GrabcutD: Improved Grabcut Using Depth Information

Karthikeyan Vaiapury, Anil Aksay and Ebroul Izquierdo
MMV Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
United Kingdom
{karthike.vaiapury,anil.aksay,ebroul.izquierdo}@elec.qmul.ac.uk

ABSTRACT

Popular state of the art segmentation methods such as Grabcut include a matting technique to calculate the alpha values for boundaries of segmented regions. Conventional Grabcut relies only on color information to achieve segmentation. Recently, there have been attempts to improve Grabcut using motion in video sequences. However, in stereo or multi-view analysis, there is additional information that could be also used to improve segmentation. Clearly, depth based approaches bear the potential discriminative power of ascertaining whether the object is nearer of farer. In this work, we propose and evaluate a Grabcut segmentation technique based on combination of color and depth information. We show the usefulness of the approach when stereo information is available and evaluate it using standard datasets against state of the art results.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Multimedia application

General Terms

Algorithms

Keywords

Improved Grabcut, Segmentation, Silhouette, Depth

1. INTRODUCTION

Image segmentation has been very old and active research over several decades. It can be used in silhouette generation which is used in many potential computer vision applications such as 3D reconstruction using visual hull [16], event detection [17] etc. For example, Guillemaut *et al.,* has used joint robust Graphcut optimization and reconstruction for high quality free viewpoint video [20].

SMVC'10, October 29, 2010, Firenze, Italy.

Izquierdo *et al.,* has explained the key components that are necessary for an advanced segmentation toolbox [1]. The six different schemes deployed are variance-based detection of uniform regions, real-time histogram-based segmentation, fast nonlinear diffusion, diffusion-based object segmentation, morphology-based object segmentation and object segmentation by contour matching.

Popular state of the art segmentation methods such as Grabcut [8] include a matting technique to calculate the alpha values for boundaries of segmented regions. Conventional Grabcut relies only on color information to achieve segmentation. Our goal is to enhance the capability of Grabcut technique using depth information obtained from stereo or multiview analysis. Grabcut is an improved version of Graphcut which uses energy minimization techniques for segmentation [3]. Lazy snapping [19] is an interactive image cut system which is also based on Graphcut and is based on boundary refinement.

Recently, there have been few attempts made to improve the existing Grabcut technique. For example, Corrigan *et al.,* has provided a matting using motion extended Grabcut. However, their work is based on videos [4].

Han *et al.,* has extended the GrabCut integrating multi-scale nonlinear structure tensor [6]. Chen *et al.,* has provided improved Grabcut using Gaussian Mixture Model optimization [5]. Prakash *et al.,* has provided an combined approach based on both active contour and Grabcut for automatic foreground object segmentation [7]. Sun *et al.,* has proposed Flashcut for foreground segmentation based on flash, motion, and color information [15].

Up to our knowledge, no work has been attempted so far to include depth information with color segmentation, specifically for improving the existing Grabcut method.

As stated by Torralba *et al.,* there exists a strong relationship between structure of the scene and depth [14]. In stereo or multi-view analysis, there is additional information that could be also used to improve segmentation. In fact, during the generation of computer generated 3D movies and animations, depth information is known prior. The depth can also be generated using ToF cameras or using stereo vision techniques. Several methods for disparity estimation [10], [2] have been proposed. They can be categorized into local and global stereo methods.

Reinhard *et al.,* has used depth of field information in which they consider object which is in focus and other with out focus [18]. Zhu *et al.,* has provided a methodology for optimized depth inference where information from both depth and stereo images are considered. Thus obtained depth map is subsequently used to enhance matting [13].

Furthermore, Corrigan has also envisaged that depth based information would provide better segmentation technique. Depth information can be used to enhance matting [4].

