Accurate Foreground Extraction Using Graph Cut with Trimap Estimation

Jung-Ho Ahn and Hyeran Byun

Dept. of Computer Science, Yonsei University, Seoul, Korea, 126-749

Abstract. This paper describes an accurate human silhouette extraction method as applied to video sequences. In computer vision applications that use a static camera, the background subtraction method is one of the most effective ways of extracting human silhouettes. However it is prone to errors so performance of silhouette-based gait and gesture recognition often decreases significantly. In this paper we propose two-step segmentation method: trimap estimation and fine segmentation using a graph cut. We first estimated foreground, background and unknown regions with an acceptable level of confidence. Then, the energy function was identified by focussing on the unknown region, and it was minimized via the graph cut method to achieve optimal segmentation. The proposed algorithm was evaluated with respect to ground truth data and it was shown to produce high quality human silhouettes.

1 Introduction

Background subtraction has been widely used to detect and track moving objects obtained from a static camera. Recently background subtraction methods have been used to assist in human behavior analysis such as surveillance and gait and gesture recognition[10, 13, 18]. In gait and gesture recognition silhouettes are used to determine configurations of the human body. However these human silhouettes are often not accurate enough and recognition performance can decrease significantly when using these silhouettes [11]. Especially, shadows can not always be properly removed, and some parts of the silhouette can be lost when occluding object parts have similar colors as the occluded background areas. Figure 1 demonstrates these problems.

Background subtraction has long been an active area of research. Horprasert *et al.* [6] proposed a robust background subtraction and shadow detection method which was applied to an object tracking system together with an appearance model [13]. The adaptive background subtraction method using the Gaussian Mixture Model (GMM) [14] was presented and it was applied to foreground analysis with intensity and texture information [16]. A joint color-with-depth observation space [7] and an efficient nonparametric background color model [5] were also proposed to extract foreground objects.

In this paper we propose a foreground segmentation that consists of two steps; trimap estimation and fine segmentation using a graph cut. In trimap estimation we estimate the confident foreground and background regions and leave the

L.-W. Chang, W.-N. Lie, and R. Chiang (Eds.): PSIVT 2006, LNCS 4319, pp. 1185–1194, 2006.

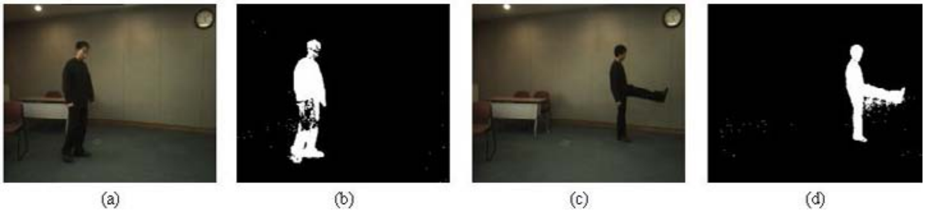

Fig. 1. Problems of the background subtraction method to extract human silhouettes. In (a) and (b) some parts of the human silhouette disappear when the color of these parts is too similar to that of the occluded background area. (c) and (d) shows the shadow problem. These silhouettes were obtained with the Horprasert's algorithm [6].

dubious area unknown. The amount of the estimated foreground and background regions is related to the level of confidence. Fine segmentation focus on the unknown regions and uses the energy minimization technique. Throughout this paper it is assumed that a static background is available, images are captured by a static mono camera, and only one person enters a scene.

As a pioneer work of object segmentation using graph cut, the user-interactive segmentation technique was proposed by Boykov and Jolly [1]. GrabCut [12] and lazy snapping[9] were other user-interactive image cutout systems based on a graph cut. Lazy snapping used a graph cut formulation based on regions generated by the watershed algorithm, instead of the image pixels. In this paper we also perform image segmentation but it is for estimating the trimap. Recently, the Layered Graph Cut(LGC) method based on color, contrast and stereo matching [8] information has been proposed to infer the foreground by using a pair of fixed webcams.

Section 2 describes the trimap concept and the proposed graph cut framework using the trimap information. Section 3 explains the concrete trimap estimation method using region likelihood. Experimental results are given in section 4and then conclusions are presented in section 5.

