AUTOMATIC VIDEO OBJECT SEGMENTATION USING GRAPH CUT[1]

Ying Mu[1], *Hong Zhang*[1*], *Helong Wang*[2], *Wei Zuo*[1]

[1] Image Processing Center, Beihang University, Beijing, 100083, China
[2] Luoyang electro-optical equipment research institute,471009,China

ABSTRACT

This paper presents an algorithm for automatic video object planes extraction from coarse to fine. For the case of single moving object in a scene, block-based segmentation defines regions of foreground, background and boundary blocks. Then, the segmentation problem is formulated as an energy minimization problem which is settled by using graph cut algorithm. Automatic segmentation can be realized by obtaining prior knowledge from foreground and background blocks and computation complex is reduced by restricting the refined segmentation region to boundary blocks. Experimental results show the effectiveness of proposed algorithm. It is can be implemented in the head-and-shoulder video sequence segmentation applications.

Index Terms—MPEG-4, video object, segmentation, automatic, graph cut

1. INTRODUCTION

The traditional video coding standards, such as MPEG-1, MPEG-2, H.263, is based on block coding scheme. They are widely implemented in media storage and broadcasting. But advances in multimedia technology require more efficient video coding methods. While the standard MPEG-4 provides content-based functionalities by introducing concept of video object. Video sequence is treated as an organized collection of video objects which are encoded and decoded separately. A video object (VO) is a semantically meaningful entity. Video object plane (VOP) is the representation of VO in a certain frame in video sequence. Decomposing each frame into VOP is the key issue in the content based video coding.

The VOP extraction is a challenging task either by means of automatic or semiautomatic method. Semiautomatic methods get foreground and background information from user's input. And then many algorithms [1-2] are proposed to obtain the refined segmentation results. The result also can be improved by user adjustment. A novel algorithm based on graph cut is proposed by Boykov and Jolly [3]. It is based on graph theory and an energy function to be minimized according to the user-imposed constrains. Semi-automatic method is used effectively in non-real-time applications while automatic one is desirable for real-time applications such as video conferences, etc. Gunsel et al. [4] proposed a K-means clustering algorithm using the difference of histogram to determine the location of boundaries. A binary object mask is generated by edge pixels detected by canny operator in [5]. The model is updated by using change detection mask. Beung-Chan Kim proposed a block-based segmentation algorithm for VOP extraction in [9]. In the recent years, Hidden Markov model is used in the video object segmentation and obtains effective results [6].

In this paper, an algorithm is proposed for single video object segmentation. Initially, block-based segmentation is performed on each frame by using color and motion information. The blocks are classified into foreground, background and boundary blocks. More accurate pixel-wise segmentation result can be obtained by introducing the graph cut algorithm which is used in semi-automatic segmentation. Refined segmentation region is restricted to boundary blocks. So the number of the nodes can be reduced in the graph to be constructed.

This paper is organized as follows. Section 1 gives a review of previous algorithms used in semi-automatic segmentation and automatic segmentation. In section 2, the proposed algorithm is described in details. Experimental results are represented in section 3 and section 4 gives conclusions.

2. VIDEO OBJECT SEGMENTATION

A VOP consists of three kinds of information to be encoded: texture, shape and motion. The shape information is a binary frame which is called alpha-plane. In the alpha-plane, the value of '0' ('1') indicates pixels outside (inside) the

[1] Supported by national key laboratory (51476010105HK0101)

[*] Corresponding author
Email: dmrzhang@buaa.edu.cn
Tel. (+86)10-82338991

object. The segmentation problem can be viewed as a binary labeling problem.

To reduce the computation load, the proposed algorithm uses block as processing unit. Firstly, the blocks are classified. Different types of blocks are indicated in Fig. 1.

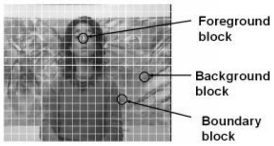

Fig. 1. three types of block (silent sequence)

2.1. First step: block-based video object segmentation

In this paper, the algorithm for block-based segmentation utilizes color and motion information at the same time. Taking the video sequence named silent for example, the block-based object segmentation is performed as follows.