Figure 1: Ballet sequence Image [12]

1.1 Grabcut Technique

Existing Grabcut technique works as follows: Initial trimap is created by user selecting a rectangle. Background class B is represented by the pixels outside rectangle and outer are unknown which belongs to foreground class A. The corresponding pixels are assigned to each class which is created using Orchard Bouman clustering algorithm. The GMMs are thrown away and new GMMs are learned from the pixel sets created in the previous set. The segmentation is estimated using graphcut which provides tentative classification of pixels belonging to the respective classes. The above process is iterated until convergence.

The dataset of ballet sequence is as shown in Fig.1. The disparity map of image can be obtained using many state of the art local or global methods. For example, the depth map of dancer image is as shown in Fig.1 above.

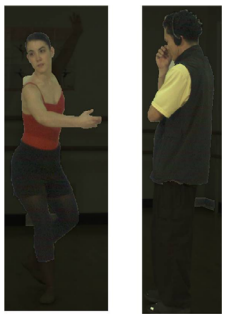

Figure 2: Existing method(Grabcut) results for Ballet sequence

As one can see from the Fig.2, quality of the existing Grabcut over the dataset is not satisfactory especially in the case of dancer. The hand portion is totally missing. This problem can be alleviated using depth information along with the available color based segmentation model.

2. PROPOSED METHOD

As discussed earlier, our proposed method is based on both depth and color segmentation model. Firstly, we discuss about depth based segmentation in order to show justification for using disparity along with color information later in our framework.

2.1 Depth Based Segmentation (DBS)

We have already proposed an optimal focused disparity framework in model based 3D reconstruction [2]. Using any of the available techniques, disparity map can be found. The algorithm used in generation of disparity maps by Zitnick *et al.,* [11] consists of three main steps: a) segmentation of image (smooth image using ansiotropic diffusion function), b) find initial disparity distribution (DSD) for each segment where DSD is set of probabilities over all disparities for individual segment in image and c) disparity smoothing using constraints which states that neighbouring segments with similiar color also has same disparity.

Disparity range information provides details regarding whether the object is nearer or farer. The depth range of interest can be obtained by finding the pixels with specific disparity range interval and subsequently, silhouette can be formed by assigning the pixels inside a region 255 value for highlighting while all others are assigned 0.

For example, in tsukuba image, we can isolate lamp and head separately Fig. 3,4. However, in dancer dataset, one can see that the man leg portion failed to segment properly (refer Fig. 3,1). This is due to the fact that too many pixels fall in the same range that makes classification of pixels more tedious. In fact, this problem could be alleviated by searching for pixels in a range within a bounding box.

Considering the various issues that have been discussed so far, we propose a novel framework that includes both color and depth information.

Figure 3: Results of Silhouette extraction from depth map

Usually, any given image can be represented as 3 channel image *R, G, B* components. RGB values encoded in 24 bits per pixel and are specified using three 8-bit unsigned integers (0 through 255) representing the intensities of red, green, and blue.

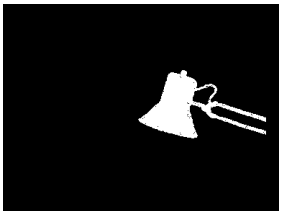

Figure 4: Segmentation of Lamp in Tsukuba using Depth

We also include the depth level (8 bit) information. In this work, we consider image as 4 channel components including disparity map [0-255]. The values 255 and 0 means nearer and farer respectively. The histogram of R,G,B and D channels of Ballet image is shown in Fig.5.

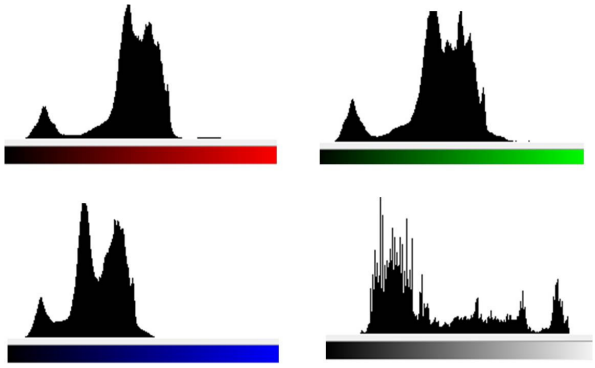

Figure 5: Histogram of a) R, b)G, c)B and d) D Channels of Ballet Image

2.2 Grabcut Using 4 channel

Let us consider image as an array $I = (I_1, ..I_n...I_N)$ which includes both R,G,B levels and depth values respectively. The segmentation is array of opacity values.