2 Graph Cut Segmentation with Estimated Trimap

2.1 Estimated Trimap

In natural image matting[4][15] with still images the trimap is supplied by the user. This trimap partitions the image into three regions: firm foreground, firm background and unknown regions. In unknown regions, the matte can be estimated using the color statistics in the known foreground and background regions. In this paper we attempt to estimate the trimap in video applications without user interaction.

Foreground segmentation refers to a binary labeling(classification) process that assigns all the pixels in a given image to either foreground or background. It is very hard to label *all* the pixels in the image correctly, but *some* parts of the

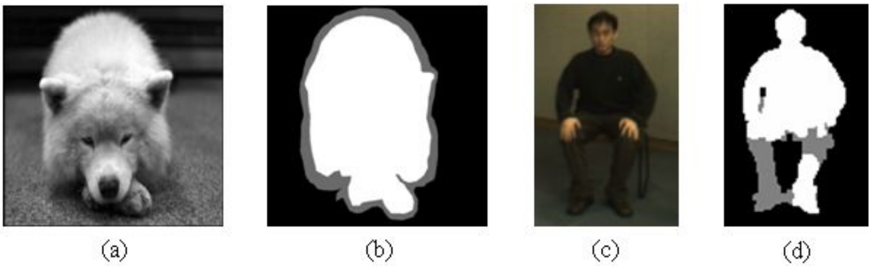

Fig. 2. Trimap Comparison. (a) and (b) show typical trimaps in natural image matting [15], whereas (c) and (d) shows the trimaps used in the proposed algorithm.

image can be easily labeled with simple ideas or features. In trimap estimation we pre-determine the labels of each of the pixels in the area that can be easily labeled. Then the trimap information is reflected into the energy function for fine segmentation. In section 3 we propose a trimap estimation method that uses the background model. Traditionally the unknown regions of the trimap is located between foreground and background areas but they can theoretically be placed anywhere in our estimated trimap. Figure 2 shows some typical examples of the trimaps of image matting and the proposed method.

The trimap $\mathcal{T}$ can be viewed as the function from the set of pixels $\mathcal{P}$ of the image to be segmented to the label set $\mathcal{L}_T$

$$\mathcal{T} : \mathcal{P} \rightarrow \mathcal{L}_T = \{-1, 0, 1\} \tag{1}$$

where -1, 0, and 1 represent unknown, background and foreground respectively. In every frame we estimate the trimap $\mathcal{T}$ and define the sets $\mathcal{O}$ and $\mathcal{B}$ by

$$\mathcal{O} = \{p \in \mathcal{P} | \mathcal{T}(p) = 1\}, \qquad \mathcal{B} = \{p \in \mathcal{P} | \mathcal{T}(p) = 0\}. \tag{2}$$

The set of unknown pixels $\mathcal{U}$ is defined by $\mathcal{P} - (\mathcal{O} \cup \mathcal{B})$. We call the pixels in $\mathcal{O}$ the foreground seeds. The pixels in $\mathcal{B}$ are called the background seeds.

2.2 Graph Cut with the Estimated Trimap

In this section we describe the energy function for fine segmentation. The energy function is similar to that used in the GrabCut method [12] but it explores both trimap information and the color likelihoods. We consider a standard neighborhood system $\mathcal{N}$ of all unordered pairs $\{p, q\}$ of neighboring pixels, and define $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_{|\mathcal{P}|})$ as the image where $\mathbf{z}_n$ is the RGB color vector for the nth pixel and $f = (f_1, f_2, \cdots, f_{|\mathcal{P}|})$ as a binary vector whose components f_p specify label assignments to pixels p in $\mathcal{P}$. Each f_p can be either 1 or 0 where 1 represents the foreground and and 0 represents the background. The vector f defines a segmentation. Given an image, we seek the labeling f that minimizes the energy

