Segmentation of Human Instances Using Grab-cut and Active Shape Model Feedback

Esmaeil Pourjam Ichiro Ide Daisuke Deguchi Hiroshi Murase
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
pourjami@gmail.com {ide,ddeguchi,murase}@is.nagoya-u.ac.jp

Abstract

For a long time, image segmentation has been of great interest for researchers. Although there have been many studies on automatic object segmentation, still a method which can cope with arbitrary input situations is out of reach. In this paper, we present "Active Shape Feedback Segmentation" (ASFSeg) method, which is a way to automatically segment human subjects (more accurately, pedestrians) from images. For this task, we try to use masks generated by the Active Shape Model (ASM) algorithm as a prior input for the Grab-cut technique to segment the desired human subject in the image without user interaction. To achieve this, we propose a feedback framework for the ASM sample generation. This feedback process compares the segmentation result mask from the Grab-cut stage and the samples generated in the ASM stage and then chooses the closest match to generate new samples for segmentation. This process is repeated until the system converges on one of the generated masks. Some experiments are carried out that shows the validity of our proposed method, which shows that the system can automatically segment the human subjects in provided pictures.

1 Introduction

For a long time, image segmentation has been of great interest for researchers. Although there have been many studies on automatic object segmentation [3,7,10,15], still a fully automatic system which can cope with arbitrary input situations is out of reach. Many factors such as scene illumination, object texture, colors, occlusions, noises, etc. affect the task of segmentation. Researchers have been trying to introduce methods that can cope with these factors as much as possible.

In contrast to automatic segmentation, in recent years, interactive image segmentation has shown some potential in the field of segmentation. Different methods have already been introduced in the literature such as graph-cut [1], obj-cut [5], lazy snapping [6], intelligent scissors [8], Grab-cut [11], TVSeg [12] and Geodesic matting [14]. Among these, those utilizing random field framework (graph-cut, Grab-cut, etc.) have shown more potential in comparison with the others. Based on that, some works have studied the application of this framework for automatic segmentation [3,7,9].

In this paper, we will introduce "Active Shape Feedback Segmentation" (ASFSeg) which is a method for automatic segmentation based on the Grab-cut segmentation framework and the Active Shape Model (ASM) method feedback for segmenting human subjects from images without involving user interaction. The paper is organized as follows: In section 2, we will consider some background researches done in this field. Also some brief explanation about ASM and Grab-cut will be presented. Section 3 will explain the ASFSeg proposed in this paper. Section 4 will show the experimental results of the proposed method.

2 Background

As mentioned earlier, various segmentation methods exist in the literature. In this section, first some related methods will be reviewed. After that, a brief introduction to ASM and Grab-cut will be presented.

2.1 Related work

Peng & Veksler [9] use a training set with different segmentation results of images (ten segmentations per image) manually labeled as good or bad to train a classifier. After the user inputs all background and foreground seeds, the system tries to find a result, classified as most confidently good segmentation. The user then may input some corrections and rerun the program to achieve better results.

Szummer et al. [13] try to learn segmentation parameters automatically using structured support vector machine $\text{SVM}_{\text{STRUCT}}$ and maximum-margin network learning. In their work, the user selects a polygon denoting the rough region of a foreground object and the system iteratively learns the parameters and segments the image.

Kuang et al. [4] try to learn two image features (color and texture) and a smoothing parameter from two polygons drawn by the user as regions for foreground and background. The method requires iterations to maximize a weighted energy function margin for estimating the parameters and at the same time segmenting the image. The interesting point in their research is that the system will learn optimized parameters specific to each input image.

Prakash et al. [10] apply active contour (snake) algorithm which is one of traditional methods of segmentation in conjunction with Grab-cut to increase the segmentation precision.

Li et al. [7] present a framework for segmenting objects in video sequences. In their work, a 3D graph cut based segmentation is proposed based on the precise segmentation provided by the user in the key frames. They also provide the user a way to correct the miss-segmentations in local frames.

Recently, Gulshan et al. [3] have taken the advantage of Microsoft Kinect and tried to propose an automatic segmentation algorithm based on that. They

first create a training dataset based on images acquired from Kinect (depth maps & image together). After that they extract HOG features from images in the data set. The extracted features are then used for training a classifier. When a new image is input to the system, this classifier generates a rough segmentation which is then given to a local Grab-cut stage for more precise segmentation.

2.2 Active Shape Models (ASM)

In our work, we will apply the original ASM method first introduced in [2]. A brief explanation about the method is presented below.