$\alpha = (\alpha_1, ...\alpha_N)$ at each pixel.

0 for background and 1 for foreground. θ is the parameter which represents foreground and background histogram distribution (histogram model).

$$\theta = h(I; \alpha), \alpha = 0, 1 \quad (1)$$

Given an image I and model θ, the segmentation task is to infer unknown opacity variables α.

The energy E is defined such that minimum represent good segmentation and it captures coherence in both color space and depth.

GMM components is a full covariance Gaussian mixture with K components. A vector $k = k1, ..kn, ..kN$ is defined and k_n assigns unique GMM component to each pixel either from background or foreground.

The Gibbs energy is of following form

$$E(\alpha, k, \theta, I) = U(\alpha, k, \theta, I) + V(\alpha, I) \quad (2)$$

The data term U which considers both color GMM and depth GMM models evaluates the fit of opacity distribution α to data I. It is defined as follows.

$$U(\alpha, k, \theta, I) = \sum_n D(\alpha_n, k_n, \theta, I_n) \quad (3)$$

where

$$D(\alpha_n, k_n, \theta, I_n) = -logp(I_n|, \alpha_n, k_n, \theta) - log\pi(\alpha_n, k_n) \quad (4)$$

$p(.)$ and $\pi(.)$ represent the Gaussian probability distribution and mixture weighting coefficient respectively.

Smoothness factor is defined as follows.

$$V(\alpha, I) = \gamma \sum_{m,n \in C} [\alpha_n \neq \alpha_m] exp - \beta ||I_m - I_n||^2 \quad (5)$$

C is the set of pairs of neighbouring pixels m, n. In this work, we propose to use scaling function in smoothness factor thereby emphasizing the importance of depth.

This can be achieved by using L2 norm

$$||I_m - I_n||_\tau^2 = \sqrt{\sum_{i=1}^{m,n} \tau_i (c(i,m) - c(i,n))^2} \quad (6)$$

Where τ_0, τ_1 and τ_2 denote scaling weightage factor assigned to R,G and B channels.

$$\tau_0 = \tau_1 = \tau_2 = (1 - \psi)/3; \quad (7)$$

The tuning parameter ψ represents the weightage factor given to depth channel component which is τ_3.

3. EXPERIMENTAL RESULTS

We have evaluated the proposed methodology using two publicly available standard datasets a) MSR (Ballet dancer) [Fig.1] and b) Middlebury Dataset (baby, Midd1, Tsukuba, Teddy, Art, Moebius) [Table 1]. As shown in Fig.6, hand

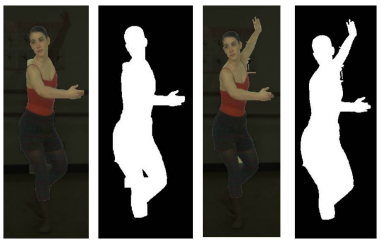

Figure 6: Ballet sequence Dancer: Grabcut and (b) GrabcutD (Color and Depth)

portion of dancer is not identified using Grabcut method(refer a), while our proposed method performed relatively better (refer b). Also, while considering the segmentation of man Fig.7, bottom portion of the image has not been identified using Grabcut while ours is able to segment.

