Towards Automatic Object Segmentation with Sequential Multiple Views

Hu He **David Mckinnon** **Ben Upcroft**
Queensland University of Technology
Brisbane, Australia
hu.he@student.qut.edu.au, {d.mckinnon,ben.upcroft}@qut.edu.au

Abstract

Object segmentation is one of the fundamental steps for a number of robotic applications such as manipulation, object detection, and obstacle avoidance. This paper proposes a visual method for incorporating colour and depth information from sequential multiview stereo images to segment objects of interest from complex and cluttered environments. Rather than segmenting objects using information from a single frame in the sequence, we incorporate information from neighbouring views to increase the reliability of the information and improve the overall segmentation result. Specifically, dense depth information of a scene is computed using multiple view stereo. Depths from neighbouring views are reprojected into the reference frame to be segmented compensating for imperfect depth computations for individual frames. The multiple depth layers are then combined with color information from the reference frame to create a Markov random field to model the segmentation problem. Finally, graphcut optimisation is employed to infer pixels belonging to the object to be segmented. The segmentation accuracy is evaluated over images from an outdoor video sequence demonstrating the viability for automatic object segmentation for mobile robots using monocular cameras as a primary sensor.

1 Introduction

Segmenting objects of interest in an image is of great significance in many robotic applications and has been a long sought after goal. The computer vision community have demonstrated excellent results from individual images using either manual initialisation from human input [Boykov and Jolly, 2001; Rother *et al.*, 2004; Li *et al.*, 2004] or pre-learnt models [Shotton *et al.*, 2006; Sun *et al.*, 2006; Yin *et al.*, 2010]. A number of recent results have demonstrated very complex dynamic objects such as dancers to be segmented with unprecedented accuracy [Guillemaut and Hilton, 2011]. However, in all these cases, camera positions have been either pre-calibrated or can be easily computed. This is rarely the case for any robotic application and we introduce a technique to potentially enable similar segmentation results with videos obtained from moving robotic platforms.

Traditional segmentation commonly relies on discriminative appearance between foreground and background, using cues such as color, motion and shape [Shotton *et al.*, 2006; Kolmogorov *et al.*, 2005; Yin *et al.*, 2010; Sun *et al.*, 2006]. However, these cues are not robust to variance in global illumination, cluttered and dynamic environments. It could therefore be argued that additional cues would aid in segmentation reliability and robustness. Recently depth information such as 3D point clouds from multiple view stereo or RGBD cameras has been made more accessible to the robotics and computer vision community. Depth provides an additional cue which is somewhat independent of colour, motion, and perspective change, and allows further improvements on the quality of segmentation that can be achieved [Zhang *et al.*, 2010; Wang *et al.*, 2010; He *et al.*, 2010].

In this paper, we focus on object segmentation of a single image from a video sequence using colour and depth information from not only the single image but also using information from neighbouring views within the video sequence. We combine the initial depth and color, as well as reprojected depths from surrounding views into a unified Markov Random Field (MRF) and infer the likelihood of particular pixels belonging to the object of interest using the Graph Cuts method [Boykov and Jolly, 2001]. The graph structure enables the costs from the reprojected depths to be easily incorporated into the inference problem and we show the improvements in segmentation with this additional information.

The rest of paper is organized as follows. In Section 2, we discuss related work. The overview of this method is addressed in Section 3. Some results and conclusion are given in Section 4 and 5, respectively.

2 Related Work

Recently, many methods have been exploited to achieve accurate and robust segmentation. Manipulation applications, especially vision based manipulation, require targets to be detected quickly and accurately. Robust and accurate object segmentation can assist robot systems to learn target objects efficiently improving interactive manipulation between object and manipulators [Li and Kleeman, 2011; Kenney *et al.*, 2009]. In human motion capture, segmented human bodies can prevent incorrect tracking and ambiguity between human and non-human motion [Liu *et al.*, 2011; Bray *et al.*, 2006]. In addition, there is significant work on segmentation in urban environments [Zhang *et al.*, 2010; Xiao and Quan, 2009; Pollefeys *et al.*, 2008] which can subsequently be used for object recognition and understanding, leading to urban 3D reconstruction and detecting loop closures for Simultaneous Localisation and Mapping.