$$E(f) = \gamma D(f) + V(f) \tag{3}$$

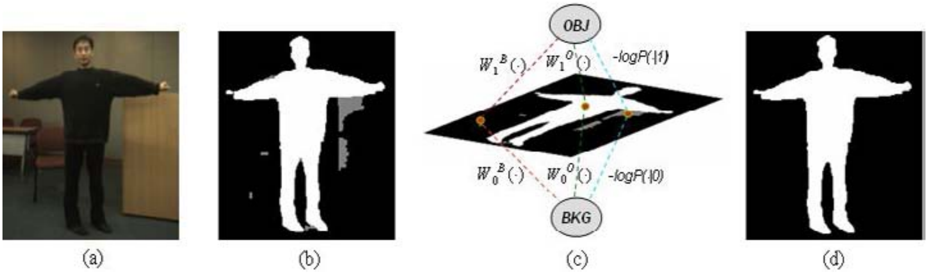

Fig. 3. Graph cut with the estimated trimap information. (a) input image, (b) trimap; white, black, gray areas indicate estimated foreground, background and unknown regions, respectively. (c) graph cut framework with the estimated trimap information, (d) the extracted foreground region.

where coefficient γ specifies the relative importance of the data term $D(f)$ and the smoothness term $V(f)$. The form of $D(f)$ and $V(f)$ are given by

$$D(f) = \sum_{p \in \mathcal{P}} D_p(f_p)$$
$$V(f) = \sum_{\{p,q\} \in \mathcal{N}} \delta(f_p, f_q) V_{p,q}(f_p, f_q).$$

where $\delta(f_p, f_q)$ denotes the delta function defined by 1 if $f_p \neq f_q$ and 0 otherwise. Given an image, it is necessary that the segmentation boundaries align with contours of high image contrast. This process is modeled using the smoothness term $V_{p,q}$ defined by

$$V_{p,q}(f_p, f_q) = \exp\left(-||\mathbf{z}_p - \mathbf{z}_q||^2/\beta\right) \tag{4}$$

where the constant β is chosen by the expectation of $2||\mathbf{z}_p - \mathbf{z}_q||^2$ over all $\{p, q\} \in \mathcal{N}$. The data term D_p measures how well label f_p fits pixel p given the observed data $\mathbf{z}_p$. We model the data term by using the given trimap $\mathcal{T}$ and the foreground and background color likelihoods of $P(\cdot|1)$ and $P(\cdot|0)$, respectively. The likelihoods are modeled by using Gaussian mixtures in the RGB color space, learned from image frames labelled from earlier in the sequence. Using the trimap information the data term D_p can be defined as

$$D_p(f_p) = \begin{cases} -\log P(\mathbf{z}_p|f_p) & \text{if } p \in \mathcal{U} \\ W^{\mathcal{O}}_{f_p}(\mathbf{z}_p|f_p) & \text{if } p \in \mathcal{O} \\ W^{\mathcal{B}}_{f_p}(\mathbf{z}_p|f_p) & \text{if } p \in \mathcal{B} \end{cases} \tag{5}$$

In the above equation $W^{\mathcal{O}}_{f_p}(\mathbf{z}_p)$'s and $W^{\mathcal{B}}_{f_p}(\mathbf{z}_p)$ are defined by

$$\begin{aligned} W^{\mathcal{O}}_{f_p}(\mathbf{z}_p) &= -W_{f_p} \log P(\mathbf{z}_p|f_p) \\ W^{\mathcal{B}}_{f_p}(\mathbf{z}_p) &= -W_{1-f_p} \log P(\mathbf{z}_p|f_p) \end{aligned} \tag{6}$$

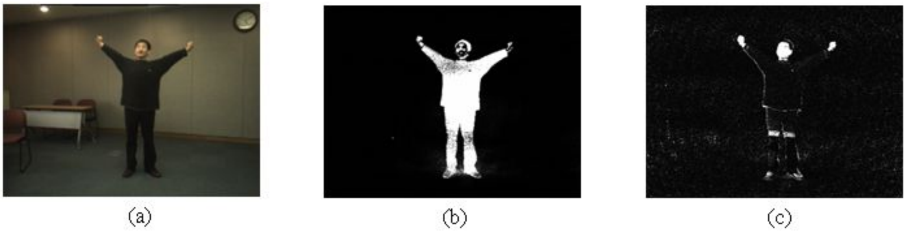

Fig. 4. Pixel foreground likelihoods: (a) image of the 250th frame in the test data $JH1$, (b) brightness distortion likelihood $-\log f^b(\alpha_p|p)$, (c) scaled chromaticity distortion likelihood $-\eta \log f^c(\gamma_p|p)$, where $\eta = 5$. The likelihoods are truncated by 255.

