Video Segmentation Using Iterated Graph Cuts Based on Spatio-temporal Volumes

Tomoyuki Nagahashi[1], Hironobu Fujiyoshi[1], and Takeo Kanade[2]

[1] Dept. of Computer Science, Chubu University.
Matsumoto 1200, Kasugai, Aichi, 487-8501 Japan.
nagahashi@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp
http://www.vision.cs.chubu.ac.jp
[2] The Robotics Institute, Carnegie Mellon University.
Pittsburgh, Pennsylvania, 15213-3890 USA.
tk@cs.cmu.edu

Abstract. We present a novel approach to segmenting video using iterated graph cuts based on spatio-temporal volumes. We use the mean shift clustering algorithm to build the spatio-temporal volumes with different bandwidths from the input video. We compute the prior probability obtained by the likelihood from a color histogram and a distance transform using the segmentation results from graph cuts in the previous process, and set the probability as the t-link of the graph for the next process. The proposed method can segment regions of an object with a stepwise process from global to local segmentation by iterating the graph-cuts process with mean shift clustering using a different bandwidth. It is possible to reduce the number of nodes and edges to about 1/25 compared to the conventional method with the same segmentation rate.

1 Introduction

The video segmentation that extracts object's region in a video sequence captured by a hand-held camera is a difficult problem. This technique is extremely important because it is often used in preprocessing for object recognition, and gesture recognition.

The interactive graph cuts proposed by Boykov *et al.* [1][2] haves been used in recent years for segmenting images. The energy function in interactive graph cuts is minimized by creating the graph from the correct-answer label and the input image that the user gave, and using a minimum cut/maximum flow algorithm. Nagahashi *et al* proposed image segmentation using iterated graph cuts based on multi-scale smoothing[3].

This segmentation of image based on graph cuts can be applied to video segmentation using the same framework. However, the size of the graph for a video sequence increases because we have to create the graph by making all pixels in the video. This causes, two main problems, i.e., we need large amounts of memory and it increases the computation cost. To overcome these problems, a graph constructed from spatio-temporal volumes has been used to reduce the

size of the graph[4][5]. However, it is difficult to precisely video segment video due to its low resolution.

We propose a method that represents spatio-temporal space as video that extends the technique of iterated graph cuts based on multi-scale smoothing[3] to spatio-temporal volumes obtain a stepwise process from global to local segmentation by iteration. Our approach uses mean shift clustering to build the spatio-temporal volumes with different bandwidths from the input video. We compute the prior probability obtained by the likelihood from a color histogram and a distance transform using the segmentation results from graph cuts in the previous process, and set the probability as the t-link of the graph for the next process. The proposed method can segment the regions of an object with a stepwise process from global to local segmentation by iterating the graph-cuts process with mean shift clustering using different bandwidth.

2 Graph Cuts for Video Segmentation

This section describes the graph-cuts-based segmentation proposed by Boykov and Jolly[1].

2.1 Graph Cuts for Image Segmentation

An image-segmentation problem can be posed as a binary-labeling problem. Let us assum that the image is a graph $G = (V, E)$, where V is the set of all nodes and E is the set of all arcs connecting adjacent nodes. The nodes are usually pixels p on the image P and the arcs have adjacency relationships with four or eight connections between neighboring pixels $q \in N$. The labeling problem is to assign a unique label L_i to each node $i \in V$, i.e., $L_i \in \{$"*obj*", "*bkg*"$\}$. The solution $\boldsymbol{L} = \{L_1, L_2, \ldots, L_p, \ldots, L_{|P|}\}$ can be obtained by minimizing the Gibbs energy $E(\boldsymbol{L})$:

$$E(\boldsymbol{L}) = \lambda \cdot \sum_{p \in P} R_p(L_p) + \sum_{\{p,q\} \in N} B_{\{p,q\}} \cdot \delta(L_p, L_q) \tag{1}$$

where

$$\delta(L_p, L_q) = \begin{cases} 1 \text{ if } \quad L_p \neq L_q \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