With respect to fully automatic segmentation, most work focusses on motion cues to infer foreground [Meier and Ngan, 1998; Tsaig and Averbuch, 2002]. However, this technique is prone to failure if the foreground is static or the background is moving. Due to this disadvantage of automatic segmentation, semi-automatic segmentation has also received a lot of attention. Semi-automatic segmentation can be divided into two research directions. The first is classifier based segmentation, in which the foreground and/or background is learned *a priori* based on sample datasets. The learnt model is then applied to segment objects in previously unseen testing data [Shotton *et al.*, 2006; Sun *et al.*, 2006; Yin *et al.*, 2010]. This method requires good training data and a clean background to model the scene. Failures in segmentation can be caused because of dramatical scene changes, common in robotic applications.

The second semi-automatic segmentation research direction requires manual human input to discriminate foreground and background, *i.e.* interactive segmentation [Boykov and Jolly, 2001; Rother *et al.*, 2004; Li *et al.*, 2004]. These methods are commonly based on color modelling, and as previously mentioned, is not robust to illumination variation or color similarity between the foreground and background. Even though the work by Yu et al. uses spatial-color information to improve the robustness of segmentation [Yu *et al.*, 2007], it is assumed the color of the foreground and background changes rarely. Furthermore, as mentioned in their paper, the technique is vulnerable to error propagation. Our work presented in this paper focusses on interactive segmentation to initialise the foreground and background models but we also wish to highlight that the algorithms presented in our paper can be based on either technique without any changes to the framework.

Due to the advantages of depth as an additional cue to aid segmentation, a number of groups have employed this information. Approaches most related to the work presented in our paper are by Wang *et al.* [Wang *et al.*, 2010] and Zhang *et al.* [Zhang *et al.*, 2010]. Our method is inspired by [Wang *et al.*, 2010] with the following distinctions: 1) Depth is computed from a monocular RGB camera instead of a time-of-flight camera or stereo camera; 2) We test our method in an outdoor environment on a moving platform, not just with a static camera in an indoor environment; 3) We model depth information with Gaussian Mixture Models(GMMs) over manually selected seeds, while in [Wang *et al.*, 2010] a threshold is chosen for foreground and background pixels resulting in inaccuracies if part of the background shares the same depth with the foreground. With respect to [Zhang *et al.*, 2010], depth cues are used for multiple label segmentation, while we focus on binary segmentation. It should also be noted that depth information from neighbouring frames is used which we believe is the first time such a method has been presented in the literature.

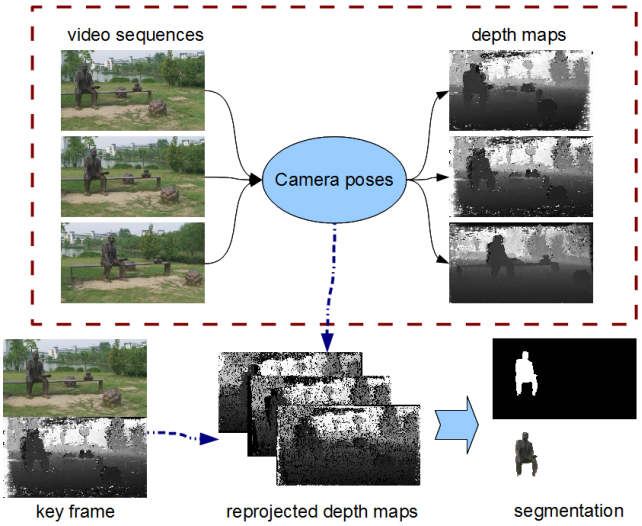

Figure 1: Overview of the proposed framework

3 Method Overview

Figure 1 shows an overview of the proposed method in this paper. It starts from estimating camera poses from multiple views using structure from motion. Initial dense depth maps for every frame are then estimated [McKinnon *et al.*, 2011]. Foreground and background seeds in a reference frame from which an object is to be segmented are then manually marked [Li *et al.*, 2004]. Note that we model foreground and background from the colour and depth seeds. Depth maps from neighbouring views are then reprojected into the reference frame from which a graph is constructed. Refinement of the foreground and background is then performed using Graph Cuts using colour, depth, and reprojected depth information.

