AUTOMATIC EXTRACTION OF MOVING OBJECTS USING MULTIPLE FEATURES AND MULTIPLE FRAMES

Jinhui Pan[*] Shipeng Li, Ya-Qin Zhang
pan_jinhui@hotmail.com spli,yzhang@microsoft.com
Tsinghua University *Microsoft Research, China*

ABSTRACT

This paper introduces a novel automatic video object extraction algorithm based on combination of color and motion segmentation results. The algorithm includes five parts: pre-processing, color segmentation, motion segmentation, combination of color and motion segmentation of multiple frames, post-processing. The performance of this algorithm is very promising, resulting in pixel-wise accuracy of extracted objects. Since it is an automatic extraction algorithm, it can be very useful in some real time video processing system based on video objects.

1. INTRODUCTION

Content-based representation and coding of the visual information is currently becoming an extremely active field of research [8][9]. The ISO MPEG-4 standard has attracted much attention recently for providing a standardized solution for content-based access and manipulation for these applications. The standard enables content-based functions by introducing the concept of video object planes (VOPs). A video sequence can be decomposed into VOPs by extracting semantic objects in it [10][11].

Currently, existing segmentation methods include: region-based approach using color homogeneity [5][18], object-based approach using motion criterion [12][13][17] and object tracking [1][3][5]. A semantic object may contain multiple regions with different colors and motion. So a single color or motion criterion cannot lead to satisfying result. Multiple features of a semantic object such as color and motion should be integrated together to form the final result.

There are many methods using multiple features in object extraction. Many research works focus on interactive object extraction [1][2][3]. These methods require human interactivity at least for the segmentation of the initial frame. They are flexible and relatively accurate. But human interactivity adds burden to users. And it is not suitable for real time processing systems, such as conferencing systems. Another kind of methods is automatic segmentation [4][5][6]. Algorithms in this group can extract semantic object automatically. But the results of most of these algorithms are rough and not suitable for general use.

Developing a generic automatic segmentation algorithm for all kinds of video sequences is not optimistic at present due to the difficulty in semantic segmentation. So we narrow down the problem scope to moving objects extraction, specifically, we developed an automatic video object segmentation algorithm for initial frames. The same object tracking algorithm used in [3] can be borrowed to extract the video objects in a whole video sequence with the initial frames. Color and motion features of multiple frames are combined in this algorithm, resulting in an accurate extraction of moving objects in a video sequence. Since it is an automatic algorithm, it is a good candidate for real time video processing, such as object based video conferencing systems. It will definitely cause an initial delay (no more than 1 second for QCIF sequences) for extracting the first frame because this algorithm uses information of multiple frames. However, it is much faster than interactive extraction. This automatic extraction algorithm is suitable for separating moving foreground from still background, especially for video sequences with relatively slow motion.

This paper is organized as follows. Section 2 gives an overview of the object extraction system. Section 3 demonstrates the combination of color and motion of multiple frames to extract the moving objects. Section 4 displays results of the algorithm. Section 5 gives some discussions and future works.

2. OVERVIEW OF THE SYSTEM

To extract a video object accurately, we combine the intermediate results of color segmentation and motion segmentation. The result of motion segmentation is affected by the precision of motion estimation. Motion estimation itself is not accurate due to the noise in the frames. It can only give a rough region of the moving object. On the other hand, color segmentation can give more accurate region edges, but the color-segmented regions usually scatter in a frame. So we need to combine color and motion information of multiple frames to generate a satisfying result (Fig 1).

Before segmentation, a pre-processing is necessary to smooth out some random noise. Here we use a vector median filter to fulfil this task [3]. This vector median filter is better than other low pass filter for segmentation pre-processing in that it simplifies the input data while preserving the boundary.

[*] This work was done while the author was with Microsoft Research, China

After pre-processing, the input sequence is segmented using region growing to get color-segmented mask [3]. Meanwhile, it is also segmented according to motion. There are many methods on motion estimation and segmentation [12][13][14][15][16]. Here, we adopt region-based motion estimation put forward by Gu, etc[3]. Then the color masks and motion masks of multiple frames are combined into a mapping operator to get the rough object mask. This object mask is post-processed to remove small holes in the object and small background residual regions.