First we create vectors from the boundary of the objects in the training set which are aligned beforehand. Thus for each image, we will have a vector with n points like:

$$\mathbf{x}_i = [x_1, y_1, \ldots, x_n, y_n]^T \tag{2.1}$$

After aligning the shapes in the training set, we can calculate the mean model for the shape domain as:

$$\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \tag{2.2}$$

Based on these, we can calculate the covariance matrix:

$$\boldsymbol{S} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{2.3}$$

By analyzing this $2n \times 2n$ matrix and calculating its eigenvalues (λ_i) and corresponding eigenvectors ($\mathbf{p}_i$) and selecting a small set of them, we can generate new samples which approximate the original training samples with the following equation:

$$\mathbf{x}_{\mathsf{new}} = \overline{\mathbf{x}} + \boldsymbol{P}\mathbf{b} \tag{2.4}$$

Matrix $\boldsymbol{P}$ is made by setting the selected eigenvectors as columns, and $\mathbf{b}$ is a vector of weights like:

$$\begin{aligned} \boldsymbol{P} &= [\mathbf{p}_1, \ldots, \mathbf{p}_t] \\ \mathbf{b} &= [b_1, \ldots, b_t] \end{aligned} \tag{2.5}$$

A suitable limit for the weights can be described as:

$$-3\sqrt{\lambda_k} \leq b_k \leq 3\sqrt{\lambda_k},\ k \in [1, \ldots, t] \tag{2.6}$$

In Figure 1, some samples generated by changing the values of b_k are presented. Note that each set of samples is created by changing just one value, for example, $\mathbf{b} = [b_1, 0, \ldots, 0]^T$.

2.3 Grab-cut

In our work, we will also apply the Grab-cut method first introduced by Rother et al. [11]. This segmentation technique is an upgraded model of the graph-cut segmentation algorithm [1], which incorporates the color features and a better iterative energy minimization procedure. However, there is one main problem here; this method cannot segment the image completely just by itself and relies on the user for further foreground and background seeds selection. The basic process flow for this method is presented in Figure 2.

(a) Case of changing b_1

(b) Case of changing b_2

Figure 1. Some samples generated with ASM.

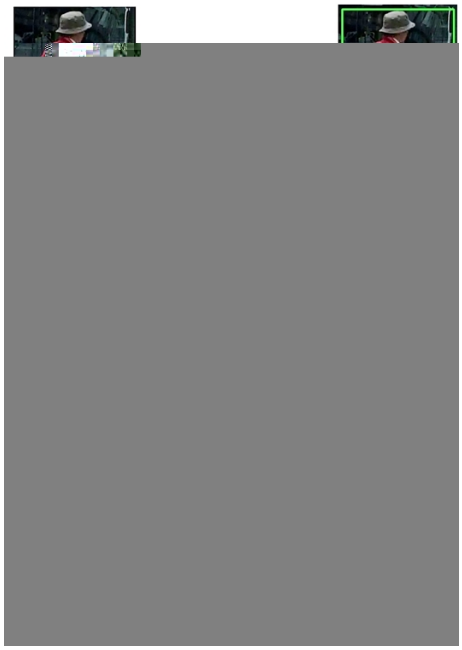

Figure 2. Basic process flow in the Grab-cut segmentation framework.

3 Active Shape Feedback Segmentation

In this work, we try to exploit the possibility of using the generated samples from an ASM system, based on real human training samples, as a basis for human object segmentation. We call our proposed method "Active Shape Feedback Segmentation" and in the rest of the paper we will use the abbreviated form "ASFSeg" to refer to it. The general process flow of the system is shown in Figure 3.

Two main points presented in this work are 1) the usage of ASM generated masks as priors for Grab-cut segmentation, and 2) the implementation of a feedback system which provides the mask generation step with needed information for generating better masks.

As it can be seen in Figure 3, the system works in the following manner: First some new samples (N samples) are generated based on the existing training dataset. One of these generated samples is selected randomly and is converted into a trimap mask (It will be explained in more detail later) and is used as a prior mask for the Grab-cut's segmentation step. After segmentation, the resulted output for foreground will be compared with the input prior mask and the error rate will be calculated. Also, the output will be compared

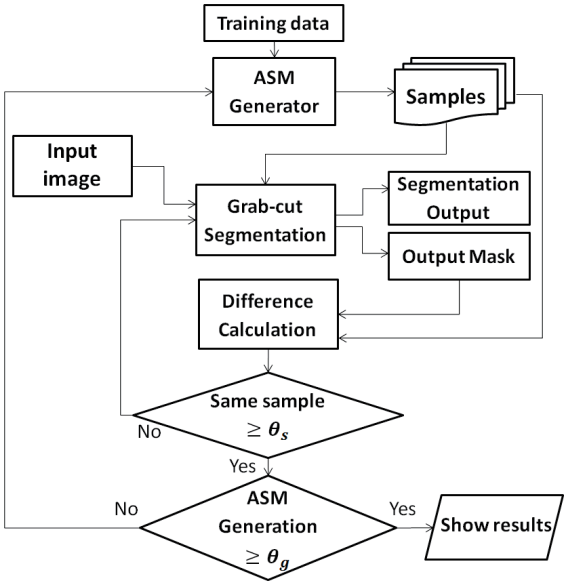

Figure 3. Process flow of the ASFSeg method.

