Bulletin of the College of Engineering, N.T.U., No. 90, February 2004, pp. 25–34
國立臺灣大學「台大工程」學刊　第九十期　民國九十三年二月　第 25–34 頁

COARSE-TO-FINE VIDEO OBJECT SEGMENTATION BY MAP LABELING OF WATERSHED REGIONS

階層式視訊物件分割方法—
使用分水嶺區域的 MAP 標記技術

Ping-Che Chen[*]　*Jin-Jen Su*[*]　*Yu-Pao Tsai*[†]　*Yi-Ping Hung*[‡]
陳秉哲[*]　蘇敬仁[*]　蔡玉寶[†]　洪一平[‡]

[*]Master　[†]Research Assistant　[‡] Professor
[*‡] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617, R.O.C.
[†]Institute of Information Science, Academia Sinica, Taipei County, Taiwan 115, R.O.C.
[*]碩士　[†]研究助理　[‡] 教授
[*‡] 國立台灣大學資訊工程學系
[†]中央研究院資訊所

摘　要

視訊物件分割常被視一種圖的標記問題。本篇論文中，我們採用階層式的機制，從低解析度到高解析度循序以最大事後機率 (MAP) 的方法，將最佳的標記指定給視訊中的每一個分水嶺區域。我們的方法分爲兩個階段：(1) 預備處理階段以及 (2) 標記階段。在預備處理階段中，我們先將視訊中的每一張影像分割成不同解析度的分水嶺區域，然後估算每一個分水嶺區域的移動向量。利用估算出的移動向量，接著對每一階層中所有分水嶺區域建立一個空間時間區域相鄰圖 (ST-RAG)，每一個空間時間區域相鄰圖中的節點，都代表一個分水嶺區域，當兩個分水嶺區域是時間或空間上的鄰居時，他們所對應的節點間，就會有一個邊將他們連結起來。在標記階段中，最大事後機率的標記程序會由最低解析度的空間時間區域相鄰圖開始，並延伸至高解析度。我們用高斯分佈來表示最大事後機率的觀察項，此一高斯分佈的機率則統計其相接鄰的分水嶺區域之運動向量而得到。最大事後機率的事前資訊項則是以馬可夫自由場來表示。我們的實驗顯示本論文的方法可以有效且精確地得到良好的分割結果。

關鍵詞：視訊物件、影像分割、貝氏法、馬可夫隨機場、分水嶺演算法。

Abstract

In this paper, we formulate video object segmentation as a graph-labeling problem and solve it with the MAP approach through a coarse-to-fine scheme based on multi-scale watershed segmentation. Our method is divided into two stages: the preprocessing stage and the labeling stage. In the preprocessing stage, we first partition each image frame into a set of multi-scale watershed regions and then estimate the motion vectors for each watershed region. Next, based on the estimated motion vectors, all the watershed regions at each resolution scale are used to build one Spatio-Temporal Region Adjacency Graph (ST-RAG), one for each resolution scale. Here, each node of a ST-RAG corresponds to one and only one watershed region, and two nodes are connected by an edge if their corresponding regions are either spatial or temporal neighbors. In the labeling stage, the MAP labeling process starts from the ST-RAG of the coarsest scale, and gradually proceeds to the finer scale. The observation term is modeled as a Gaussian distribution, learned from the estimated motion vectors of neighboring regions, and the prior term modeled as a Markov random field with some skillfully designed potential. Our experiments showed that the proposed method could efficiently produce accurate results of video object segmentation.

Keywords: video object, image segmentation, Bayesian approach, Markov random field, watershed algorithm.

1. INTRODUCTION

Video object is an important primitive in MPEG-4 standard, and can provide more flexible manipulations. For example, video objects extracted from different video sources can be synthesized into one new video signal. Another application of video objects is in augmented reality, where the augmented video objects serve as dynamic textures or active components in a virtual world for enhancing user perception.

Many methods have been proposed for video object segmentation. Among them, some are unsupervised [1~4]. However, unsupervised methods may not be able to always extract the exact video object as desired. To solve this problem, some researchers propose semi-automatic methods by allowing user interaction to guide the segmentation [1,5~8]. For example, the method proposed in [8] use modified Intelligent Scissors to segment an initial frame and apply automatic tracking to extract the video object in the subsequent frames. This is a typical "initial single-frame segmentation + tracking" method. While this type of method is quite popular, most researchers use only local temporal information to track the video object frame by frame. Therefore, the quality of segmentation result strongly relies on the success of frame-by-frame tracking, which can fail occasionally due to poor image data.