for all $p \in \mathcal{P}$ and $W_1 > 1 > W_0$. This data term model boosts the likelihood of $-\log P(\cdot|f_p)$ but suppresses $-\log P(\cdot|1 - f_p)$ for the pixel p of $\mathcal{T}(p) = f_p$ where f_p= 0 or 1. The estimated foreground and background seeds can be incorrectly labeled, thus the trimap information is used to distort the color likelihoods in the data term and the final segmentation label is given by a graph cut.

Minimization of the energy (3) is done by a standard min-cut/max-flow algorithm [2]. Figure 3 (c) illustrates the proposed graph cut framework using the estimated trimap information.

3 The Trimap Estimation Method Using Region Likelihood

To extract an accurate foreground silhouette, it is very important that the set $\mathcal{O}$ in the trimap $\mathcal{T}$ should not contain any background pixels but should contain as many foreground pixels as possible, and vice versa for the set $\mathcal{B}$. To accomplish this, we estimate the seeds of $\mathcal{T}$ in the region units, instead of in the pixel units. Region unit processing can have the a denoising effect which can allow for recovering clear outlines of the foreground objects. This processing was motivated from the perceptual grouping principles for object segmentation of still images[17]. We first propose the background color distribution model in section 3.1, and the region likelihood is determined in section 3.2. Finally, the region decision function for trimap estimation is proposed in section 3.3.

3.1 Brightness and Chromaticity Background Likelihood

Horprasert et. al[6] proposed the statistical background model that separated the brightness from the chromaticity component. They modeled a pixel p by 4-tuple $< \mu_p, \sigma_p, a_p, b_p >$ where μ_p was the expected color value, σ_p was the standard deviation of the RGB color value, a_p was the variation of the brightness distortion, and b_p was the variation of the chromaticity distortion of pixel p.

Using the background model, we propose the foreground likelihood $l(p)$ for each pixel p. The object likelihood is decomposed of the brightness and chromaticity likelihoods. From the definitions of the distortions , the distribution

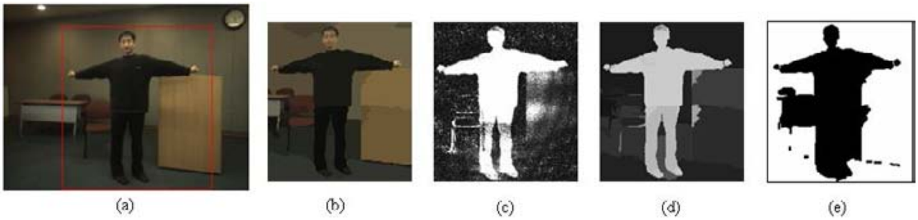

Fig. 5. Region unit processing for trimap estimation. (a) original image of the 581st frame in the test data $JH3$ with a bounding box, (b) mean shift segmentation of the image within the bounding box, (c) pixel foreground likelihoods $l(p)$'s, (d) naive region likelihood $L_p(R_i)$ (e) white area shows the regions that touch the bounding box.

f^b of the brightness distortion α of pixel p can be modelled using normal distribution of mean 1 and variance a_p^2, $\alpha \sim N(1, a_p^2)$. The distribution f^c of the chromaticity distortion γ at pixel p can be approximated by one-sided normal distribution of mean 0 and variance ${b_p}^2$,

$$f^c(\gamma|p) = \frac{2}{\sqrt{2\pi}b_p}\exp(-\gamma^2/(2{b_p}^2)), \qquad \gamma \geq 0. \tag{7}$$

Assuming that brightness distortion α_p and chromatic distortion γ_p are independent, the *naive* background probability density function f_B of pixel p can be given by

$$f_B(C_p|p) = f^b(\alpha_p|p)f^c(\gamma_p|p)\,, \tag{8}$$

where C_p is the RGB color vector of pixel p, and α_p and γ_p are calculated as in [6]. Figure 4 shows that the brightness and chromaticity distortions complement each other, thus our independence assumption can be empirically supported. The pixel foreground likelihood l of pixel p is given by

$$l(p) = -\log(f^b(\alpha_p|p)) - \eta\log(f^c(\gamma_p|p)), \tag{9}$$

where a constant η is introduced since the chromaticity distortion is relatively smaller than the brightness distortion in practice. Note that $l(p) = -\log(f_B(C_p|p))$ when $\eta = 1$.