3.1 Camera Pose and Depth Estimation

Camera poses and dense depth maps estimation from a video sequence has long been an active research topic in computer

vision. As the primary focus of this work is to investigate robust segmentation, we only briefly outline multi camera pose estimation and computation of the dense depth map for each view of the video sequence. The basic methods used in this paper are summarised as follows. SFM is used to calculate the camera poses from each view point in the video [Snavely *et al.*, 2006; Warren *et al.*, 2010]. For a reference frame in the video sequence, its neighbouring views are computed in terms of the corresponding camera pose. The initial dense depth maps for the reference frame are then estimated using a local cross-based method [Lu *et al.*, 2009; McKinnon *et al.*, 2011]. Depth fusion is also employed to refine the initial depth [McKinnon *et al.*, 2011; Merrell *et al.*, 2007] .

3.2 Segmentation Model

Once the depths and colour information for a reference frame have been computed, a graph is constructed representing the pixels and their initial foreground/background labels. This model is described below.

Smoothness Term

Let $L = (L_1, \cdots, L_s, \cdots, L_{|\mathscr{V}|})$ be the binary vector whose elements L_s specify assignments to pixel s in $\mathscr{V}$ which represents the set of all pixels in a given image. Each L_s can be either background ($L_s = 0$) or foreground ($L_s = 1$). Here the optimized vector L defines the final segmentation.

To compute the likelihood of a pixel belonging to foreground or background, the following energy function, similar to the Gibbs energy function described in [Geman and Geman, 1984], is used:

$$E(L) = \sum_{s \in \mathscr{V}} E_{data}(L_s) + \lambda \sum_{(s,t) \in \mathscr{N}} E_{smoothness}(L_s, L_t) \qquad (1)$$

where $E_{data}(L_s)$ is the likelihood energy encoding the cost when the label of the node s is L_s, and $E_{smoothness}(L_s, L_t)$ is the prior energy, representing the cost when the label of adjacent nodes (pixels) s and t are L_s and L_t, respectively. $\mathscr{N}$ denotes the set of 4-connected pairwise neighbouring pixels, and $\lambda \in [0,1]$ indicates the relative importance of the region-based energy versus the boundary-based energy.

The basic graph is illustrated in Figure 2. The term $E_{data}(L_s)$ corresponds to the likelihood cost between foreground and background in terms of the pixel label. The black solid arrow denotes the cost between pixel and foreground, while the black dashed arrow denotes the cost between pixel and background. The blue links in the graph represent prior information $E_{smoothness}(L_s, L_t)$, which preserves the discontinuity property in the image.

As the prior information is an inherent property preserved by the image itself, we construct this energy term as:

$$E_{smoothness}(L_s, L_t) = \frac{1}{d(s,t)} exp(-\frac{||I_s - I_t||^2}{2\sigma^2}) \qquad (2)$$

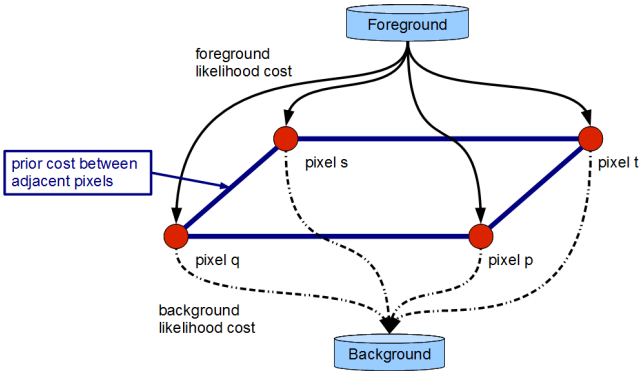

Figure 2: Illustration of graph cut method.

where I_s denotes the RGB value of a pixel s, $||I_s - I_t||^2$ is the Euclidean norm of the intensity difference, σ is the average intensity difference between neighbouring pixels in the image, which can be estimated as pixel noise introduced by the camera [Boykov and Jolly, 2001], and $d(s,t)$ is the spatial distance between two pixels s and t. This smoothness term favors the segmentation boundary where neighbouring pixels have large difference in terms of color.

Data Term

Here we will focus on modeling the data term by combining color and multiple view depths. In previous work [Sun *et al.*, 2006; Kolmogorov *et al.*, 2005; Rother *et al.*, 2004], GMMs were employed to model color distribution of foreground and background, respectively. In practice, foreground and background pixels always have significant overlap in the RGB color space. In this paper, depth is fused into the data term to increase the distinction between foreground and background.

More specifically, depth maps from the selected neighbouring views are reprojected to the reference frame using the previously computed camera poses. The foreground and background of the reprojected depth maps are also modelled with GMMs. The depth terms from the reference frame and reprojections are then combined as a weighted sum to form the depth data term. This is then combined with the color data term to generate the final data term in the segmentation model.