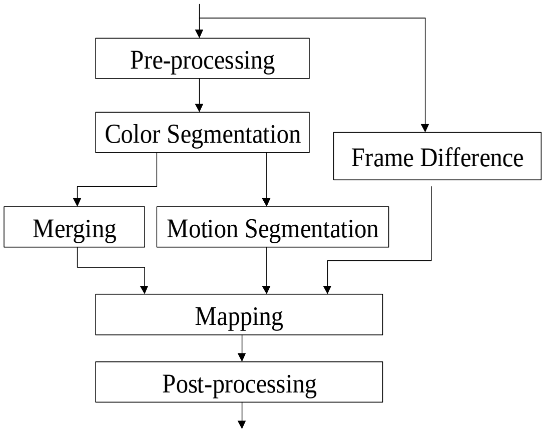

Fig 1. Flow Chart of the System

3. COMBINATION OF MULTIPLE FEATURES AND MULTIPLE FRAMES

Both color and motion belong to low-level features for segmentation. Neither of them alone can lead to satisfying extraction of object. So we combine multiple features and multiple frames according to the rules described as follows.

(1) Combination of Multiple Frames

To get motion mask, we always do motion estimation using two continuous frames. But the motion vector is not accurate enough, especially in regions at the boundary of the object. This is caused by many reasons. If a moving object contains uniform regions, it is difficult to get precise motion estimation. At the same time, random noise adds to the inaccuracy in matching. And the uncovered background would be always assigned to a moving object. All these can make motion estimation inaccurate, or even wrong.

To solve these problems, we combine information from multiple frames to get a refined mask. We can remove the ”motion” caused by random noise by checking multiple motion masks because random noise will not always be in the same place in multiple frames. And the uncovered background can also be distinguished with true moving regions according to this method. Of course, in a higher motion sequence, the number of frames needed is also smaller.

What we do here is to calculate the frequency that a pixel is assigned as a point in a moving object in multiple (usually, 10 or more) motion masks. If the frequency is higher than a certain threshold, then this pixel is considered as a moving one. Otherwise, it is designated as background and is removed from the final motion mask. The method can be formulated as follows. (Fig 2 shows the results of combination of multiple frames.)

$$C(i,j) = \sum_{s=1}^{S} MS_S(i,j)\,. \qquad (3.1)$$

$$M(i,j) = \begin{cases} 1\,, & \text{if } C(i,j)/S > T2; \\ 0\,, & \text{otherwise}\,, \end{cases} \qquad (3.2)$$

where $MS_S(i,j)$ represents each motion mask being combined, $MS_S(i,j)$ =1 means pixel (i,j) is in a moving region and $MS_S(i,j)$ =0 means pixel (i,j) is in the background. S denotes the number of total motion masks used in the combination, C(i,j) denotes the total times that pixel (i,j) is assigned as moving pixel in S motion masks. T2 is the threshold, which is often set to 50%.

(2) Combination of Multiple Features

Color segmentation provides accurate edges of objects, but it always results in over-segmentation. While motion segmentation provides a coarse mask of moving objects, but its boundary is too rough to lead to an accurate extraction.

To get moving object with pixel-wise accuracy, we combine the color and motion information by mapping operation [6]. For each region generated in color segmentation, we find its correspondence in motion mask. If the percent of the region that is assigned to a moving object exceeds a certain threshold, the whole region belongs to moving object. This method can be formulated as follows:

$$B(N) = \sum_{(i,j)\in N} M(i,j)\,. \qquad (3.3)$$

$$J(N) = \begin{cases} 1, & \text{if } B(N)/A(N) > T3; \\ 0\,, & \text{otherwise}\,. \end{cases} \qquad (3.4)$$

$$FM(i,j) = J(N)\,, \qquad \text{if } (i,j) \in N\,, \qquad (3.5)$$

where N represents color segmented region, A(N) is the area of region N. M(i,j) is the pixel in the combined motion mask, FM(i,j) is the pixel in mapping mask. T3 is the threshold mentioned above, which is often set to 50%-60%.

Sometimes, there could be errors in motion estimation due to the coarse motion estimation algorithm. To remove regions incorrectly assigned to moving objects, difference mask is added into mapping operator. Then the results of motion mask mapping and difference mask mapping are combined by a logical operation AND to get the final object mask. The difference mask DM(i,j) is obtained as follows:

$$D(i,j) = | I_t(i,j) - I_{t+m}(i,j) | . \quad (3.6)$$

$$DM(i,j) = \begin{cases} 1, & \text{if } D(i,j) > T4; \\ 0, & \text{otherwise.} \end{cases} \quad (3.7)$$

Then we can get the difference mapping result FD(i,j) using a similar mapping method described in formula 3.3-3.5, as follows:

$$B(N) = \sum_{(i,j)\in N} DM(i,j) . \quad (3.8)$$

$$J(N) = \begin{cases} 1, & \text{if } B(N)/A(N) > T3; \\ 0, & \text{otherwise.} \end{cases} \quad (3.9)$$

$$FD(i,j) = J(N), \qquad \text{if } (i,j) \in N. \quad (3.10)$$

Then, we can get the result mask combining color, motion and difference by:

$$F(i,j) = \begin{cases} 1, & \text{if } FM(i,j) = 1 \text{ and } FD(i,j) = 1; \\ 0, & \text{otherwise.} \end{cases} \quad (3.11)$$