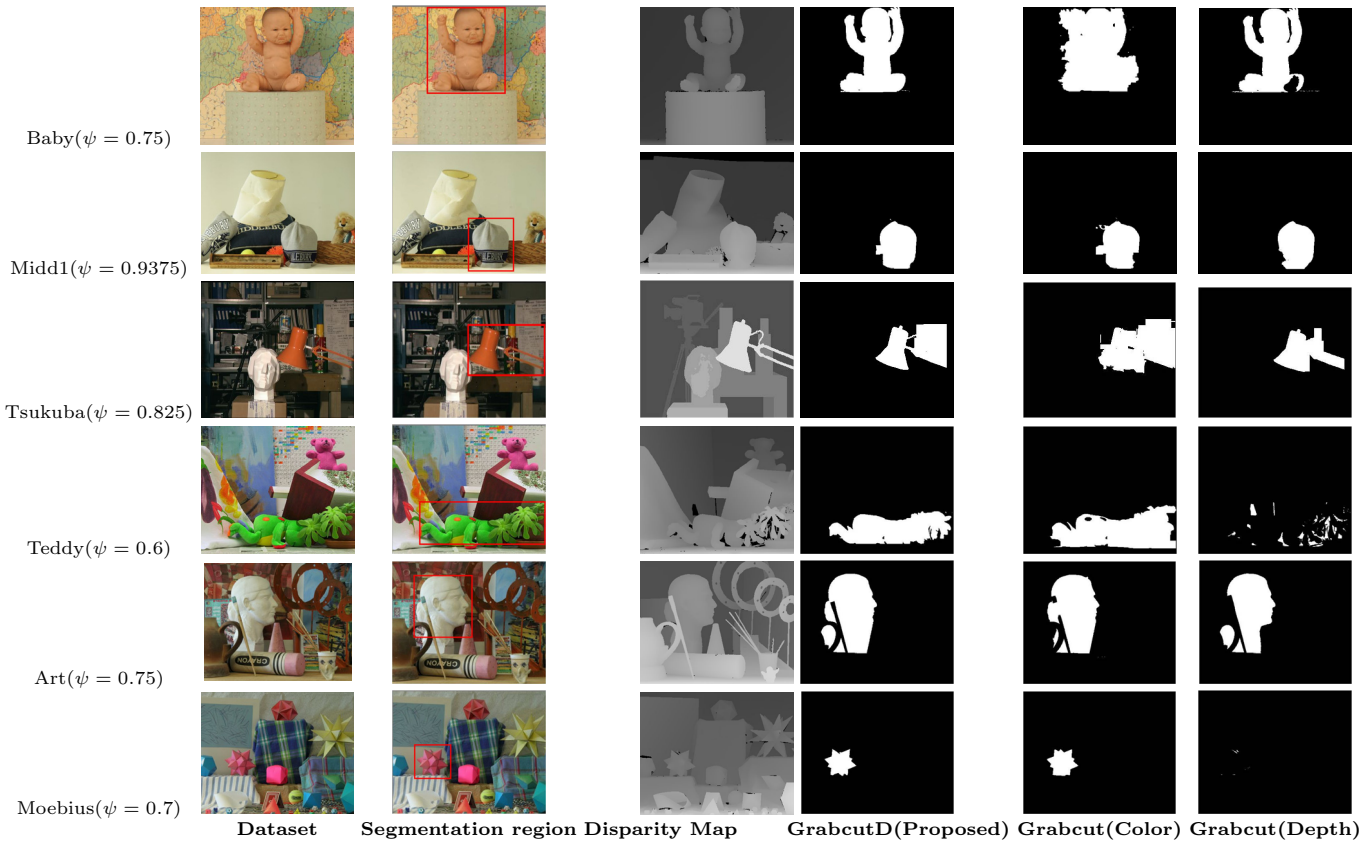

Table 1: Middelbury Dataset Results comparison using Grabcut (color), Proposed GrabcutD(color and depth) and Grabcut(depth)

Figure 7: Ballet sequence Man: Grabcut and (b) GrabcutD (Color and Depth)

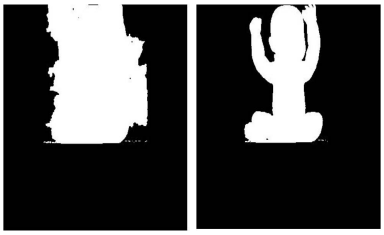

Figure 8: Baby ($\psi = 0.25$) **and** ($\psi = 0.75$)

As shown in Table 1, our algorithm is able to segment the objects better in atmost all dataset images.