To achieve this we define the data as:

$$\sum_{s \in \mathscr{V}} E_{data}(L_s) = \lambda^{rgb} \sum_{s \in \mathscr{V}} D_{rgb}(L_s)$$

$$+\lambda^{depth} \sum_{s \in \mathscr{V}} \left[\frac{\lambda^{ref} D^{ref}_{depth}(L_s) + \sum_{i=1}^{n} \lambda_i^{reprj} D^{i,ref}_{depth}(L_s)}{n_{valid} + 1} \right] \qquad (3)$$

where $D_{rgb}(L_s)$ is color likelihood term, which models the foreground and background color likelihood. The depth likelihood term consists of two parts; $D^{ref}_{depth}(L_s)$ models the depth discrimination between foreground and background in the

reference frame, while $D^{i,ref}_{depth}(L_s)$ represents the likelihood of foreground and background depth reprojected from frame i to the reference frame. n is the number of neighbouring frames that are used to improve depth consistency and segmentation. λ^{rgb}, λ^{depth}, λ^{ref} and λ^{reprj}_i are weights for corresponding likelihood energy terms which control the influences of these terms. As RGB color and depth values are normalised into the same domain, *i.e.*, $[0,255]$, λ^{rgb} and λ^{depth} sum to 1, and the same rule is applied to λ^{ref} and $\sum_{i=1}^{n}\lambda^{reprj}_i$. It should be noted that n_{valid} is not necessary equal to n. This is due to possible occlusions where pixel s can not be reprojected from a neighbouring frame to the reference frame, which would result in $0 \leqslant n_{valid} \leqslant n$. In particular, λ^{ref} is enforced to 1 if $n_{valid} = 0$, otherwise the significance of the depth likelihood for foreground and background in the reference frame is decreased accordingly.

Here we employ GMMs to model likelihood energy terms. More specifically, these likelihood energy terms are defined as:

$$D_{rgb}(L_s = 1/0) = \sum_{i=1}^{K_{f/b}} w_i^{f/b} N(I_s|\mu_i^{f/b}, \Sigma_i^{f/b}) \quad (4)$$

$$D_{depth}(L_s = 1/0) = \sum_{i=1}^{K_{f/b}} w_i^{f/b} N(d_s|\mu_i^{f/b}, \Sigma_i^{f/b}) \quad (5)$$

$$D^{i,ref}_{depth}(L_s = 1/0) = \sum_{i=1}^{K_{f/b}} w^{f/b}_{i,ref} N(d_s|\mu^{f/b}_{i,ref}, \Sigma^{f/b}_{i,ref}) \quad (6)$$

where $N(\bullet)$ is a Gaussian distribution and (w_i, μ_i, Σ_i) denotes the mixture coefficient, the mean color or depth, and the color or depth covariance matrix of the ith component of the foreground and background GMMs. $K_{f/b}$ is the number of components of GMMs with respect to foreground and background, respectively. All these parameters (w_i, μ_i, Σ_i) can be learned from the manually selected pixel seeds.

The model now contains energy terms for colour, depth, and reprojected depths and can now be solved using the Graph Cuts Method as is the standard method for colour segmentation [Boykov and Jolly, 2001].

4 Results and Analysis

4.1 Data sets

We test the proposed method on an outdoor dataset (181 frames) filmed from a moving platform. We use 5 components in the GMM models for the foreground and background likelihood terms. The parameter λ between the data term and smoothness term is set as 200 in this paper. Other weights denoting the importance of depth data terms are addressed in Section 4.2. In terms of depth reprojection, 1, 4 and 12 neighbouring views are investigated. Typical individual frames from the data set are shown in Figure 3. For the following results, frame #35 is considered as the reference frame.

Figure 3: Data set used in this paper. The resolution is 960×540, in particular, the frame #35 is the reference frame used here.

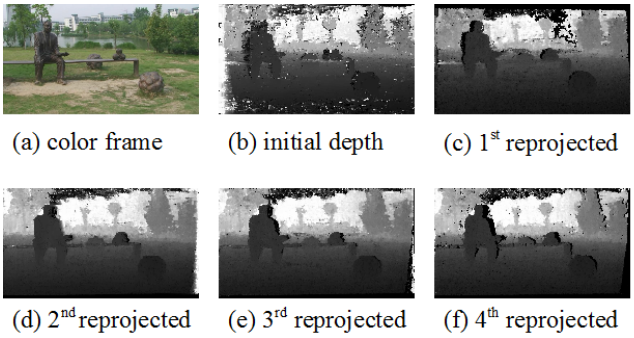

Figure 4: Original color and depth map of the reference frame, as well as the reprojected depth maps from neighbouring views. (a) is the color frame; (b) is the initial depth map; (c) - (f) are the reprojected depth maps from four different views to the reference frame by estimated camera poses.