2.1.1. Object block segmentation

Initially, moving region search is performed on the difference image obtained from two successive frames in YUV color space. The sum of absolute differences (SAD) is calculated on the each block in the image. The SAD of the block at (x, y), which is the left top position of a block, is calculated as:

$$SAD(x,y)=\sum_{i=0}^{15}\sum_{j=0}^{15}|I_k(x+i,y+j)-I_{k-1}(x+i,y+j)| \tag{1}$$

where $I_k(x,y)$ and $I_{k-1}(x,y)$ are the gray level in Y color space at (x, y) of the current and previous frames. To select the block with reasonable value of SAD as moving region, optimal threshold is selected by applying Ostu method [7]. According to this method, the optimal threshold T satisfies the following function:

$$\eta(T)=\max[\frac{\sigma_B^2(T)}{\sigma_W^2(T)}] \tag{2}$$

where $\sigma_w^2(T)$ is the in-class variance, and $\sigma_B^2(T)$ is the between-class variance. Since a moving object usually introduces significant intensity change between two successive frames, the block satisfied with $SAD(x,y)>T$ is considered as foreground block. The result is shown in Fig. 2.

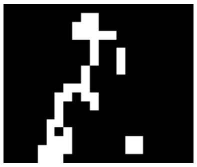

Fig. 2. segmentation by using the value of SAD (silent sequence)

Note that one single object is assumed in head-and-shoulder sequence, the largest connected region is defined as the object region. Since the color is homogeneous in the object region, the result should be adjusted by color information. Pixels in a frame are clustered by the K-means method in YUV space. The mean colors of each cluster are denoted as $\{K_i\}$. For each block in the frame, the average value of gray level is $C(j)$. The block is classified into the class i if $d_i=\min\|C(j)-K_i\|$. The largest connected region is considered as class k if $d_k=\max\{d_i(j)\}$. Then the block adjacent to the largest connected region is merged into the region if it belongs to class k.

And then a post-processing operation is performed to remove the small regions and filling the holes of object region. The final result is given as a block-based mask which indicates the block region. The Fig. 3 shows the binary mask and the region of object in the original image.

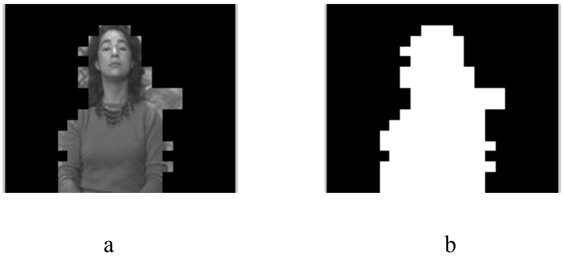

Fig. 3. result of block-based segmentation (silent sequence) (a) region of object in original image (b) binary mask

2.1.2. Block classification

According to the result, boundary blocks are specified for further segmentation. Since the boundary blocks constitute the contours of the object, the result of the block classification in this initial step is very important for the overall performance. Therefore, it should be ensured that the frame is over-segmented rather than under-segmented. Fig. 4 shows the classification result:

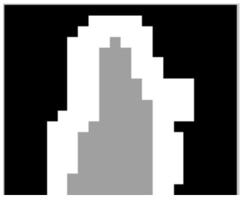

Fig. 4. block classification (silent sequence)

In the Fig. 4, blocks in a frame are classified into three classes: foreground (gray blocks), boundary blocks (white blocks) and background blocks (black blocks). Further segmentation restricts the segmentation region to boundary blocks. Pixels in the foreground and background blocks are used as seed pixels to obtain the prior knowledge of foreground and background.

2.2. Second step: refined segmentation on boundary blocks

Inspired by algorithm proposed by Boykov and Jolly [3] for interactive image segmentation, the image is supposed to be a graph $G = (V, E)$, where V is the set of nodes and E is the set of arcs connecting adjacent nodes. In the image, the nodes are pixels and the arcs are the connection relationship between adjacent pixels. The pixel-wise segmentation is to obtain the alpha-plane mentioned above by assigning value x_i to each node $i \in V$. x_i is represented as follows:

$$x_i \begin{cases} 1 & i \in \mathit{foreground} \\ 0 & i \in \mathit{background} \end{cases} \tag{3}$$

The set of $X = \{x_i\}$ can be obtained by minimize the energy function [3]:

$$E(X) = \sum_{i \in V} E_u(x_i) + \lambda \sum_{(i,j) \in E} E_v(x_i, x_j) \tag{4}$$

where $E_u(x_i)$ represents the cost when assigning node i to x_i . $E_v(x_i, x_j)$ represents the cost of when assigning adjacent nodes i and j to x_i and x_j respectively.