Instead of tracking the video object frame by frame, we formulate video object segmentation as a graph-labeling problem, and find the globally optimal solution by using the maximum *a posteriori* probability (MAP) approach. The work most related to ours is the one done by Patras, *et al*. [2]. They first partition each image frame into a set of watershed regions (WRs) and build a spatial region adjacency graph (S-RAG) based on these WRs, and then find the optimal labeling by using the MAP approach, frame by frame. By contrast, we first partition each image frame into a set of multi-scale watershed regions (MSWRs) and then utilize the motion information of MSWRs to build a spatio-temporal region adjacency graph (ST-RAG), one for each scale. Here, each node of a ST-RAG corresponds to one and only one watershed region, and two nodes are connected by an edge if their corresponding regions are either spatial or temporal neighbors. Our method has two major contributions. First, we use ST-RAGs (instead of using S-RAGs) to achieve the globally optimal labeling, with which we can consider the labeling of WRs in all frames simultaneously. Second, by adopting the coarse-to-fine scheme, we can reduce the computational time for MAP labeling.

As shown in Fig. 1, our method consists of two stages: the preprocessing stage and the labeling stage. The preprocessing stage includes three operations: (i) constructing MSWRs; (ii) estimating motion vectors of every MSWR; (iii) building a set of ST-RAGs from the coarsest scale to the finest scale. We will describe the preprocessing stage in more details in Section 2. In the labeling stage, the MAP labeling process starts from the ST-RAG of the coarsest scale, and gradually proceeds to the finer scale. Here we introduce an "uncertain region" label that indicates, up to the current scale, a region can neither be classified into the desired video object nor into the background, and hence needs to be determined at a finer scale. We will describe the initial labeling of the ST-RAG in Section 3 and our method for MAP labeling in Section 4. At the finest scale, if the segmented result produced by MAP labeling is not consistent with user's expectations, the user can adjust the segmentation result in one or more frames and restart the process of MAP labeling to propagate the modified result to all frames. Here, user adjustment can be performed by using any interactive segmentation method, and in our current implementation we adopt a modified intelligent scissors [9]. In this paper, we show some experimental results in Section 5, and give a conclusion in Section 6.

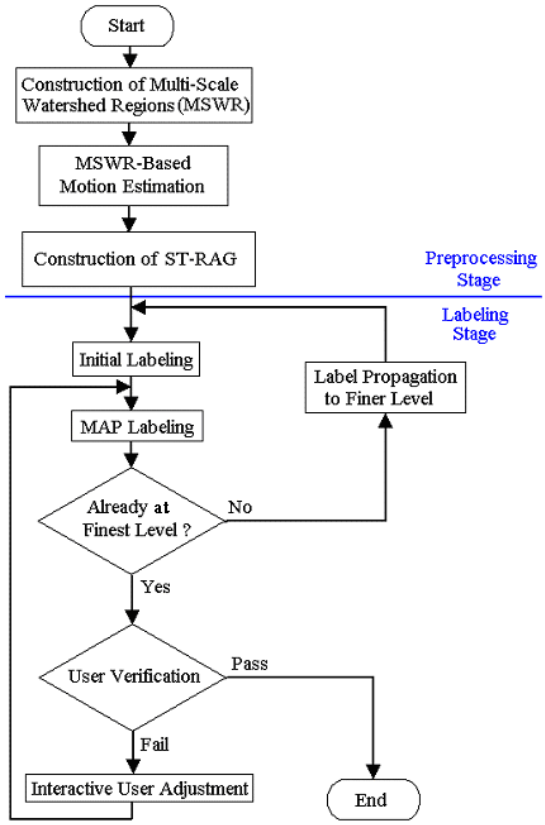


Fig. 1 Workflow of the proposed video object segmentation method

2. PREPROCESSING STAGE

In this stage, we first partition each frame in a video sequence into a set of multi-scale watershed regions by using the watershed segmentation algorithm. Next, we estimate the motion vectors of each watershed region using WR-based template matching. The estimated motion vectors will be used to construct ST-RAGs in Section 2.3 and be used as observation features in the labeling stage.

2.1 Construction of Multi-Scale Watershed Regions

An image can be partitioned into a set of regions by using the watershed segmentation algorithm [10]. However, the basic watershed algorithm tends to produce over-segmentation due to noise or local irregularities in the gradient image. Since overly segmented regions may not be reliable enough for motion estimation in the next step, we apply geodesic reconstruction to alleviate the over-segmentation problem [11]. Noticing that a watershed region at a coarser scale is obtained by merging some watershed regions at a finer scale, we can easily build a multi-scale structure from the finest scale to the coarsest scale.