The coefficient, $\lambda \geq 0$, in Eq. (1) specifies the relative importance of the region properties term $R_p(L_p)$ versus the boundary properties term $B_{\{p,q\}}$. Regional term assumes that the individual penalties for assigning pixel p to "*obj*" and "*bkg*", corresponding to $R_p($"*obj*"$)$ and $R_p($"*bkg*"$)$ are given. For example, $R_p(\cdot)$ may reflect how the intensity of pixel p fits into a known intensity model (e.g., a histogram) of the object and background. Term $B_{\{p,q\}}$ comprises the "boundary" properties of segmentation $\boldsymbol{L}$. Coefficient $B_{\{p,q\}} \geq 0$ should be interpreted as a penalty for the discontinuity between p and q. $B_{\{p,q\}}$ is normally large when

pixels p and q are similar (e.g., in intensity) and $B_{\{p,q\}}$ is close to zero when these two differ greatly. The penalty $B_{\{p,q\}}$ can also decrease as a function of distance between p and q. Costs $B_{\{p,q\}}$ may be based on the local intensity gradient, Laplacian zero-crossing, gradient direction, or other criteria.

Table 1 lists the edge cost of the graph. The regional and boundary terms in

Table 1. Edge cost.

Edge		Cost	For
n-link	$\{p,q\}$	$B_{\{p,q\}}$	$\{p,q\} \in N$
t-link	$\{p,S\}$	$\lambda \cdot R_p(\text{"}bkg\text{"})$	$p \in P, p \notin \mathcal{O} \cup \mathcal{B}$
		K	$p \in \mathcal{O}$
		0	$p \in \mathcal{B}$
	$\{p,T\}$	$\lambda \cdot R_p(\text{"}obj\text{"})$	$p \in P, p \notin \mathcal{O} \cup \mathcal{B}$
		0	$p \in \mathcal{O}$
		K	$p \in \mathcal{B}$

Table 1 are calculated by

$$R_p(\text{“}obj\text{”}) = -\ln \Pr(I_p|\mathcal{O}) \tag{3}$$

$$R_p(\text{“}bkg\text{”}) = -\ln \Pr(I_p|\mathcal{B}) \tag{4}$$

$$B_{\{p,q\}} \propto \exp\left(-\frac{(I_p - I_q)^2}{2\sigma^2}\right) \cdot \frac{1}{dist(p,q)} \tag{5}$$

$$K = 1 + \max_{p \in P} \sum_{q:\{p,q\} \in N} B_{\{p,q\}}. \tag{6}$$

Let $\mathcal{O}$ and $\mathcal{B}$ define the “object” and “background” seeds. The seeds are given by the user. The boundary between the object and the background is segmented by finding the minimum cost cut [6] on the graph, G.

2.2 Problems with Conventional Method

Interactive Graph Cuts [2] create a graph from video.Thus, the size of the graph from video increases when placing individual pixels into a node. For example, the total number of the edges will be 25 million, when the input video is 360×240 with 100 frames. Therefore, we need copious amounts of memory and it takes a long time for processing with the minimum cut/maximum flow algorithm. One common technique of solving such this problem is to reduce the size of the graph by using a spatio-temporal volume. However, it is difficult to precisely segment regions and boundaries because segmentation using spatio-temporal volumes has low resolution. To overcome this problem, we propose a method that represents spatio-temporal space as video that extends the technique of iterated graph cuts based on multi-scale smoothing[3] to spatio-temporal volumes to obtain a stepwise process from global to local segmentation by iteration.

3 Iterated Graph Cuts Using Spatio-temporal Volumes

3.1 Proposed Method

We extend the technique of iterated graph cuts based on multi-scale smoothing[3] to spatio-temporal volumes.

Objects that move fast may be divided into different volumes between frames in a row when using spatio-temporal volumes. Therefore, it is difficult to create an optimal graph by only using adjoining volumes. To solve these problems, we introduced two kinds of edges, i.e., a volume that adjoins as an n-link, and a volume obtained from a search for corresponding points between frames.