In this paper, we focus on how to define the energy term $E_u(x_i)$ and $E_v(x_i, x_j)$ according to the result of first step. Once block-based segmentation is done, pixels in foreground and background blocks are defined as foreground and background seeds which are represented by F and B respectively. The boundary blocks are defined as uncertain region which is represented by U . The value of pixels in F and B is confirmed in the set of X . Only pixels in U are needed to be assigned according to the following discussion.

$E_u(x_i)$ is defined as the color similarity of a node. The pixels in foreground and background are used as seeds to obtain the histograms for foreground and background intensity distributions. The energy term defined for uncertain region is as follows:

$$\begin{cases} E_u(x_i = 1) = -\log p(i \mid F) \\ E_u(x_i = 0) = -\log p(i \mid B) \end{cases} \quad i \in U \tag{5}$$

The equation is to ensure the node has similar color information to foreground or background.

$E_v(x_i, x_j)$ represents the color gradient between two adjacent nodes. Based on color and position information of pixels, $E_v(x_i, x_j)$ is defined as follows:

$$E_v(x_i, x_j) = | x_i - x_j | \cdot \exp(-\frac{\| F_i - F_j \|^2}{\sigma^2}) \cdot \frac{1}{\| d_i - d_j \|^2}$$
$$i, j \in V \tag{6}$$

where F_i is the gray level in YUV space of pixel i . σ^2 is the standard variance of Gaussian distribution of gray level. $\| d_i - d_j \|^2$ is the distance between pixel i and j . $| x_i - x_j |$ is the difference between the labels of pixel i and j . This item restricts the gradient information calculation to the segmentation boundary. Because the segmentation is only performed on the uncertain area, the $E_v(x_i, x_j)$ is defined when the adjacent pixels are assigned different labels.

To minimize energy function $E(X)$ in equation (4), min-cut/max-flow algorithm [8] is used. The algorithm is powerful in optimization problems used in minimizing certain energy function in vision problems. The final alpha-plane is shown in Fig. 5.

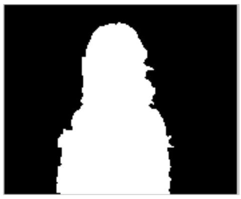

Fig. 5. alpha plane of silent sequence

By redefining the energy terms in graph cut algorithm, refined segmentation is performed on the uncertain region. The alpha-plane is obtained which indicates the pixel-wise region of the object. This method makes it is possible that extracting the object from the video sequence automatically. It is an important step in object-based video coding scheme.

3. EXPERIMENTAL RESULTS

The proposed algorithm is applied to automatic segmentation of video object for MPEG-4 applications. This algorithm is effective in single moving object segmentation. Head-and-shoulder video sequences in size of 176×144 are used for testing. In our experiments the block size is defined as 8×8. The results of block based segmentation step are shown above taking the example of silent sequence. Fig. 6 shows the final result of segmentation.

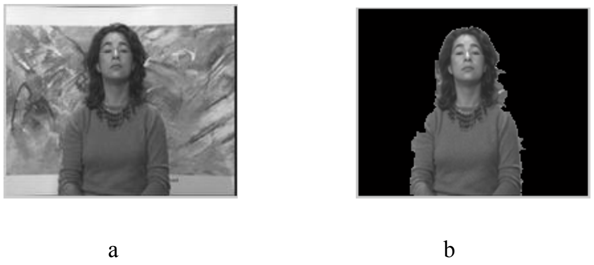

Fig. 6. video object in video sequence. (a) a frame in video sequence. (b) automatic VOP extraction results.

In Fig. 6, pixel-wise segmentation results are compared with the original video sequences. We can see from the figure, the proposed method can get nearly closed video object's boundary. Similar results can be obtained in other head-and-shoulder sequences. The results demonstrate that the proposed method is effective in head-and-shoulder sequence.