For the first example (baby), the Grabcut technique (color) performs poorly since color of map in the background and that of baby significantly coincides. On the other hand, just using depth information, it fails to segment the leg portion since leg is closer to camera.

However, using both depth and color with tuning parameter value ($\psi = 0.75$) , our algorithm performs better as expected. In order to illustrate the convergence performance, we show different results of varying the parameters with corresponding silhouette information (Fig.8,9).

In the middl1 image as shown in Table 1, the right side curve portion of the hat is clearly seen in GrabcutD whereas in Grabcut, it is not the case. It can also be inferred that depth alone might not be sufficient for segmentation in some challenging datasets especially if there is no distinct depth information of the object of interest(refer teddy image). For this example, color information is also needed. ($\psi = 0.60$).

Further, in moebius image as shown in Table 1, star like structure on the top portion is not having clear depth which might affect the resulting quality. However, bottom portion of the image is segmented clearly using our method. The graph shown in Fig.10 displays the energy convergence of different middlebury datasets using the proposed GrabcutD method. As shown in the graph, Teddy took maximum of iterations n = 8 to converge to the minimum energy for segmentation while Midd1 image converged in n = 3 iterations.

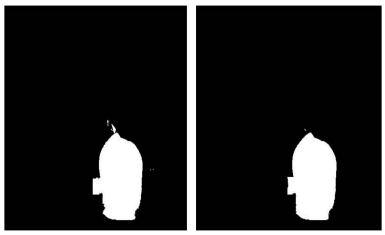

Figure 9: Midd1 ($\psi = 0.25$) **and** ($\psi = 0.9375$)

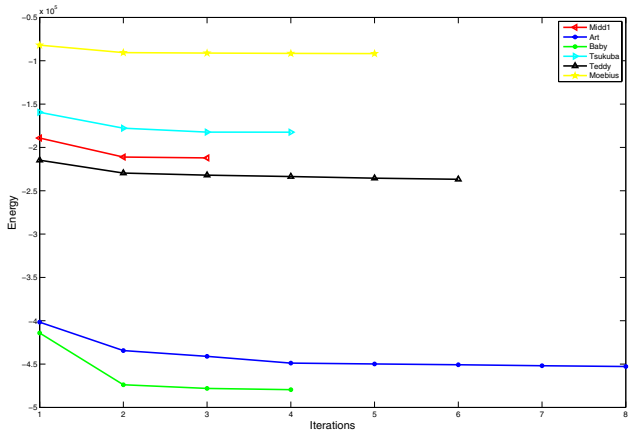

Figure 10: Energy Vs Iteration (GrabcutD)

4. CONCLUSION

In conclusion, based on the above results, we proved that depth based information will improve the Grabcut technique. In this work, we have proposed a novel method extended with depth information. The limitations of the Grabcut is overcome by integrating depth information. If there are challenging situations like having erroneous depth, it might affect the resulting quality. In order to show usefulness of the approach, we have conducted experiments on different standard datasets. The efficiency of the proposed methodology is clearly justified.

As future work, we would further investigate to learn the tuning parameter adaptively for any given model based on color and disparity information.

5. ACKNOWLEDGEMENTS

This research was partially supported by the European Commission under contract FP7-247688 3DLife. Our thanks to Microsoft Research and Middlebury University for posting the datasets.