with other generated samples and the most simillar one (the sample whose error rate is less than the others) would be selected for the segmentation process. This step would be repeated until the system converges to one of the samples (the same sample is selected repeatedly; more than θ_s times). Next, parameters for the first and the second samples with the least error rate are calculated and based on that, a new set of samples (again, N samples) are generated. The same process is repeated for finding samples with the least error. The whole process of sample generation and image segmentation will be repeated for more than θ_g times and the final result would be presented to the user.

Since our assumption is that the system cannot use any kind of user provided data, we here use the generated sample as a ground truth for segmentation provided that the desired object should be similar to the provided mask to some degree.

The Grab-cut also can take in different types of masks as its input. The mask pixels can be assigned with four different states of "Probably foreground", "Probably background", "Foreground", or "Background". In this work the Grab-cut mask is created so that the eroded part of the sample would be labeled as "Foreground", the dilated part would be "Probably foreground" and the rest would be labeled as "Probably background" (as in Figure 4(b)).

Although the Grab-cut algorithm used here is almost the same as in the original paper [11], we slightly modified the smoothness parameter by adding a distance penalty between the generated mask boundary and input image pixels.

4 Experimental results

In this section, results of experiments for validating the ASFSeg method are presented. For these experiments, 13 samples for training the ASM model have been used. At each stage of sample generation, we create N=30 samples from which, one is selected randomly for segmentation. As for the criteria to stop the

(a) Generated sample (b) Resulted trimap

Figure 4. Converting a generated sample to a trimap.

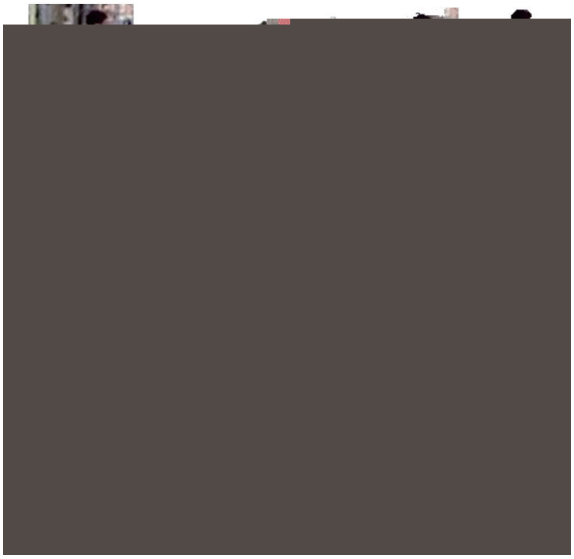

Figure 5. Comparison between the segmentations by Grab-cut, N-cut and ASFSeg methods. The images have been resized for better presentation. Actual size is respectivly: 82×134, 145×206 and 138×228 pixels.

segmentation and the generation process, experiments show that if we set $\theta_s = 5$ and $\theta_g = 3$, usually desired results would be achieved.

The comparison is done between the original Grab-cut segmentation, Normalized Cut (N-cut) segmentaion [15,16] and the ASFSeg method, with and without ASM feedback. The precision of the methods is calculated based on the number of pixels that have been segmented correctly as foreground or background in comparison to the ground truth provided by manual segmentation of the desired object.

For Grab-cut segmentation, the code provided by OpenCV opensource library, and for N-cut Segmentaion, the code provided by [16] were used.

Table 1 shows the segmentation precision for each set of pictures in Figure 5. Two types of errors are calculated here. "Foreground error" is the percentage of foreground pixels which were segmented as background in comparison with the actual total number of foreground pixels. "Background error" is also calculated in the same way, but this time the percentage of background pixels segmented as foreground is concerned.

Since in our work, we consider that the training data contains samples in which just human body information exists, if a subject holds an object, for example a carry-bag or the like, there would be some segmentation error. As a result, Table 1 uses the ground truth with just body segmentation. Meanwhile in Table 2, a

Table 1. Comparison of the precision between the segmentation methods.

	FG error (%)			BG error (%)		
Image set	1	2	3	1	2	3
Grab-cut	100.00	2.02	1.50	0.00	14.57	16.76
ASM (no feedback)	16.33	3.01	14.25	34.74	3.00	13.81
N-Cut	4.32	8.73	17.89	20.79	9.61	8.51
ASFSeg	4.52	3.16	2.32	4.06	2.15	3.09

* FG: Foreground, BG: Background

Table 2. Comparison of the precision between the segmentation methods when the holding objects are considered.