2.2 MSWR-Based Motion Estimation

For every watershed region, we use the template-matching algorithm to estimate its motion vector. That is, we estimate the motion vector by minimizing the sum of absolute difference (SAD) between the template and the target. Here, the template window for motion estimation is determined by the dilation of the watershed region. The dilation uses a unit circle as its kernel. We use the sum of the gradient magnitude of all pixels within the template window to determine whether the template window is large enough. If the sum is smaller than a threshold, the template window is dilated by unit circle. This process is repeated until the sum is larger than the threshold.

If the average absolute difference (*i.e.*, the SAD divided by the number of pixels in the template window) of the target region is too large, the motion vector may be inaccurate and should not be used. This information of inaccuracy will be stored in the node corresponding to the WR when constructing ST-RAGs, which will be described in the next subsection. Notice that both forward motion vector and backward motion vector need to be estimated for each WR, except for the boundary frames, because motion vectors of both directions will be used in our method.

2.3 Construction of ST-RAGs

Let $G = (V, E)$ be a ST-RAG at a certain scale, where V is the set of the nodes corresponding to watershed regions, and E is the set of edges connecting either the spatial neighbors or the temporal neighbors. Here, two regions are considered to be spatial neighbors only if they are in the same frame and share a common boundary, while two regions are temporal neighbors only if they are in adjacent frames and become overlapped after projection with the corresponding motion vector. Figure 2 shows an example of two ST-RAGs constructed at two different scales. Each node of the graph has two flags: the flag "*ReliableMotion*" and the flag "*Fixed*". The flag "*ReliableMotion*" is set to be true only if the region has a reliable estimate of motion vector. This information will be used in MAP labeling, as described in Section 4. The flag "*Fixed*" is set to be true (either via initial labeling or via user adjustment) only if the region's label is not allowed to be changed afterwards during performing MAP labeling.

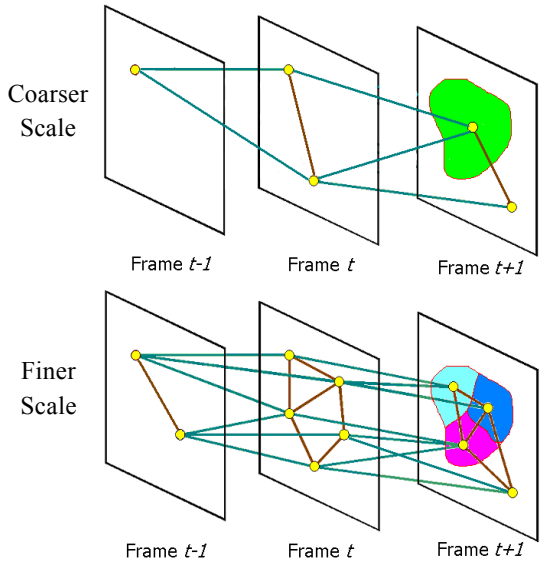


Fig. 2 An example of two ST-RAGs at two different scales

3. INITIAL LABELING

After preprocessing, the labeling stage starts with initial labeling of the graph. The label field K is defined as the following:

$$K = \{ \text{“}F\text{”}, \text{“}B\text{”}, \text{“}U\text{”} \} ,$$

where “*F*”, “*B*”, “*U*” represent foreground region, background region, and uncertain region, respectively.

A region is labeled as uncertain if, at the present time, the region cannot be decided either as a foreground region or as a background region. The initial labels of the ST-RAG at the coarsest scale can be obtained by label propagation from key frames, as described in Section 3.1. The initial labels of ST-RAGs at the finer scale are obtained by label propagation from the coarser scale, which will be described in Section 3.2.

3.1 Label Propagation from Key Frames

The initial label of the ST-RAG can be obtained by using segmentation results of a set of key frames, which can be generated either by an automatic video object segmentation algorithm [1] or by an interactive tool for single image segmentation [9] that is an extension of intelligent scissors [12]. After initial segmentation, all regions in the key frames are labeled to be "*F*", "*B*", or "*U*". If a region is labeled to be either "*F*" or "*B*", we set its flag "*Fixed*" to be true. If we use the automatic video object segmentation algorithm [1], all frames will be key frames and will have initial labeling. If we use the modified intelligent scissors to segment only a couple of key frames, the unlabeled regions in the remaining frames can be labeled by "label propagation from the key frames".