As mentioned in Section 3, not only the original depth map for the reference frame is employed, but reprojected depth maps from neighbouring views are used. With respect to the reference frame used in this paper, the relevant depth map and reprojected depth maps are shown in Figure 4. It can be seen that the reference depth map is quite noisy, while the redundant information from reprojected depth maps should enhance pixel-wise depth estimation in the reference frame.

4.2 Segmentation results

To demonstrate the proposed method in Section 3, we investigate the use of depth as an additional cue. Specifically, we compare segmentation results using only color, only depth, joint color and reference depth, and joint color, reference depth as well as reprojected depth maps.

For modeling the data terms in Eq.3, we need some prior information of foreground and background to estimate the parameters in Eq.s 4, 5 and 6. Inspired by the interactive segmentation method in [Boykov and Jolly, 2001], we use masks to select a small number of pixel seeds. Two masks are shown in Figure 5 which are created from human input marking the reference frame with a virtual pen.

Once all the terms were initialised, we applied the graph cut method ([Boykov and Jolly, 2001]) to minimize the energy function described in Eq.1 to achieve our final segmentation. Segmentation results are shown in Figure 6. It should

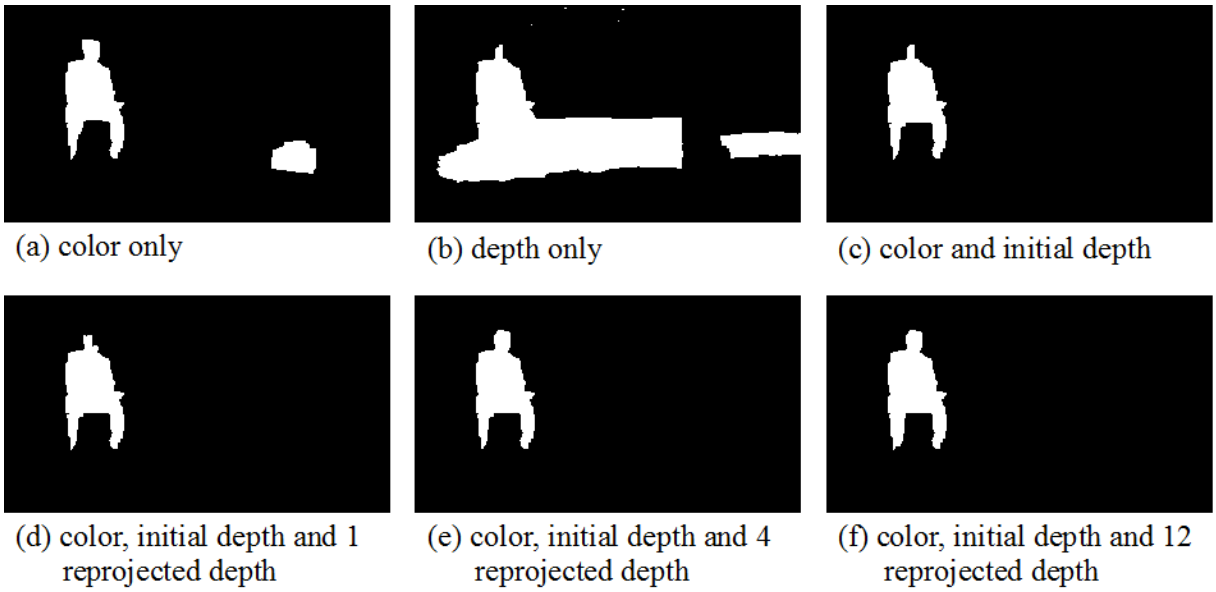

Figure 6: Segmentation results in different scenarios. (a) Only using color information; (b) Only using depth information; (c) Joint of color and initial depth information; (d) Joint of color, initial depth and only one reprojected depth map; (e) Joint of color, initial depth and four reprojected depth maps; (d) Joint of color, initial depth and twelve reprojected depth maps.

