VIDEO SEGMENTATION BASED ON SPATIAL AND TEMPORAL INFORMATION

Jae Gark Choi

Electronics and Telecommunications Research Institute, 161 Kajong-Dong, Yusung-Gu, Taejon, 305-350, KOREA

Si-Woong Lee and Seong-Dae Kim

Dept. of Electrical Engineering, KAIST, 373-1 Kusong-Dong, Yusung-Gu, Taejon, 305-701, KOREA

ABSTRACT

The paper presents a morphological spatio-temporal segmentation method, which is based on a new similarity measure. This similarity measure considers jointly spatial and temporal information and consists therefore of two terms. The first term minimizes the displaced frame difference, considering the affine motion model. By the second term the spatial homogenity of the luminance values of every region is maximized. The procedure toward complete segmentation consists of three steps: joint marker extraction, boundary decision, and motion-based region fusion. By incorporating spatial and temporal information simultaneously, we can obtain visually meaningful segmentation results. Simulation results demonstrates the efficiency of the proposed method.

1. INTRODUCTION

Object-based coding algorithms segment image contents into a set of objects according to a given model and estimate their parameters which can then be encoded [1][2][3]. The segmentation approaches for such codings range from split and merge method [4] to morphological segmentation [5]. Among them, morphological segmentation techniques are of particular interest because they rely on morphological tools which are very attractive to deal with object-oriented criteria such as size and contrast. However, morphological filters operate only on luminance component and use size and contrast as criterions. The segmentation produced by only spatial information may have false contours.

Recently, an attempt has been made to use the informations from both spatial domain and temporal domain which results in a more meaningful segmentation for perception [6][7][8]. However, the spatio-temporal segmentation algorithms did not utilize joint similarity measure which may be simultaneously handled during segmentation. An efficient segmentation result will be expected if luminance and motion information are simultaneously used as a similarity measure.

This paper presents an efficient spatio-temporal segmentation algorithm using morphological tools and a joint similarity measure. The algorithm consists of three steps: joint marker extraction, boundary decision and motion-based region fusion. First, joint markers are extracted. A joint marker is a germ which is coherent in both motion and luminance. Second, region boundaries are decided by watershed algorithm which incorporates motion and luminance information simultaneously. For such a segmentation, a new joint similarity measure is proposed. Finally, regions with similar motion are merged into single entities and thus define the objects of the scene. This corresponds to motion-based region fusion where an effective region merging method is used. This algorithm makes no assumption about the content of images. Also, it gives more meaningful results for perception.

2. JOINT SIMILARITY MEASURE

In this section, we describe the joint similarity measure for the morphological spatio-temporal segmentation. We first point out the foreseen problems of the conventional joint similarity measure [10] and then propose a new joint similarity measure to overcome the problems.

A possible similarity measure for spatio-temporal segmentation is the weighted sum of the intensity difference plus the motion difference. The intensity difference is the gray level difference between the pixel under consideration and the mean of the pixels that have already been assigned to the region. The motion difference is the motion error between the estimated motion vector $(d_x(x,y), d_y(x,y))$ at pixel (x,y) and the motion vector $(d_x^\theta(x,y), d_y^\theta(x,y))$ generated at pixel (x,y) by the parametric motion model θ of the region. However, the existing techniques generating optical flow reveal inherent noise problem especially near

motion boundary [9]. As the optical flow is not very accurate at edge boundaries, the resulting segmentation using the similarity criterion loses object boundary precision. Besides, as the intensity difference and the motion difference have different units, the scaling between them is required.

To solve the problems, we use the displaced frame difference as indirect motion similarity instead of the motion difference. Thus the motion similarity between the pixel (x, y) under consideration and the region R is defined as the displaced frame difference,

$$S_m(x, y; R) = I_k(x, y) - I_{k-1}(x - d_x^\theta(x, y), y - d_y^\theta(x, y)) \quad (1)$$

where $I_k(x, y)$ is a gray value at (x, y) in kth frame. Note that it uses not the estimated motion vector at pixel (x, y) but the motion vector at pixel (x, y) generated by the motion parameters θ of the region R. Thus a new joint similarity measure can be defined as the weighted sum of the motion similarity plus the intensity similarity,

$$S(x, y; R) = \alpha \; S_m(x, y; R) \; + \; (1 - \alpha) \; S_i(x, y; R) \quad (2)$$

where α is a weight factor and $S_i(x, y; R)$ is the intensity difference between the pixel under consideration and the mean of the region R.

3. THE SEGMENTATION ALGORITHM

In this section, we describe our spatio-temporal segmentation algorithm. It is based on morphological segmentation which uses morphological filters and watershed algorithm as basic tools. The block diagram is illustrated in Fig. 1. Each process is described in the following subsections.