In order to determine the initial label of an unlabeled region, we define two variables l_b and l_f, where l_b denotes the initial label propagated from the nearest antecedent key frame and l_f denotes the initial label propagated from the nearest subsequent key frame. To determine l_b, we successively project all the pixels in the unlabeled region to the nearest antecedent key frame by using backward motion vectors. If most of the projected pixels are located at the regions labeled as "*F*" or "*B*", we set l_b to be "*F*" or "*B*". Otherwise, we set l_b to be "*U*". Here, "most" means the number of projected pixels having label "*F*" or "*B*" is larger than a threshold. Figure 3 shows an example of determining l_b while frame *t*-1 is an antecedent key frame of frame *t*. To determine l_f, we apply the same process of determining l_b except that we project the pixels using forward motion vectors instead of using backward motion vectors. Once l_b and l_f are determined, we set the initial label of the unlabeled region based on the values of l_b and l_f according to Table 1.

3.2 Label Propagation from Coarser Level

When the labeling procedure enters a finer scale from a coarser scale, the initial label of the ST-RAG at the finer scale can be determined by propagating labels from the coarser scale to save processing time. For example, if a watershed region is labeled as "*F*" and then splits into three smaller regions at the finer scale,

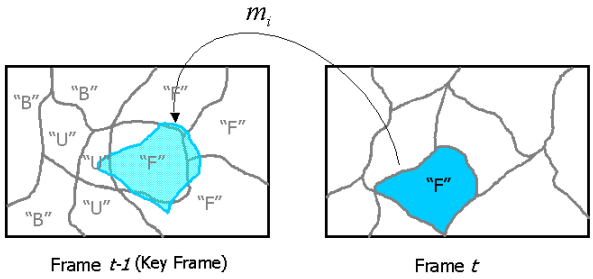


Fig. 3 An example of label propagation from the nearest antecedent key frame

Table 1 The value for initial labeling based on different values of l_b and l_f

l_b \ l_f	*F*	*U*	*B*
F	*F*	*F*	*U*
U	*F*	*U*	*B*
B	*U*	*B*	*B*

the corresponding nodes of these smaller regions will be still labeled as "*F*" and their flag "*Fixed*" set to be true. That is, a node labeled as "*F*" or "*B*" at the coarser scale cannot be changed during the MAP labeling process at the finer scale, unless purposely adjusted by the user.

4. MAP LABELING

After initial labeling, we find the optimal labeling of the ST-RAG by using maximum *a posteriori* probability (MAP) approach. Using Bayes rule, we have:

$$\max_L P(L \mid M) \propto \max_L P(M \mid L)\, P(L)$$

where $M = \{m_i \mid i \in V\}$ is the observed motion features, and $L = \{l_i \mid l_i \in K, i \in V\}$ is a label field on the graph. In this paper, we model the observation term by a Gaussian distribution and the prior term by a Markov random field (MRF).

4.1 Observation Term: Gaussian Model

Since the motion for each watershed region has been estimated, we can compute the observation term for each region by utilizing the motion information. If the motion of a region is reliable (flag "ReliableMotion" is true), we can compute its observation term by first modeling the neighboring regions' motion with a Gaussian distribution and then evaluating the probability of its motion vector in the distribution. If the motion of a region is not reliable (*i.e.*, flag "ReliableMotion" is

false), we simply ignore the observation term and completely rely on the prior term. Note that if one region has both backward and forward motion vector, we will construct two Gaussian distributions for both directions and get two probabilities according to its motion vectors. Furthermore, the observation term of the region is equivalent to the multiplication of the two probabilities. In order to estimate the Gaussian distribution for a given region *i*, we need to determine its neighboring regions. In our implementation, we apply the breath-first-search algorithm on the ST-RAG, starting from node *i*, in order to find a fixed number of neighbors (currently fixed to 30) with the same label as region *i*. To estimate the distribution more robustly, we weight the contribution of each chosen neighboring regions according to its area and the region-distance. The region-distance is the number of edges on the shortest path from node *i* to the chosen neighboring node.