	FG error (%)			BG error (%)		
Image set	1	2	3	1	2	3
Grab-cut	100.00	1.50	0.83	0.00	13.52	13.81
ASM (no feedback)	22.51	2.37	14.41	36.23	1.81	11.52
N-Cut	3.82	8.36	17.05	19.05	8.65	5.76
ASFSeg	11.90	2.48	6.94	5.16	0.94	1.87

* FG: Foreground, BG: Background

complete foreground object (human & bag) is used as the ground truth.

5 Conclusions

In this paper, we presented a method that can perform segmentation of pedestrians automatically. The main idea is to make the process automatic by using the ASM model generation algorithm to generate some prior masks for the Grab-cut segmentation step instead of asking the user to identify the background and foreground seeds. It should be mentioned that even if the ASFSeg method can perform the segmentation automatically and sometimes with better accuracy in comparison with the Grab-cut method, still there are some problems that have to be solved so that it becomes applicable in real situations. For example, since the Grab-cut just uses color features for foreground and background segmentation, if the color distribution between an object and its background is not very different, we will not bo able to obtain a satisfactory result.

As for the future work, we would like to:

(1) Make a more complete training dataset for the ASM generation step which includes more variations in the model.

(2) Use more features (color, texture, distance penalty, etc.) for modeling the object features.

(3) Optimize the code, since the time consumption is one of the problems here.

6 Acknowledgments

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research. Also parts of work was developed based on the MIST library (`http://mist.murase.m.is.nagoya-u.ac.jp/`).

References

[1] Y.Y. Boykov and M.-P. Jolly: "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in ND Images", *In Proc. 8th IEEE Conf. on Computer Vision*, vol.1, pp.105–112, 2001.

[2] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham: "Active Shape Models—Their Training and Application", *Computer Vision and Image Understanding*, vol.61, no.1, pp.38–59, 1995.

[3] V. Gulshan, V. Lempitsky, and A. Zisserman: "Humanising GrabCut: Learning to Segment Humans Using the Kinect", *In Proc. 13th Int. Conf. on Computer Vision Workshops*, pp.1127–1133, 2011.

[4] Z. Kuang, D. Schnieders, H. Zhou, K.-Y.K. Wong, Y. Yizhou, and B. Peng: "Learning Image-Specific Parameters for Interactive Segmentation", *In Proc. 25th IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp.590–597, 2012.

[5] M.P. Kumar, P.H.S. Torr, and A. Zisserman: "Obj Cut", *In Proc. 18th IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol.1, pp.18–25, 2005.

[6] Y. Li, J. Sun, C.K. Tang, and H.-Y. Shum: "Lazy Snapping", *ACM Trans. on Graphics*, vol.23, no.3, pp.303–308, 2004.

[7] Y. Li, J. Sun, and H.-Y. Shum: "Video Object Cut and Paste", *ACM Trans. on Graphics*, vol.24, no.3, pp.595–600, 2005.

[8] E.N. Mortensen and W.A. Barrett: "Intelligent Scissors for Image Composition", *In Proc. 22nd Conf. on Computer Graphics and Interactive Techniques*, pp.191–198, 1995.

[9] B. Peng and O. Veksler: "Parameter Selection for Graph Cut Based Image Segmentation", *In Proc. 19th British Machine Vision Conf.*, pp.160–170, 2008.

[10] 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*, vol.6, no.3, pp.13–28, 2007.

[11] C. Rother, V. Kolmogorov, and A. Blake: "Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts", *ACM Trans. on Graphics*, vol.23, no.3, pp.309–314, 2004.

[12] M. Unger, T. Pock, W. Torbin, D. Cremers, and H. Bischof: "TVSeg—Interactive Total Variation Based Image Segmentation", *In Proc. 19th British Machine Vision Conf.*, vol.2, pp.335–354, 2008.

[13] M. Szummer, P. Kohli, and D. Hoiem: "Learning CRFs Using Graph Cuts", *In Proc. 10th European Conf. on Computer Vision, Part 2, Lecture Notes in Computer Science*, vol.5303, pp.582–595, 2008.

[14] G. Varun, C. Rother, A. Criminisi, A. Blake, and A. Zisserman: "Geodesic Star Convexity for Interactive Image Segmentation", *In Proc. 23rd IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp.3129–3136, 2010.

[15] J. Shi and J. Malik: "Normalized Cuts and Image Segmentation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp.888–905, 2000.

[16] T. Cour, S. Yu, and J. Shi: "Normalized Cut Segmentation Code", `http://www.timotheecour.com/software/ncut/ncut.html`, retrieved 2013.

[17] "Open Source Computer Vision Library", `http://www.opencv.org/`, retrieved 2013.