However, difference is an optional feature. In MPEG-4 test sequences, color and motion are good enough for extracting objects. In this case, difference mask is not included in the system and F(i,j) can be obtained simply by

$$F(i,j) = FM(i,j) . \quad (3.12)$$

(3) Post Processing

From the former steps, we obtain an object mask image that roughly describes the moving object to be extracted. A post processing is used for a better extracting result. Here isolated small regions were merged to neighborhood to get a clean mask. Then morphological operators[7] (open and close) can be adopted according to different situations. Thus, we get a clean object mask, which we can use to extract moving objects in a video sequence(Fig3).

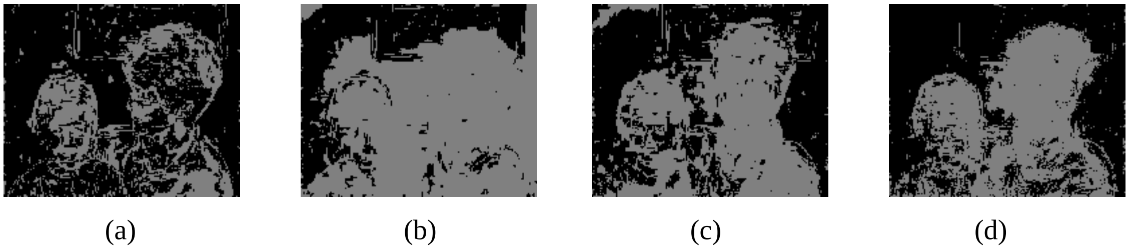

Fig 2 Mother and Daughter: result of combination of multiple frames.

(a)—(c) Motion mask from every two frames, (d) result of combination of (a)—(c).

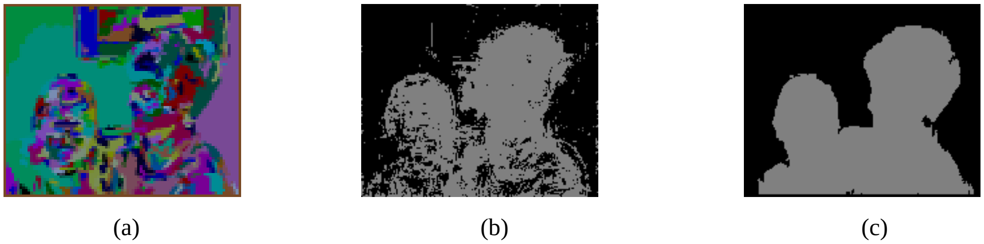

Fig 3 The result of mapping and post-processing:

(a) the result of color segmentation , (b) the motion mask, and (c) final object mask.

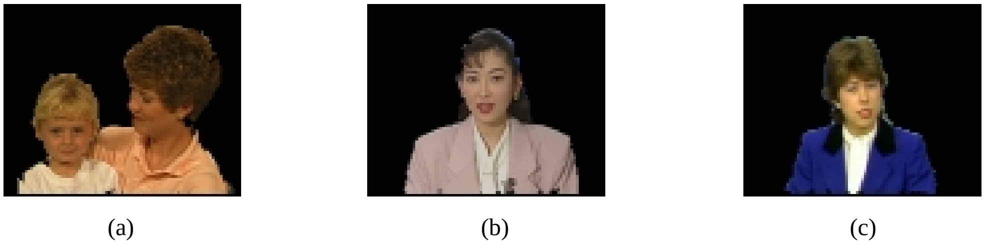

Fig. 4: Experimental results on MPEG-4 test sequences: extracted objects.

(a) Mother and Daughter: Frame 0. (b) Akiyo: Frame 20. (c) Claire: Frame 150.

4. EXPERIMENTAL RESULTS

We have tested the proposed algorithm using MPEG-4 test sequences that contain moving foreground objects and still background and some self-captured real life video scenes, the proposed algorithm works well for all the cases. We choose a typical MPEG-4 sequence: Mother-and-Daughter to show the algorithm results. The video sequence is in QCIF format. Threshold for color segmentation is set to 8, and threshold for motion segmentation is set to 1. 10 motion masks are combined to get the motion segmentation result. The interval of the motion masks is set 10. Mapping threshold is set 60%. Frame difference is not included in the system for this sequence. The results of the extraction algorithm on other two MPEG-4 test sequences: Akiyo and Claire are also shown in Fig. 4.

Subjectively, we can see from Fig 4 that moving objects in the sequences are extracted accurately. The extracted objects are in pixel-wise precision. It produces very good initial segmentation result for further processing, such as object-based tracking [1][3]. More work is being done to develop an objective measurement of the segmentation results.