Energy Function A volume pair that adjoins as the n-link, and a volume pair obtained by searching for the corresponding points between frames are used in the proposed method. Therefore, we extend the energy function using the graph cuts discussed Section 2.1 as follows:

$$E(\boldsymbol{L}) = \lambda \cdot \sum_{p \in P} R_p(L_p) + \sum_{\{p,q\} \in N} B_{N\{p,q\}} \cdot \delta(L_p, L_q) + \sum_{\{p,q\} \in C} B_{C\{p,q\}} \cdot \delta(L_p, L_q) \quad (7)$$

where $p, q \in P$ is a spatio-temporal volume, N is a neighboring volume of p and C represents corresponding points between frames. By using $B_{C\{p,q\}}$ in the energy function, we obtain robust segmentation results even if divided into different volume between frames.

Flow of Proposed Method Figure 1 shows the flow for of the new approach. First, the seeds, "foreground" and "background", are given by the user. Next, we

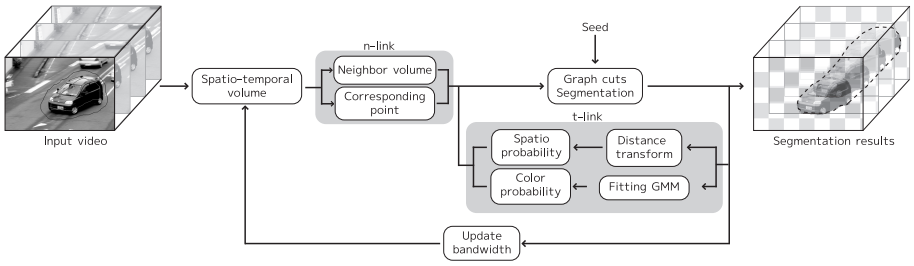

Fig. 1. Overview of proposed method.

obtain the spatio-temporal volume using mean shift clustering using bandwidth h. Graph cuts are done to segment the video into an object or a background. The Gaussian Mixture Model (GMM) is then used to make a color distribution model for the object and background classes from the segmentation results obtained from the graph cuts. The prior probability is updated from the distance

transform by the object and background classes of GMM. The t-links for the next graph-cuts process are calculated as a posterior probability which is computed a prior probability and GMMs, and h is updated as, $h = \alpha \cdot h$. These processes are repeated until $h < th$.

The processes are as follows.

Step 1 Input seeds
Step 2 Create spatio-temporal volume
Step 3 Search corresponding points
Step 4 Do graph cuts
Step 5 Calculate the posterior probability from the segmentation results and set as the t-link
Step 6 Update $h = \alpha \cdot h$, and Steps 1-5 are repeated until $h < th$.

The details of each process are given in what follows.

3.2 Spatio-temporal Volume

We employ mean shift clustering[7] to obtain the spatio-temporal volume. Let the space, time, and color information vector denote $\mathbf{x}_i = \{\mathbf{x}_i^s, \mathbf{x}_i^t, \mathbf{x}_i^r\}$, the filtering result denote $\mathbf{z}_i$, and each label denote L_i. $\{\mathbf{y}_j\}_{j=1,2,\ldots}$ is defined as

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)} \tag{8}$$

$$g(\mathbf{x}) = \frac{C}{h_s^2 h_t h_r^p} k\left(\left\|\frac{\mathbf{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathbf{x}^t}{h_t}\right\|^2\right) k\left(\left\|\frac{\mathbf{x}^r}{h_r}\right\|^2\right), \tag{9}$$

where h_s, h_t, h_r is the bandwidth by space, time, and color, C is the normalizing constant, $k(\mathbf{x})$ is the kernel function (e.g., Gaussian distribution). Mean shift clustering involves main four steps and an optional one.

1. Initialize $y_{i,j} = x_i$
2. Compute $y_{i,j+1}$, $k \leftarrow k + 1$ until convergence $\mathbf{z}_i = \mathbf{y}_{i,c}$ is reached.
3. Identify clusters $\{C_p\}_{p=1,\ldots,m}$ of convergence points by linking together all $\mathbf{z}_i$ that are closer than 0: 5 from one an other in the joint domain.
4. $L_i = \{p | \mathbf{z}_i \in C_p\}$
5. Optional: Eliminate spatial regions smaller than M pixels.