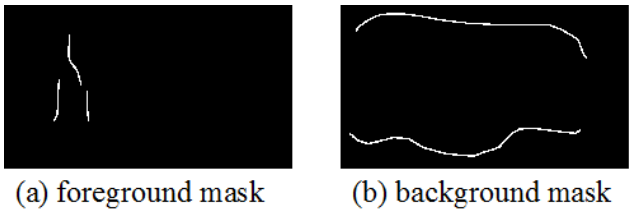

Figure 5: Foreground and background masks. In particular, while pixels are the corresponding pixel seeds which are used to model foreground and background of the reference frame. (a) Foreground mask; (b) Background mask.

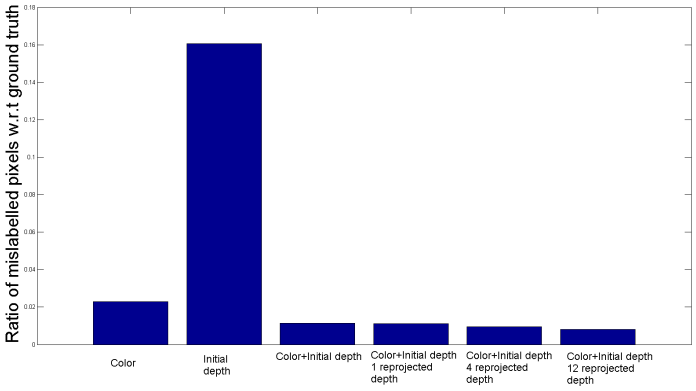

Figure 8: Segmentation obtained in different scenarios along with ground truth comparison, *i.e.*, color only, depth only, color + initial depth, color + initial depth + one reprojected depth, color + initial depth + four reprojected depths, color + initial depth + twelve reprojected depths

be noted that the relative weight for color λ^{rgb} is set as 0.6 in the experiment when using joint color and depth information. Qualitatively, segmentation is improved by using joint depth and color due to the color ambiguity between the statue and rock while there is similarity between the ground and the feet of the statue. The segmentation is further refined as reprojected depth maps are introduced into the model improving the reference depth estimation which is noisy and incomplete.

4.3 Evaluation and analysis

To achieve a quantitative evaluation of segmentation, we manually generated the ground truth of the reference frame shown in Figure 7. As in [He *et al.*, 2010], the number of pixels mislabelled as compared to the ground truth is expressed as a ratio with the total pixels in the original image:

$$\varepsilon = \frac{N_{error}}{N_{total}} \tag{7}$$

where N_{error} is the number of misclassified pixels compared to the ground truth and N_{total} is the number of total pixels in the original image.

In Figure 8, we plot the error histograms of the segmentation results in different scenarios with respect to the ground

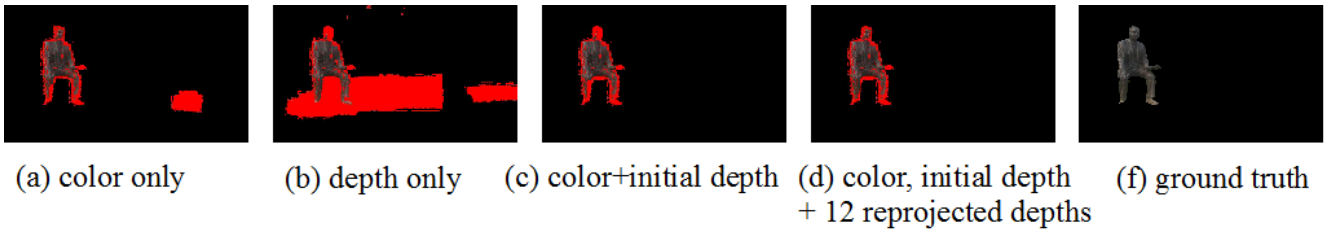

Figure 7: Segmentation obtained in different scenarios along with ground truth comparison. The segmentation error is shown in red.

truth. We find the combination of color and depth information improves the segmentation accuracy as compared to either color or depth alone. Furthermore, segmentation is improved by combining depth information from neighbouring views. However, we also notice that the quantitative accuracy of segmentation does not increase significantly even when additional reprojected depth maps are included in the computation. We find improvement is negligible for more than 12 neighbouring views are used.