5. CONCLUSION AND FUTURE WORK

The proposed algorithm concentrates on extraction of moving objects in a video sequence. It makes use of multiple features and multiple frames to generate the final result. Since it combines color and motion information according to their characteristics, the performance of the algorithm is very promising. And the pre-processing and post-processing also refine the result.

This algorithm extracts a moving object accurately. The result can be used as an initial object to replace human interactivity in semi-automatic segmentation system [3]. This can leads to a totally automatic segmentation, which is especially suitable for real time video object processing once the computing speed is improved.

Many parameters need to be set in this algorithm. Some are fixed according to experience for all video sequences. But some should be adjusted according to different video sequences. In fact, some parameters can be set automatically if motion is estimated roughly at the beginning. And dominant motion detection and global motion compensation are not included in the system. So it can only handles video sequences with still background at present. Global motion compensation will add to its generality.

Since this algorithm is based on rules of combination, it is easy to be extended by adding new information into the mapping operator. For example, depth is an important clue for segmentation. If depth calculation is robust enough to be combined into the algorithm, it will improve the final result.

6. REFERENCES

[1] C. Gu and M.-C. Lee, “Semiautomatic Segmentation and Tracking of Semantic Video Objects”, IEEE Trans. Circuits Syst. Video Technol. VOL 8, NO. 5, Sept. 1998.

[2] R. Castango, T. Ebrahimi and M. Kunt, “Video Segmentation Based on Multiple Features for Interactive Multimedia Application”, IEEE Trans. Circuits and Syst. Video Technol. VOL. 8, NO. 5, Sept. 1998.

[3] C. Gu and M.-C. Lee, “Tracking of Multiple Semantic Video Objects for Internet Applications”, SPIE, Visual Communications and Image Processing’99, Volume 3653.

[4] T. Meier and K. N. Ngan, “Automatic Segmentation of Moving Objects for Video Object Plane Generation”, IEEE Trans. Circuits Syst. Video Technol. VOL. 8, NO. 5, Sept. 1998.

[5] D. Wang, “Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking”, IEEE Trans. Circuits Syst. Video Technol. VOL. 8, NO. 5, Sept. 1998.

[6] A. A. Alatan, L. Onural, M. Wollborn,etc. “Image Sequence Analysis for Emerging Interactive Multimedia Services—The European COST 211 Framework”, IEEE Trans. Circuits Syst. Video Technol. VOL 8, NO. 7, Nov. 1998.

[7] J. Serra, Image Analysis and Mathematical Morphology. New York: Academic, 1982.

[8] M. Kunt, A. Ikonomopoulos, and M. Kocher, “Second generation image coding techniques”, Proc. IEEE VOL. 73,pp. 549-575, Apr. 1985.

[9] H. G. Musmann, M. Otter, and J. Ostermann, “ Object-oriented analysis-synthesis coding of moving images”, Signal Processing: Image Comm. VOL. 1,pp. 121-130,1989.

[10] T. Sikora, “The MPEG-4 video standard verification model”, IEEE Trans. Circuits Syst. Video Technol. VOL. 7, pp. 19-31, Feb. 1997.

[11] MPEG Group, “Overview of the MPEG-4 version 1 standard”, ISO /IEC JTC1/SC29/WG11, Doc. no. 1909, Oct. 1997 [Online].

[12] A. M. Tekalp, Digital Video Processing, Prentice Hall, Inc, 1996.

[13] M. M. Chang, A. M. Tekalp, and M. I. Sezan, “Simultaneous motion estimation and segmentation,” IEEE Trans. Image Processing, VOL .6, pp.1326—1333,Sept. 1997.

[14] D. Wang, L. Wang, “Global motion parameter estimation using a fast and robust algorithm”, IEEE Trans. Circuits Syst. Video Technol. VOL.7, NO.5, Oct. 1997.

[15] H. Jozawa, K. Kamikura, etc. “Two-stage motion compensation using adaptive global MC and local affine MC”, IEEE Trans. Circuits Syst. Video Technol. VOL. 7,NO. 1, Feb. 1997.

[16] C.-H. Lin, J.-L. Wu, “A lightweight genetic block-matching algorithm for video coding”, IEEE Trans. Circuits Syst. Video Technol. VOL. 8,NO. 4, August 1998.

[17] J. Y. A. Wang and E. H. Adelson, “Representing moving images with layers”, IEEE Trans. Circuits Syst. Video Technol. VOL. 3, NO. 5, Sept. 1994.

[18] J. Scharcanski and A. N. Venetsanopoulos, “Edge detection of color image using directional operators”, IEEE Trans. Circuits Syst. Video Technol. VOL. 7, NO. 2, April 1997.