If the region *i* is labeled as "*F*" or "*B*", the mean and the covariance of the Gaussian distribution can be estimated by Eq. (1). For each region, we should compute the probability of every possible label and determine which label has the maximum probability. As mentioned in Section 3, if a region is set to "*U*" it means the motion information cannot reliably determine the label of this region at this moment. In this case, we model the conditional probability by using a uniform distribution.

$$\mu_i(k) = \frac{\sum_{j \in S_i(k)} \omega_{ij} * m_j}{\sum_{j \in S_i(k)} \omega_{ij}},$$

$$\Sigma_i(k) = \frac{\sum_{j \in S_i(k)} \omega_{ij} * [m_j - \mu_i(k)][m_j - \mu_i(k)]^t}{\sum \omega_i}$$

$$\omega_{ij} = \frac{\text{Area}_j}{\text{BlockDistance}_{ij}} \tag{1}$$

$k \in \{"F","B"\}$

m_i : estimated motion vector of region i

$S_i(k)$: the set of neighboring regions of region i that has the same label k with region i

In summary, we use Eq. (2) to evaluate the observation term. Here, K_A and K_B are the normalization factors for label "*F*" and "*B*", and are dependent on their covariance. K_u is a user-adjustable factor for controlling the probability of occurrence of uncertain regions.

$$P\left(m_i \mid \{l_i = k\} \cup \bar{L}_i\right) = \begin{cases} \frac{1}{K_A} \exp\left[-\frac{1}{2}\left[m_i - \mu_i("F")\right]^t \Sigma_i^{-1}("F")\left[m_i - \mu_i("F")\right]\right]; & k = "F" \\ \frac{1}{K_B} \exp\left[-\frac{1}{2}\left[m_i - \mu_i("B")\right]^t \Sigma_i^{-1}("B")\left[m_i - \mu_i("B")\right]\right]; & k = "B" \\ \frac{1}{K_u}; & k = "U" \end{cases} \tag{2}$$

where $\bar{L}_i = L - \{l_i\}$ and K_F, K_B and K_U are normalization factors for label "*F*", "*B*" and "*U*".

4.2 Prior Term: MRF Model

We use a Markov random field defined on watershed regions to model the prior term as the following:

$$P(L) = \frac{e^{-\frac{1}{T}U(L)}}{Z} \tag{3}$$

where $U(L) = \sum_{i \in V} \sum_{j \in N_i} V_c(i, j, l_i, l_j)$ and Z is a normalization constant, T is the temperature that is proportional to the current scale. That is, the temperature is higher at coarser scale. $U(L)$ is the energy function which is the sum of clique potentials over all possible cliques C. For simplicity, we only use 2-cliques to define the energy function. Note that Z is a constant across all possible configurations, so there is no need to compute the value of Z.

Before defining our energy function, we first define similarity measures for two regions that are either spatially neighbors or temporally neighbors. Consider Fig. 4(a), we define spatial similarity measure (S_s) for two regions *a* and *b* by Eq. (4).

$$S_s = \begin{cases} H_1(2e^{-\delta H_2} - 1) & \text{where } v_a \text{ and } v_b \text{ are reliable} \\ H_1 & \text{otherwise} \end{cases}$$

$$H_1 = \max\left(\frac{l_{a,b}}{c_a}, \frac{l_{a,b}}{c_b}\right) \tag{4}$$

$$H_2 = \begin{cases} |v_a - v_b| & \text{where } v_a \text{ and } v_b \text{ are reliable} \\ \infty & \text{otherwise} \end{cases}$$

where δ is the factor for weighting the difference between motion vectors; *a* and *b* denote the two watershed regions that are spatial neighbors; $l_{a,b}$ is the length of the arc between *a* and *b*; c_i is the length of the contour of region *i*, $i = a, b$; and v_i is the motion vector of region *i*, $i = a, b$.

Consider Fig. 4(b). The temporal similarity measure (S_t) for two regions x and y is defined by Eq. (5):

$$S_t = \max\left(\frac{r_{x,y}}{\text{Area}_x}, \frac{r_{x,y}}{\text{Area}_y}\right) \tag{5}$$

where x and y denote two watershed regions that are temporal neighbors; $r_{x,y}$ is the overlapping area of x and y; and Area_i is the area of region i, $i = x, y$.

