in red; (c) Segmentation using traditional unsupervised method (K-means); (d) Segmentation using GrabCutSFM. (This figure is best viewed in color).

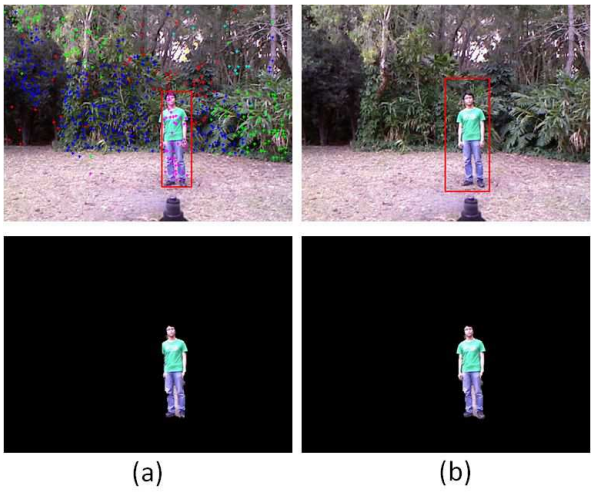


Fig. 8. Comparison between GrabCut and proposed GrabCutSFM. (a) Segmentation with estimated bounding box automatically; (b) GrabCut segmentation with manually provided bounding box. (This figure is best viewed in color).

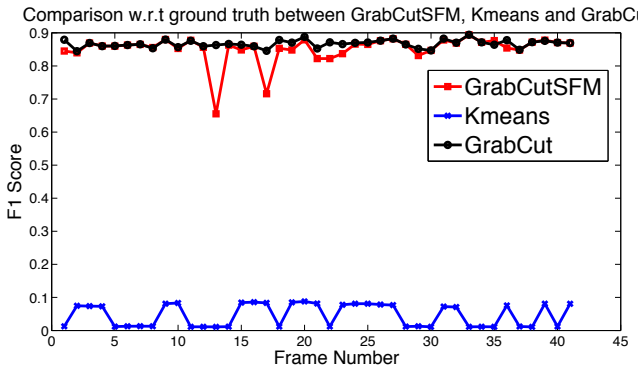


Fig. 9. Quantitative comparison of segmentation using K-means, GrabCut and GrabCutSFM.

meanwhile trivial difference on segmentation results from GrabCutSFM and GrabCut encourages extending the state-of-the-art interactive segmentation to 3D space with fully automatic initialization.

V. Conclusion and Future Work

This paper presents an unsupervised object segmentation method, GrabCutSFM, using 3D cues to obtain meaningful segmentation automatically. This method does not require training data or human intervention, *i.e.*, creating a solution for fully automatic unsupervised segmentation. We evaluated our method on real video data qualitatively and quantitatively. For future work, we would like to extend the method to more complicated environment and incorporate with high level robotic applications, such as object detection and recognition.

Acknowledgment

The authors would like to thank Inkyu Sa for his help in data capture. Thanks to Alex Bewley, Timothy Morris and Michael Warren for proofreading this paper.

References

[1] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in ND images," in *International Conference on Computer Vision*, vol. 1. Citeseer, 2001, pp. 105–112.
[2] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH 2004 Papers*. ACM, 2004, p. 314.
[3] Y. Li, J. Sun, C. Tang, and H. Shum, "Lazy snapping," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 303–308, 2004.
[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," *Computer Vision–ECCV 2006*, pp. 1–15, 2006.
[5] P. Yin, A. Criminisi, J. Winn, and I. Essa, "Bilayer segmentation of webcam videos using tree-based classifiers," *IEEE transactions on pattern analysis and machine intelligence*, pp. 30–42, 2010.
[6] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888 –905, aug 2000.
[7] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.
[8] M. Johnson-Roberson, J. Bohg, M. Björkman, and D. Kragic, "Attention based active 3d point cloud segmentation," in *IROS 2010*, 2010.
[9] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.
[10] H. He, D. McKinnon, M. Warren, and B. Upcroft, "Graphcut-based interactive segmentation using colour and depth cues," *Australasian Conference on Robotics and Automation*, 2010.
[11] J. Xiao and L. Quan, "Multiple view semantic segmentation for street view images," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 686–693.
[12] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge Univ Press, 2000, vol. 2.
[13] A. Jain and R. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
[14] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.
[15] L. Vese and T. Chan, "A multiphase level set framework for image segmentation using the mumford and shah model," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.
[16] J. Sun, W. Zhang, X. Tang, and H. Shum, "Background cut," *Computer Vision–ECCV 2006*, pp. 628–641, 2006.
[17] N. Bergstrom, M. Bjorkman, and D. Kragic, "Generating object hypotheses in natural scenes through human-robot interaction," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 827–833.
[18] M. Bjorkman and D. Kragic, "Active 3d scene segmentation and detection of unknown objects," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 2010, pp. 3114–3120.
[19] H. He, D. N. McKinnon, and B. Upcroft, "Towards automatic object segmentation with sequential multiple views," *ACRA 2011 Proceedings*, pp. 1–7, 2011.
[20] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 835–846.
[21] J. Bouguet, "Camera calibration toolbox for matlab," 2004.
[22] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 8, pp. 1362–1376, 2010.
[23] J. Hartigan and M. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
[24] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[25] S. Geman and D. Geman, "Stochastic relaxation,Gibbs distributions, and the Bayesian restoration of images," *IEEE trans on pattern analysis and machine inteligence*, 1984.