Figure 2 shows examples of spatio-temporal volumes with different bandwidths. In Fig. 2, we can see that each volume is decreased by decreasing the bandwidth. We represent global and local information using a spatio-temporal volume with different bandwidths in the proposed method. Then, a graph is created from nodes that correspond to spatio-temporal volumes segmented by mean shift clustering.

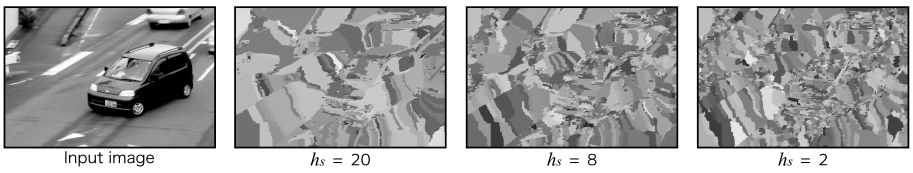

Fig. 2. Examples of Spatio-temporal volumes.

3.3 Add Edges Using Corresponding Points

Objects that move fast may be divided into different volumes between frames when using spatio-temporal volumes. Figure 3 shows an example of adding an edge using corresponding points. The edge has not been calculated because two the volumes are not neighbors. In our approach, we add an edge from the corresponding points. The corresponding points are computed by matching keypoints using SIFT [8] in two frames. This helps to correct two volumes, that are not in the neighborhood, corresponding to the same object. Consequently, volumes that are not in the neighborhood are represented as the same object.

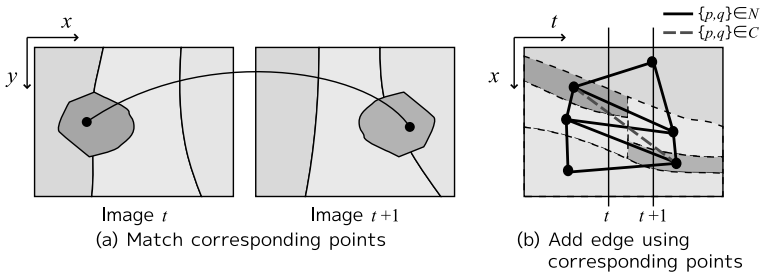

Fig. 3. Example of adding edge using corresponding point.

3.4 Iterated Graph Cuts

We have discussed segmenting of video using graph cuts using a spatio-temporal volume that is created from video using mean shift clustering and employing iteration from large to small bandwidths. We will not describe the method of updating the n- and t-links, and the effect of iterated processing.

Update n-link The n-link represents information between neighboring nodes. The volume pair that adjoins n-link $B_N(\boldsymbol{L})$, and the volume pair obtained by

searching for corresponding points between frames $B_C(\boldsymbol{L})$ are used in the new method. $B_N(\boldsymbol{L})$, $B_C(\boldsymbol{L})$ is given by

$$B_{\{p,q\}} = \exp\left(-\frac{\|\boldsymbol{I}_p - \boldsymbol{I}_q\|^2}{2\sigma^2}\right), \tag{10}$$

where I_p is the color in volume p.

Update t-link We compute the prior probability obtained by the likelihood from a color histogram and a distance transform using the segmentation results from the graph cuts in the previous process, and set the probability as the t-link using

$$R'_p(\text{“}obj\text{”}) = -\ln \Pr(\mathcal{O}|I_p) \tag{11}$$
$$R'_p(\text{“}bkg\text{”}) = -\ln \Pr(\mathcal{B}|I_p) \tag{12}$$

where $\Pr(\mathcal{O}|I_p)$ and $\Pr(\mathcal{B}|I_p)$ are given by

$$\Pr(\mathcal{O}|I_p) = \frac{\Pr(\mathcal{O})\Pr(I_p|\mathcal{O})}{\Pr(I_p)} \tag{13}$$
$$\Pr(\mathcal{B}|I_p) = \frac{\Pr(\mathcal{B})\Pr(I_p|\mathcal{B})}{\Pr(I_p)}. \tag{14}$$

$\Pr(I_p|\mathcal{O})$ and $\Pr(I_p|\mathcal{B})$ are the computed color probabilities and $\Pr(\mathcal{O})$ and $\Pr(\mathcal{B})$ are computed spatial probabilities from the segmentation results using graph cuts in the previous process.