3.2 Region Likelihoods

In every frame we perform image segmentation in order to partition each image into homogeneous small regions. We define $\mathcal{R} = \{R_i\}_{i\in I}$ as the set of the regions. For computational efficiency we perform it in the bounding box surrounding the foreground object. The bounding boxes are obtained by the maximal connected component of the foreground regions that are extracted by pixel-wise background subtraction. The foreground region likelihood $L(R_i)$ of region $R_i \in \mathcal{R}$ is calculated by

$$L(R_i) = \lambda_1 L_p(R_i) + \lambda_2 L_o(R_i). \tag{10}$$

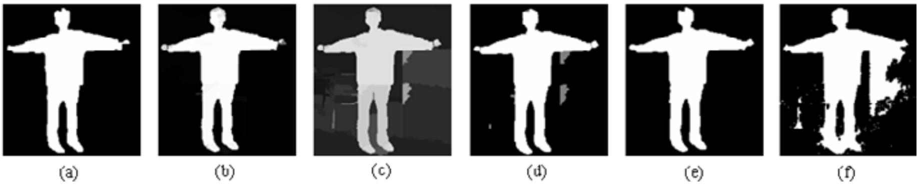

Fig. 6. Foreground segmentation with trimap estimation. (a) previous foreground object silhouette of the 580th frame, (b) regularization term $L_o(R_i)$ scaled by 255, (c) foreground region likelihood $L(R_i)$ truncated by 255. (d) the estimated trimap, (e) silhouette obtained by the proposed algorithm, (g) silhouette obtained by the Horprasert's algorithm[6]; deep shadow are on the floor and the teaching desk.

.

In the above equation, $L_p(R_i)$ is the *naive* foreground region likelihood given by the arithmetic mean of the pixel foreground likelihoods $l(p)$'s, i.e. $\sum_{p \in R_i} l(p)/n_{R_i}$, here n_{R_i} is the number of pixels in the region R_i. The naive region likelihood is not enough to decide the foreground object region especially when the occluding foreground parts are a similar color to the occluded background parts or when there are some deep shadows. Figure 5 shows an example of this. case. To overcome this problem we use the regularization term $L_o(R_i)$. This is the overlapping ratio of region R_i given by $n^o_{R_i}/n_{R_i}$, where $n^o_{R_i}$ is the number of pixels in region R_i that belong to the previous foreground object region. λ_2 is a regularization parameter. Figure 5 shows the pixel and naive foreground region likelihoods of an image. Some foreground regions with lower naive region likelihoods were supplemented by L_o in Fig. 6 (c). In the experiments we set $\lambda_1 = 0.8$ and $\lambda_2 = 30$.

3.3 Trimap Decision

By using the region likelihoods $L(R_i)$ in (10), the trimap can be estimated by using the region decision function $f_D : \mathcal{R} \rightarrow \{-1, 0, 1\}$ defined by,

$$f_D(R_i) = \begin{cases} 1 & \text{if } L(R_i) > T_U \\ 0 & \text{if } L(R_i) < T_L \\ -1 & \text{otherwise.} \end{cases} \tag{11}$$

where the thresholds T_U and T_L are related to the level of confidence on the trimap information. We set $T_U = 170$ and $T_L = 100$ in the experiments. Furthermore we define $\mathcal{R}_B$ as a collection of the regions that touch the bounding box. The regions in $\mathcal{R}_B$ are assumed to belong to the background area but when some region R_js belong to $\mathcal{R}_B$ and $f_D(R_i) = 1$ the regions are labeled as unknown. Figure 5 (e) shows the regions in $\mathcal{R}_B$ and an example of the estimated trimap is shown in Fig. 6 (d).