Based on the spatial and temporal similarity measures, Eqs. (6) and (7) define the prior energy functions for two regions i and j that are either spatial neighbors or temporal neighbors, respectively.

$$V_c(i, j, l_i, l_j) = \begin{cases} k_1 S_s^5 \; ; l_i = l_j \text{ and } l_i, l_j \in \{"F", "B"\} \\ -k_1 S_s^5 ; l_i \neq l_j \text{ and } l_i, l_j \in \{"F", "B"\} \\ k_1 \min\left(S_s^4 - k_a, -\frac{k_a}{2}\right); \text{otherwise} \end{cases} \tag{6}$$

k_a: factor for "U" labels
k_1: normalization factor

$$V_c(i, j, l_i, l_j) = \begin{cases} k_2 S_t^5 \; ; l_i = l_j \text{ and } l_i, l_j \in \{"F", "B"\} \\ -k_{21} S_t^5 ; l_i \neq l_j \text{ and } l_i, l_j \in \{"F", "B"\} \\ k_2 \min\left(S_t^4 - k_b, -\frac{k_b}{2}\right); \text{otherwise} \end{cases} \tag{7}$$

k_b: factor for "U" labels
k_2: normalization factor

Note that the value of the spatial similarity measure ranges between –1 and 1, and the value of the temporal similarity measure ranges between 0 and 1. For both measures, two similar regions have a larger value, and two dissimilar regions have a smaller value. A similarity value close to zero implies that it is not clear whether the two regions are similar or dissimilar.

The potential function should have a large negative value if the labeling for the two regions agrees with their similarity measure. In the contrary situation, the potential function should have a large positive value. For example, if two regions are both labeled as "F" and the similarity measure between them is large, the potential will be assigned a large negative value. If the similarity measure is close to zero, the potential function should have a relatively large negative value for uncertainty labeling. Figure 5 shows the potential functions for different combination of the value of l_i and l_j. To find the configuration of labeling by maximizing *a posteriori* probability, we use the modified ICM (Iterative Conditional Modes) algorithm [13].

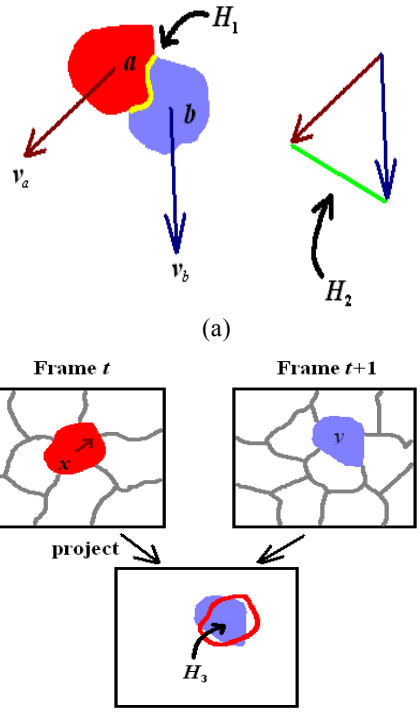


Fig. 4 Illustration for defining the similarity measures of two regions that are (a) the spatial neighborhoods and (b) the temporal neighborhoods

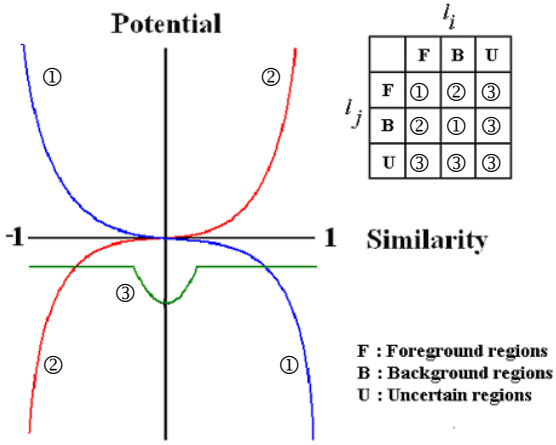


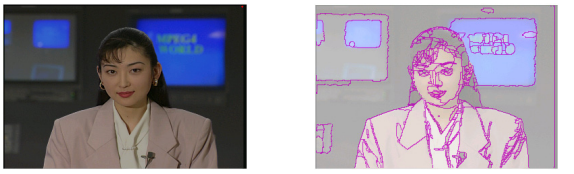
Fig. 5 The clique potential function

5. EXPERIMENTAL RESULTS

In this section, we show some results obtained by using the proposed method for video object segmentation. We choose three of the MPEG-4 validation sequences (the "Akiyo", the "Foreman" and the "Bream") to demonstrate our method.

Figure 6 shows an example of multi-scale watershed regions obtained by using different scaling parameters. Figure 6(a) is the first frame of the "Akiyo" sequence. Figures 6(b), 6(c), and 6(d) show the watershed regions generated at a coarser scale, a middle scale, and a finer scale, respectively.