6. REFERENCES

[1] E. Izquierdo and M. Ghanbari, Key Components for an Advanced Segmentation Toolbox, *In IEEE Transactions on Multimedia*, 2002, Vol. 4, No. 1, pp 97-113.

[2] K.Vaiapury and E.Izquierdo, A OFDP Framework in Model based Reconstruction, *The 12th International Asia-Pacific Web Conference,(APWEB)*, 2010.

[3] Y.Boykov, O.Veksler and R.Zabih, Fast Approximate Energy Minimization via Graph Cuts, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*,2001, Vol. 23, No. 11, pp. 1222-1239.

[4] D.Corrigan, S.Robinson and A.Kokaram, Video Matting using Motion Extended Grabcut, *IET European Conference on Visual Media Production (CVMP)*, 2008.

[5] D.Chen, B.Chen, G.Mamic, C.Fookes and S.Sridharan, Improved Grabcut Segmentation via GMM Optimisation, *Digital Image Computing: Techniques and Applications*, 2008, pp. 39-45.

[6] S.Han, W.Tao, D.Wang, T.Cheng and X.Wu, Image Segmentation Based on Grabcut Framework Integrating Multiscale Nonlinear Structure Tensor, *IEEE transactions on Image Processing*, 2009, Vol. 18, No.10, pp. 2289-2302.

[7] S.Prakash, R. Abhilash and S.Das, SnakeCut: An Integrated Approach Based on Active Contour and Grabcut for Automatic Foreground Object Segmentation, *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, 2007, Vol. 6, No 3.

[8] C.Rother, V.Kolmogorov and A.Blake, Grabcut-interactive foreground extraction using iterated graph cuts, *Proceedings of ACM SIGGRAPH*, 2004, pp. 309 - 314.

[9] J.F. Talbot and X.Xu, Implementing Grabcut, *research.justintalbot.org/papers/Grabcut.pdf*.

[10] S.T Barnard and W.B Thompson, Disparity analysis of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence, (PAMI)*, 1980, No.4.

[11] L.Zitnick, S.B.Kang, M.Uyttendaele, S.Winder and R.Szeliski, High Quality video interpolation using a layered representation, *ACM Transactions on Graphics*, pp. 600-608, 2004.

[12] MSR 3D Video Dataset, *research.microsoft.com/en-us/um/people/sbkang/3dvideodownload*.

[13] J.Zhu, M.Liao, R.Yang and Z.Pan, Joint depth and alpha matte optimization via fusion of stereo and time-of-flight sensor, *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2009, pp. 453-460.

[14] A.Torralba and A.Olivia, Depth Estimation from Image structure, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002, vol 24, No.9.

[15] Jian Sun, Jian Sun, Sing-Bing Kang, Zongben Xu, Xiaoou Tang and Heung-Yeung Shum. Flash Cut: Foreground Extraction with Flash/No-Falsh Image Pairs, *Computer Vision and Pattern Recognition, (CVPR)*, 2007.

[16] K.Forbes, A.Voigt and N.Bodika. Visual Hulls from Single Uncalibrated Snapshots Using Two Planar Mirrors. *In Proceedings of the Fifteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2004.

[17] A.Mokhber, C.Achard and M.Milgram, Recognition of human behavior by space-time silhouette characterization, *Pattern Recognition Letters*, 2008, Vol 29, Issue 1, pp. 81-89.

[18] E.Reinhard, E. and E.A.Khan, 2005. Depth of field

based alpha-matte extraction, *In Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization, (APGV)*, 2005, Vol. 95, pp. 95-102.

[19] Y.Li, J.Sun, C.Tang and H.Shum, Lazy Snapping, *ACM Transactions on Graphics*, 2004, Vol. 23, No. 3, pp. 303-308.

[20] J.Guillemaut, J.Kilner and A.Hilton, Robust Graphcut Scene Segmentation and Reconstruction for free view point video of complex dynamic scenes, *International Conference on Computer Vision, (ICCV)*, 2009.