Updating color probability The color probabilities $\Pr(I_p|\mathcal{O})$ and $\Pr(I_p|\mathcal{B})$ are computed by using GMM [9]. The GMM for the RGB color space is obtained by

$$\Pr(I_p|\cdot) = \sum_{i=1}^{K} \alpha_i p_i(I_p|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{15}$$

where $p_i(I_p|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is gaussian distribution. We used the EM algorithm to fit the GMM [10].

Updating spatial probability The spatial probabilities $\Pr(\mathcal{O})$ and $\Pr(\mathcal{B})$ are updated by spatial information from the graph cuts in the previous process. The next segmentation label is uncertain in the vicinity of the boundary. Therefore, the spatial probability is updated by using the results of a distance transform[11]. The distance from the boundary is normalized from 0.5 to 1. Let d_{obj} denote the distance transform of the object, and d_{bkg} denote the distance transform of the background. The prior probability is given by

$$\Pr(\mathcal{O}) = \begin{cases} d_{obj} & \text{if } d_{obj} \geq d_{bkg} \\ 1 - d_{bkg} & \text{if } d_{obj} < d_{bkg} \end{cases} \tag{16}$$
$$\Pr(\mathcal{B}) = 1 - \Pr(\mathcal{O}). \tag{17}$$

Color probability can be spatially controled using spatial probability. Consequently, the segmentation that is observed in the boundary possible in the next graph cuts. Therefore, we can obtain more robust segmentation even if the video contaions the same objects.

Iteration Finally, using $\Pr(I_p|\mathcal{O})$ and $\Pr(I_p|\mathcal{B})$ from GMM, and $\Pr(\mathcal{O})$ and $\Pr(\mathcal{B})$ from the distance transform, posterior probability can be computed by means of Eqs. (11) and (12). We compute the prior probability obtained by the likelihood from a color histogram and the distance transform, and set the probability as the t-link of the graph for the next process using the segmentation results obtained by using the graph cuts in the previous process.

4 Experimental Results

4.1 Experiment Outline

We used 13 videos including those of a vehicle moving, a human walking, a flower, and a leaf captured with a hand-held camera outdoors. A seed was only given to the first frame. We evaluated the segmentation results for the 10th frame comparing them with those from a manually correct mask. We defined a true positive (TP) as the number of objects of correct detection pixels, a false positive (FP) as the number of backgrounds of missed detection pixels, and a false negative (FN) as the number of objects of missed detection pixels. We evaluated the recall , precision, and F-measure as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{F} - \text{measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{20}$$

We compared three conventional methods and two methods we propose.

Conventional method 1 This involves Boykov's graph cuts approach [2]. Each pixel is a node obtained by using a graph.

Conventional method 2 This uses the spatio-temporal volume.

Conventional method 3 This involves iterating segmentation such as GrabCut [12] with a spatio-temporal volume .

Proposed method 1 Our approach involves iterating segmentation with a spatio-temporal volume using different bandwidths. However, we did not use spatial probability.

Proposed method 2 Our approach involves iterating segmentation with a spatio-temporal volume using different bandwidths with spatial probability.

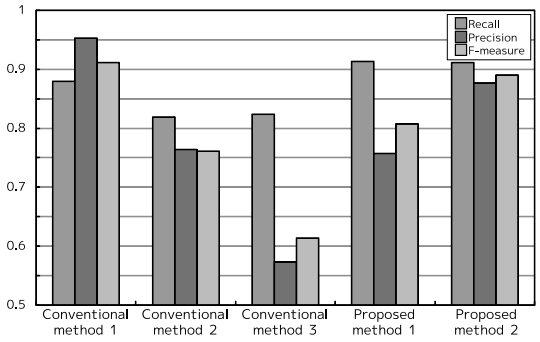

Fig. 4. Segmentation rate.