4 Experimental Results

Performance of the proposed method was evaluated with respect to the ground-truth segmentation of every tenth frame in each of five 700-frame test sequences

Table 1. Segmentation Error(%). The proposed algorithm(GT) is compared with the Horprasert's algorithm(HP) and the normal Graph Cut algorithm(GC) using color likelihoods on the test sequences of $JH1$, $JH2$, $JH3$, $GT1$ and $KC1$.

	$JH1$	$JH2$	$JH3$	$GT1$	$KC1$
GT	0.47	0.45	0.98	0.74	1.58
HP	2.59	5.22	1.67	1.66	3.01
GC	10.86	12.90	12.89	16.91	23.91

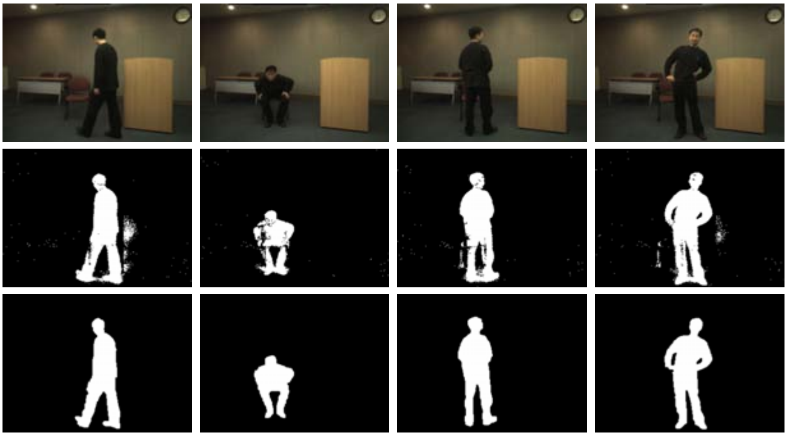

Fig. 7. Comparison of extracted human silhouettes. (Top) Four frames of the test sequence $JH3$. (Middle) Foreground extraction using HP[6]. (Bottom) Foreground extraction using the proposed algorithm.

of 320×240 images. The ground-truth data was labeled manually. Each pixel was labeled as foreground, background or unknown. The unknown label occurred in one plus pixel and one minus pixel along the ground-truth foreground object boundaries The five test sequences were labeled by $JH1$, $JH2$, $JH3$, $GT1$, $KC1$. Each was composed of different backgrounds and people. In all sequences, one person entered the scene, moved around and assumed natural poses, and the whole body was shown for gesture recognition. A person is shown under a light in *KC1* so that deep shadows are cast over the floor and wall.

Segmentation performance of the proposed method(GT) was compared with that of the Horprasert's background subtraction method(HP) [6] and the video version of the GrabCut method(GC) [12]. The only difference between GT and GC is that GT used a trimap but GC did not. Table 1 shows that the proposed method outperformed the both methods. The error rate was calculated within the bounding boxes only ignoring the mixed pixels. In the experiments we used ten Gaussian mixtures for color likelihood models and the mean shift segmentation method [3] was used to obtain the regions $\mathcal{R}^t$. Human silhouette results are

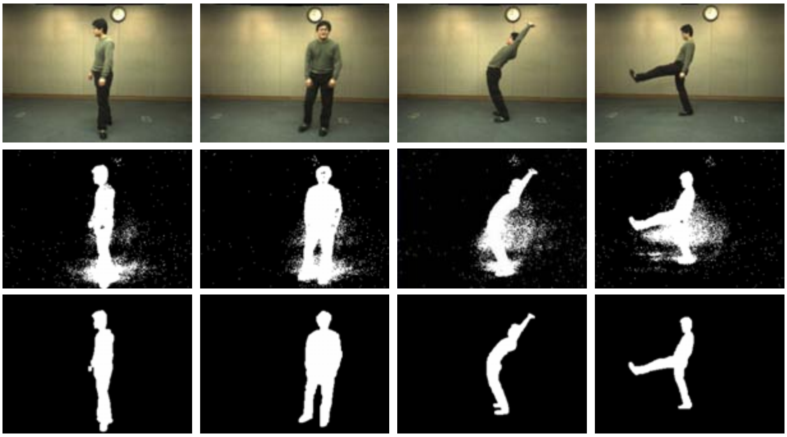

Fig. 8. Comparison of the silhouettes. (Top) Four frames of the test sequence $KC1$. (Middle) Foreground extraction using HP[6]. (Bottom) Foreground extraction using the proposed algorithm.

shown in Fig.7 and Fig. 8. The average running time of the proposed algorithm was 14.5 fps on a 2.8 GHz Pentium IV desktop machine with 1 GB RAM.