5 Conclusion and Future work

In this paper, we have presented that depth as an additional cue to colour can improve segmentation accuracy and that reprojected depth information from neighbouring views can further refine the result. Qualitative and quantitative evaluation shows that the proposed method can achieve accurate segmentation under a challenging outdoor environment with an uncalibrated camera that is itself moving.

To improve the robustness and accuracy of segmentation further, we would like to explore several directions in the future. Currently, estimated depth maps can be quite noisy for uncalibrated cameras and as a results reprojected depth maps can in some circumstances degrade the solution. A calibrated camera (intrinsics only) will address this issue. We also intend to introduce heuristic methods used to infer the various weights in the model, *e.g.*, the weight between color and depth, and between the data and smoothness term. Finally, we would like to apply the proposed method for segmentation of indoor and outdoor video sequences automatically without any human intervention.

References

[Boykov and Jolly, 2001] Y. Boykov and M.P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in ND images. In *International Conference on Computer Vision*, volume 1, pages 105–112. Citeseer, 2001.

[Bray *et al.*, 2006] M. Bray, P. Kohli, and P. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. *Computer Vision–ECCV 2006*, pages 642–655, 2006.

[Geman and Geman, 1984] S. Geman and D. Geman. Stochastic relaxation,Gibbs distributions, and the Bayesian restoration of images. *IEEE trans on pattern analysis and machine inteligence*, 1984.

[Guillemaut and Hilton, 2011] J.Y. Guillemaut and A. Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International journal of computer vision*, pages 1–28, 2011.

[He *et al.*, 2010] 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.

[Kenney *et al.*, 2009] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 1377–1382. IEEE, 2009.

[Kolmogorov *et al.*, 2005] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bi-layer segmentation of binocular stereo video. 2005.

[Li and Kleeman, 2011] W.H. Li and L. Kleeman. Segmentation and modeling of visually symmetric objects by robot actions. *The International Journal of Robotics Research*, 2011.

[Li *et al.*, 2004] Y. Li, J. Sun, C.K. Tang, and H.Y. Shum. Lazy snapping. *ACM Transactions on Graphics (TOG)*, 23(3):303–308, 2004.

[Liu *et al.*, 2011] Y. Liu, C. Stoll, J. Gall, H. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view. In *International Conference on Computer Vision, 2011*. IEEE, 2011.

[Lu *et al.*, 2009] J. Lu, K. Zhang, G. Lafruit, and F. Catthoor. Real-time stereo matching: a cross-based local approach. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 733–736. IEEE, 2009.

[McKinnon *et al.*, 2011] David McKinnon, Hu He, Ben Upcroft, and Ryan N. Smith. Towards automated and in-

situ, near-real time 3-d reconstruction of coral reef environments. In *Proceedings of the MTS/IEEE Oceans '11 - Kona*, 2011.

[Meier and Ngan, 1998] T. Meier and K.N. Ngan. Automatic segmentation of moving objects for video object plane generation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):525–538, 1998.

[Merrell *et al.*, 2007] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. 2007.

[Pollefeys *et al.*, 2008] M. Pollefeys, D. Nistér, J.M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.J. Kim, P. Merrell, et al. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2):143–167, 2008.

[Rother *et al.*, 2004] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, page 314. ACM, 2004.

[Shotton *et al.*, 2006] 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*, pages 1–15, 2006.

[Snavely *et al.*, 2006] N. Snavely, S.M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[Sun *et al.*, 2006] J. Sun, W. Zhang, X. Tang, and H.Y. Shum. Background cut. *Computer Vision–ECCV 2006*, pages 628–641, 2006.

[Tsaig and Averbuch, 2002] Y. Tsaig and A. Averbuch. Automatic segmentation of moving objects in video sequences: a region labeling approach. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(7):597–612, 2002.

[Wang *et al.*, 2010] L. Wang, C. Zhang, R. Yang, and C. Zhang. TofCut: Towards Robust Real-time Foreground Extraction Using a Time-of-Flight Camera. 2010.

[Warren *et al.*, 2010] M. Warren, D. McKinnon, H. He, and B. Upcroft. Unaided stereo vision based pose estimation. *Australasian Conference on Robotics and Automation*, 2010.

[Xiao and Quan, 2009] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 686–693. IEEE, 2009.

[Yin *et al.*, 2010] 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*, pages 30–42, 2010.

[Yu *et al.*, 2007] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu. Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models. In *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*, pages 5–5. IEEE, 2007.

[Zhang *et al.*, 2010] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. *Computer Vision–ECCV 2010*, pages 708–721, 2010.