4.2 Comparison with Conventional Method

Figure 4 is a bar chart with the segmentation rate and Fig. 5 shows example segmentation results with three of the methods.

Effect Using Spatio-temporal Volume Conventional method 1 in FIg.4 can obtain better segmentation than Conventional method 2 whose using spatio-temporal volume has lower resolution than that of the formaer. Therefore, poor segmentation is obtained with Conventional method 2.

Effect Iterating Process We compared Conventional method 2 with Conventional method 3, which had repetition processing added. The recall was same rate, but precision with Conventional method 3 was lower than with Conventional method 2. It was difficult to detect the background when a spatio-temporal volume was used (see Fig. 4, Conventionals method 1 and 2). Therefore, the background color was learned as an object color in the iterating process. Figure 5(c)(d) shows the segmentation results for Conventional methods 2 and 3. We can see that the false detection of the background has gradually been extended by the iterating process.

Effect of Iterating Process by Changing Bandwidth Proposed method 1 is better at segmentation, its recall is better at 0.09, its precision is better at 0.18, and its f-measure is better at 0.19, than those of Conventional method 3. Figure 5(e) shows the segmentation results for the iterating process obtained by changing the bandwidth. We can see that Proposed method 1 can reduce missed detection in the background. Figure 5(d) shows Conventional method 3 detects many incorrect small volumes in the background because the color looks like the object. When the bandwidth in mean shift clustering is large, the spatio-temporal volume is large as shown in Fig. 5(b). Although we obtained coarse segmentation results, these were not incorrect volumes. By changing the

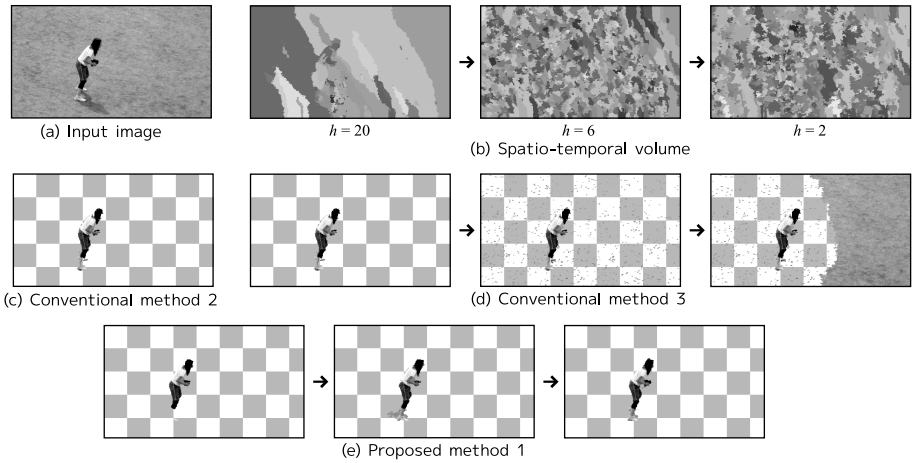

Fig. 5. Example segmentation results.

bandwidth, we could obtain more precise segmentation like that in coarse-to-fine approach.

Effect of Spatial Probability by Distance Transform Proposed method 2 using spatial probability has better Precision at 0.12 than Proposed method 1. Figure 6 shows the segmentation results in a sequence that has the same object. Proposed method 1 that only uses color probability cannot segment correctly, e.g., it detects the leaf, which has not been specified. However, we can see that Proposed method 2 can detect the leaf, which has been specified.

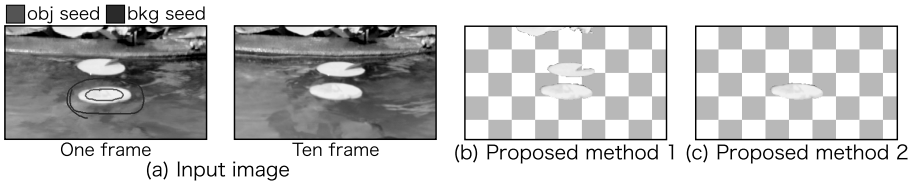

Fig. 6. Example segmentation results using distance transform.