The results are also compared with that of Kim's method [9]. The comparison is shown in Fig. 7.

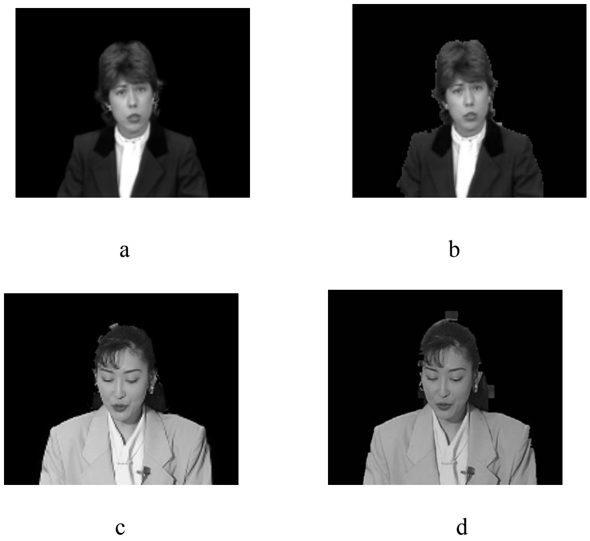

Fig. 7. VOPs extracted by using two different methods. (a) (c) VOP extracted by using the proposed method. (b) (d) VOP extracted by using Kim's method.

Kim proposed a block-based segmentation method in [9]. A relaxation algorithm is used to obtain the refined segmentation results. From Fig. 7, we can see that the contours of the object obtained by using the proposed algorithm are better than that obtained by Kim's method.

The results shown above demonstrate the effectiveness of the proposed algorithm in VOP extraction in MPEG-4 applications. It can extract the semantic object in the sequence and make it possible to realize the object based video coding scheme.

4. CONCLUSIONS

In this paper, an automatic video object segmentation method is proposed for MPEG-4 applications. Coarse result of foreground and background is obtained in the initial block-based segmentation step. Then the blocks in a frame are classified into three classes: foreground, background and boundary blocks. Segmentation region is restricted to the boundary blocks in the refined segmentation step to reduce the computation load. By redefining the energy term in graph cut algorithm, pixel-wise result is obtained automatically. Experimental results demonstrate the effectiveness of proposed algorithm. It can be applied to head-and-shoulder video type, and can be implemented in object based video coding in the further step.

5. REFERENCES

[1] Y. Y. Chuang, B. Curless, D. H. Salesin, R. Szeliski, A Bayesian approach to digital matting. In Proceeding of the IEEE Computer Vision and Pattern Recognition 2001, Hawaii, (2001), 264-271.

[2] Na Li; Shipeng Li; Chun Chen，Video object extraction using extended intelligent scissors Image Processing, 2003. ICIP 2003. Volume 2, 14-17 Sept. (2003) Page(s):II - 439-42 vol.3

[3] Boykov, Y.Y， Jolly, M.-P，Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images，Computer Vision, 2001. ICCV 2001. Volume 1, 7-14 July (2001) Page(s):105 - 112 vol.1

[4] Gunsel B, Ferman AM, Tekalp AM. Temporal video segmentation using unsupervised clustering and semantic object tracking. Journal of Electronic Imaging (1998)7(3):592－604.

[5] Meier T, Ngan K. Video segmentation for content-base coding. IEEE Transactions on Circuits and Systems for Video Technology (1999) 9:1190－203.

[6] Criminisi, A. Cross, G. Blake, A. Kolmogorov, V.，Bilayer Segmentation of Live Video，Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on Volume 1, 17-22 June (2006) Page(s):53 – 60

[7] Ostu N. A threshold selection method from gray - level histogram. IEEE Trans , (1979) SMC29 : 62～66.

[8] Boykov, Y.; Kolmogorov, V. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision; Pattern Analysis and Machine Intelligence, Volume 26, Issue 9, Sept. (2004) Page(s):1124 – 1137

[9] Beung-Chan Kim and R.-H.Rae-Hong Park, A fast automatic VOP generation using boundary block segmentation"，Real-Time Imaging, Volume 10, Issue 2, April (2004) Pages 117-125