Figure 7 shows the segmentation result of the "Akiyo" sequence. Figure 7(a) shows the original images of frame 30, 60, 90, 120, 150, and 180, and 7(b) shows the segmentation results corresponding to these original images. By using the proposed method, the obtained result looks very good even without any user adjustment, as shown in Fig. 7.



(a) Original image (b) WSs at a coarser scale

(c) WSs at a middle scale (d) WSs at a finer scale

Fig. 6 An example of multi-scale watershed regions

frame 30 frame 60 rame 90

frame 120 frame 150 frame 180

(a) Original images

frame 30 frame 60 frame 90

frame 120 frame 150 frame 180

(b) Segmentation results

Fig. 7 Segmentation results of the "Akiyo" sequence without any user adjustment

Figure 8 shows the segmentation result of the "Foreman" sequence. Figure 8(a) shows the original images of frame 30, 60, 90, 120, 150, and 180, and 8(b) shows the segmentation results corresponding to these original images. It is harder to automatically extract the foreground object from the "Foreman" sequence, because the difference in color between the foreman's helmet and the wall of building is small. Hence, some minor interactive adjustment was required for achieving a better result, which is shown in Fig. 8(b).

Figure 9 shows the segmentation results of the "Bream" sequence. It is difficult to separate the tail of the fish from the background, because the fish tail is somewhat transparent. Figure 9(a) shows the original images of frame 30, 60, 90, 120, 150, and 180, and 9(b) shows the segmentation result generated by the proposed method with some minor user adjustment.

frame 30 frame 60 frame 90

frame 120 frame 150 frame 180

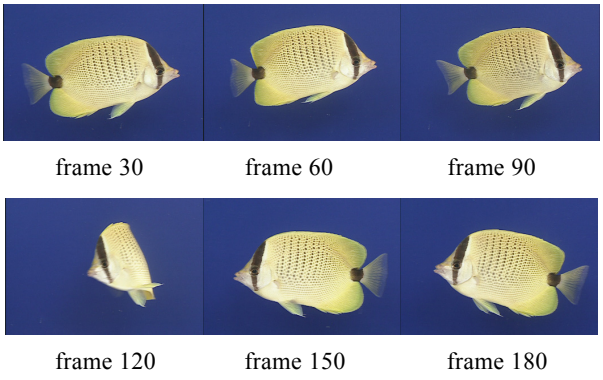
(a) Original images

frame 30 frame 60 frame 90

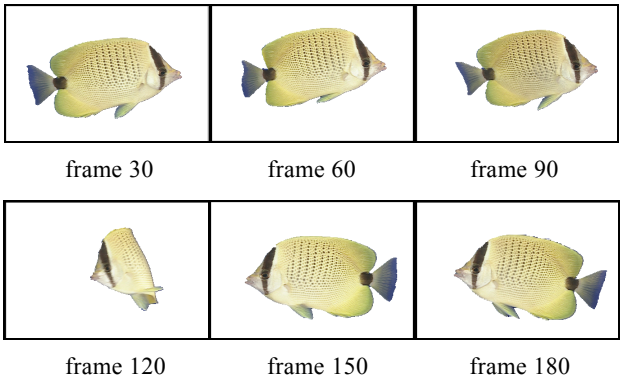
frame 120 frame 150 frame 180

(b) Segmentation results

Fig. 8 An example of segmentation result of the "Foreman" sequence



(a) Original images



(b) Segmentation results

Fig. 9 Segmentation result of the "Bream" sequence

6. CONCLUSION

In this paper, we formulate video object segmentation as a graph-labeling problem and solve it with the MAP approach through a coarse-to-fine scheme. Based on multi-scale watershed regions and their bi-directional motion estimates, we build one ST-RAG for each scale. Once the ST-RAG is constructed, the MAP approach is used to label each watershed region as foreground region, background region, or uncertain region. The method proposed in this paper has two major contributions. First, we introduce a new presentation, the ST-RAG, to globally model the spatial and temporal relationship between watershed regions, and then use it to achieve the globally optimal labeling. Second, by adopting the coarse-to-fine scheme, we can greatly reduce the computational time required for MAP labeling. Our experimental results have shown that the proposed method could efficiently produce accurate results of video object segmentation.

ACKNOWLEDGEMENTS

This work is partially supported by the National Science Council of Republic of China under the grant of NSC 90-2213-E-001-015 and NSC 91-2213-E-002-127.