Overall, Proposed method 2 using spatio-temporal volumes could obtain a segmentation rate comparable to that of Conventional method 1.

4.3 Comparison of Graph Size

Table 2 lists the graph size with each method. The bandwidth of Conventional

Table 2. Graph size.

	Conventional method 1[2]	Conventional method 2, 3	Proposed method 1, 2
Node	864,000	52,993	43,140 - 52,993
Edge	2,505,600	81,239	17,477 - 81,549

methods is $h = 2$, and the results by using Proposed method were obtained by changing bandwidth h from 20 to 2. Compared with Conventional method 1, the proposal technique was able to reduce the number of edges about 6.1% and the number of nodes to about 3.3%. It was possible to reduce the number of nodes and edges to about 1/25 compared to the conventional method with the same segmentation rate.

4.4 Effect of Adding Edge Using Corresponding Points

We evaluated how effective it was to add edges using corresponding points. It is difficult to segment objects with Conventional method when they moves fast. We used video at 6 fps in this experiment. We compared Proposed method where edges were added using corresponding points and Conventional method where edges were not added using corresponding points. The proposed method

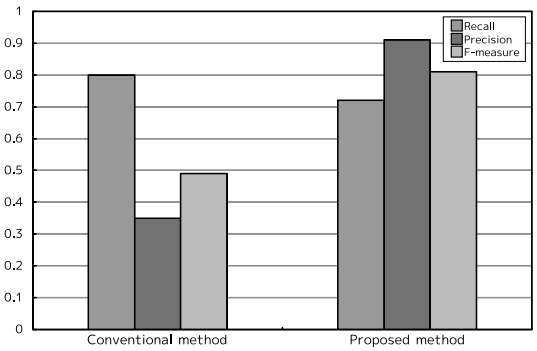

Fig. 7. Segmentation results.

could obtain better segmentation than the conventional method. Fewer errors were detected because corresponding points were matched between volumes that were not neighbors by frame.

5 Conclusion

We presented a novel approach to video segmentation using iterated graph cuts based on spatio-temporal volumes. We used the mean shift clustering algorithm to build the spatio-temporal volumes with different bandwidths from the input video. We computed the prior probability obtained by the likelihood from the color probability and the spatial probability using the segmentation results from graph cuts in the previous process, and set the probability as the t-link of the graph for the next process. It is possible to reduce the number of nodes and edges to about 1/25 comparing to the conventional method with the same segmentation rate.

We would like to investigate features other than color in the future. In addition, we would like to accelerate segmentation processing.

References

1. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. ICCV2001 **01** (2001) 105
2. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. Int. J. Comput. Vision **70**(2) (2006) 109–131
3. Nagahashi, T., Fujiyoshi, H., Kanade, T.: Image segmentation using iterated graph cuts based on multi-scale smoothing. In: ACCV 2007, Part II, LNCS 4844. (2007) pp. 806–816
4. Wang, J., Bhat, P., Colburn, R.A., Agrawala, M., Cohen, M.F.: Interactive video cutout. In: SIGGRAPH '05: ACM SIGGRAPH 2005 Papers, New York, NY, USA, ACM Press (2005) 585–594
5. Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. In: SIGGRAPH '05: ACM SIGGRAPH 2005 Papers, New York, NY, USA, ACM Press (2005) 595–600
6. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9) (2004) 1124–1137
7. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5) (2002) 603–619
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2) (2004) 91–110
9. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition (CVPR-99), Los Alamitos (1999) 246–252
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1) (1977) 1–38
11. Toyofumi, S., Junichiro, T.: Euclidean distance transformation for three dimensional digital images. The transactions of the Institute of Electronics, Information and Communication Engineers **76**(3) (1993) 445–453
12. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3) (2004) 309–314