5 Conclusions

This paper has addressed accurate foreground extraction in video applications. We proposed a novel foreground segmentation method using a graph cut with an estimated trimap. We first estimated the trimap by partitioning the image into three regions: foreground, background and unknown regions. Then the trimap information was incorporated into the graph cut framework by distorting the foreground and background color likelihoods. The proposed algorithm showed good results on the real sequences and also worked at near real-time speed. However, about 60 percent of the processing time of the proposed algorithm was taken from the mean shift segmentation. In future work we are developing an efficient image segmentation method to find proper *automic* regions.

Acknowledgements. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019), and the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

References

1. Y. Boykov, and M. Jolly. Iterative graph cuts for optimal boundary and region segmentation of objects in N-D Images: In: Proc. IEEE Int. Conf. on Computer Vision, (2001) 105-112.
2. Y. Boykov and V. Kolmogorov. An experimental comparision of min-cut/max-flow algorithms for energy minimization in vision: IEEE Trans. on Pattern Anal.and Mach. Intell. 26(9) (2004) 1124-1137.
3. D. Comaniciu, P. Meer: Mean shift: a robust approach toward feature space analysis: IEEE Trans. Pattern Anal. Mach. Intell. 24(5) (2002) 603-619.
4. Y.-Y. Chuang, B. Curless, D. Salesin and R. Szeliski, A bayesian approach to digital matting: In: Proc. Int. Conf. Computer Vison and Pattern Recognition 2 (2001) 264-271.
5. A. Elgmmal, R. Duraiswami, L.S. Davis: Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking: IEEE Trans. Pattern Anal. Mach. Intell. 25(11) (2003) 1499-1504.
6. T. Horprasert, D. Harwood, L.S. Davis: A statistical approach for real-time robust background subtraction and shadow detection. In: Proc. IEEE Frame Rate Workshop (1999) 1-19.
7. M. Harville: A framework for high-level feedback to adaptive, per-pixel, mixture-of-Gaussian background models. In: Proc. European Conf. on Computer Vision (2002) 543-560.
8. V. Kolmogorov, A. Criminisi, A. Blake, G. Cross and C. Rother: Bi-layer segmentation of binocular stereo video. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, (2005).
9. Y. Li, J. Sun, C.-K. Tang and H.-Y. Shum: Lazy snapping. ACM Trans. Graphics 23(3) (2004) 303-308.
10. H. Li, M. Greenspan: Multi-scale gesture recognition from time-varying contours. In: Int. Conf. Computer Vision (2005) 236-243.
11. Z. Liu, S. Sarkar: Effect of silhouette quality on hard problems in gait recognition. IEEE Trans. Systems, Man, and Cybernetics-Part B:Cybernetics 35(2) (2005) 170-183.
12. C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts: In: ACM Trans. Graph, 23(3)(2004) 309-314.
13. A. Senior: Tracking people with probabilistic appearance models. In: Proc. IEEE Int. Workshop on PETS, (2002) 48-55.
14. C. Stauffer, W.E.L. Grimson: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 747-757.
15. J. Sun, J. Jia, C.-K. Tang and H.-Y. Shum, Poisson Matting: ACM Transaction on Graphics, 23(3) (2004) 315-321.
16. Y.-L. Tian, M. Lu, A. Hampapur: Robust and efficient foreground analysis for real-time video surveillance. In: Proc. Int. Conf. Computer Vision and Pattern Recognition (2005) 970-975.
17. Z. Tu: An integrated framework for image segmentation and perceptual grouping. In: Int. Conf. Computer Vision (2005) 670-677.
18. L. Wang, T. Tan, H. Ning, W. Hu: Silhouette analysis-based gait recognition for human identification. IEEE Trans. Pattern Anal. Mach. Intell. 25(12) (2003) 1505-1518.