REFERENCES

[1] C. Gu and M. C. Lee, "Semi-automatic segmentation and tracking of semantic video objects," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 8, No. 5, 1998, pp. 572–584.

[2] I. Patras, E. Hendriks and I. Lagendijk, "Video segmentation by MAP labeling of watershed segments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, 2001, pp. 326–332.

[3] J. Shi and J. Malik, "Motion segmentation and tracking using normalized cuts," *Proceedings of IEEE International Conference on Computer Vision*, Bombay, India, 1998, pp. 1154–1160.

[4] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8, No. 5, 1998, pp. 539–546.

[5] S. Jehan-Besson, M. Barlaud and G. Aubert, "Video object segmentation using eulerian region-based active contours," *Proceedings of International Conference on Computer Vision 2001 (ICCV '01)*, 2001, pp. 353–361.

[6] H. Luo and A. Eleftheriadis, "An interactive authoring system for video object segmentation and annotation," *Signal Processing: Image Communication*, Vol. 17, 2001, pp. 559–572.

[7] B. Marcotegui, P. Correia, F. Marques, R. Mech, R. Rosa, M. Wollborn and F. Zanoguera, "A video object generator tool allowing friendly user interaction," *Proceedings of IEEE International Conference of Image Processing, Kobe (Japan)*, Oct. 1999.

[8] H. Zhong, L. Wenyin and S. Li, "Interactive tracker – A semiautomatic system for video object segmentation," *Proceedings of IEEE International Conference on Multimedia and Expo* (*ICME '01*), 2001, pp. 645–648.

[9] Y.-P. Hung and Y.-P. Tsai, "Trail-dependent intelligent scissors based on multi-scale image segmentation," *Proceedings of Fifth Asian Conference on Computer Vision*, Melbourne, Jan. 2002, pp. 539–544.

[10] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 6, 1991, pp. 583–598.

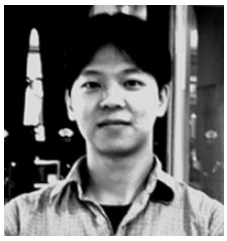
[11] L. Najman and M. Schmitt, “Geodesic saliency of watershed contours and hierarchical segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 12, 1996, pp. 1163–1173.

[12] E. N. Mortensen and W. A. Barrett, “Toboggan-based intelligent scissors with a four parameter edge model,” *Proceedings of IEEE: Computer Vision and Pattern Recognition* (*CVPR '99*), Vol. II, June 1999, pp. 452–458.

[13] S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, Tokyo, 1995.

[14] Y.-P. Hung, Y.-P. Tsai and C.-C. Lai, “A Bayesian approach to video object segmentation via merging 3D watershed volumes,” *Proceedings of International Conference on Pattern Recognition (ICPR02)*, Quebec, Canada, Vol. 3, Aug. 2002.

Ping-Che Chen (陳 秉 哲) 於 2000 年獲得台灣大學資訊工程系學士學位，於 2002 年獲得台灣大學資訊工程研究所碩士學位。他曾參與 1998 年亞洲區 ACM 程式設計比賽台北站獲得第二名。研究領域爲電腦視覺和電腦圖學。



Jin-Jen Su (蘇 敬 仁) 於 2003 年取得台灣大學資訊工程研究所碩士學位，研究領域爲電腦視覺 (Computer Vision) 與模組辨認 (Pattern Recognition)，研究方向則集中在視訊影像物件切割 (Video Object Segmentation)。

Yu-Pao Tsai (蔡 玉 寶) 中央研究院資訊科學所研究助理。1997 年獲得政治大學資訊科學系學士、1999 年獲得交通大學資訊科學系碩士。研究領域爲電腦圖學、電腦視學。

Yi-Ping Hung (洪 一 平) 1982 年獲得國立台灣大學電機工程學系學士學位，1987、1988 和 1990 年分別獲得美國 Brown University 工程碩士學位、應用數學碩士學位以及工程博士學位。1990 年起任職於中央研究院資訊科學研究所，並於 2002 年轉任國立台灣大學資訊工程系教授。他曾於 1997 年獲頒中央研究院年輕學者研究著作獎，並於 1996 到 1997 年擔任中央研究院資訊科學研究所副所長。其目前研究領域爲電腦視覺、圖形識別、影像處理、虛擬實境、多媒體以及人機界面。

收稿日期 92 年 12 月 17 日、修訂日期 93 年 2 月 7 日、接受日期 93 年 2 月 12 日
Manuscript received December 17, 2003, revised February 7, 2004, accepted February 12, 2004