3.1. Joint marker extraction

The key to the success of morphological segmentation relies on the proper selection of markers which are perceptually important. Conventional marker extractions use size and contrast criterions of the simplified luminance image, but the segmentation produced by such criterions may have false contours. If luminance and motion are simultaneously incorporated, a more perceptually meaningful segmentation can be possible.

In this paper, we propose a simple joint marker extraction method. The joint marker extraction detects the presence of homogeneous regions in both motion and luminance, and produces markers identifying the interior of the regions that will be segmented. As illustrated in Fig. 1, images are first simplified to make them easier to segment. Morphological open-close by reconstruction filters are used for simplification. These filters remove regions that are smaller than a given size but preserve the contours of the remaining objects. Second, intensity markers are extracted from simplified luminance image. We select as intensity markers flat regions whose size have greater than a given threshold. From simplified images, intensity markers can simply be identified by labeling flat regions. Third, homogeneous motion regions inside intensity markers are extracted as motion markers. The motion marker is considered as the joint marker because it is a homogeneous region in both motion and luminance. Here, the homogeneity of motion is decided in view of affine model.

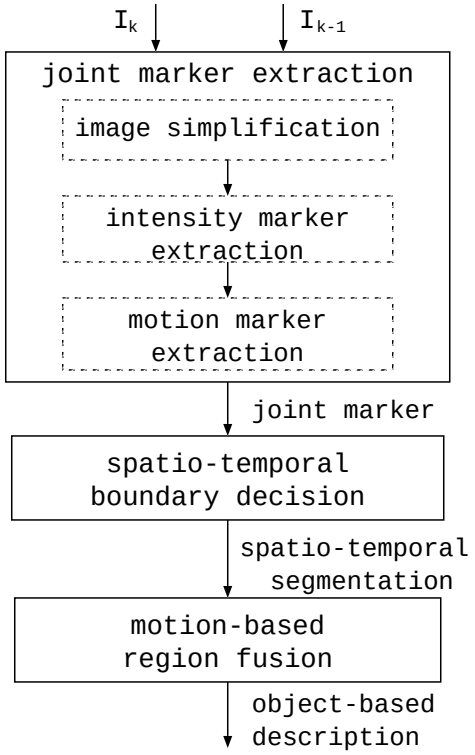


Figure 1: Structure of the segmentation algorithm

3.2. Boundary decision

After joint marker extraction, the number and the interior of the regions to be segmented are known. However, a large number of pixels are not yet assigned to any region. These pixels correspond to uncertainty areas mainly concentrated around the contours of the regions.

The decision about the actual borders of the regions are taken through the use of the watershed algorithm. It is basically a region growing algorithm, starting from markers, it successively joins pixels from uncertainty

area to the nearest similar region. It uses the joint similarity measure defined in Eq. (2) as similarity criterion. A pixel under consideration is assigned to a specific region because it is in the neighborhood of at least one marker and it is more similar (in the sense defined by the joint similarity measure) to this marker than to any other marker of its neighborhood.

Once a new pixel has been assigned to a region, the motion model θ of the region should be updated in order to accurately compute the joint similarity with respect to new pixels. However, update of the models per every assignment requires a heavy computation load. To avoid such a load, we update the models whenever only 50 pixels are added to the region. It is possible because addition of a few pixels does not abruptly change the motion parameters.

3.3. Motion-based region fusion

The segmentation in section 3.2 results in regions which are homogeneous in both motion and luminance. The segmented regions may have different average vlaue in luminance but may have the same motion parameter set. Therefore, an elimination of redundant regions is needed in view of the coding framework. This corresponds to motion-based region fusion and is expected to simplify the segmentation. This is important for coding since fewer regions permit a smaller bit rate allocated to the transmission of region shapes and motion parameters.

If regions created in the section 3.2 are consistent with the same affine motion, they should be merged together. Consistency with an affine transformation is detected by computing, using the least-squares technique, optimal parameters and related error values for a pair of adjacent regions [11].

4. SIMULATION RESULTS

In order to verify the performance of the proposed algorithm, simulations have been carried out on the "Table Tennis" and "Claire" sequences in QCIF format. Fig. 2 (a) and (b) show the original image of "Table Tennis" and its simplified image, respectively. Intensity markers are extracted from simplified images. Intensity markers are given in Fig. 3 (a) where each marker is labeled in gray level and uncertain areas are represented in white. Then, homogeneous motion regions inside intensity markers are extracted as joint markers. The resulting joint markers are given in Fig. 3 (b). We can see that the ball has been disappeared by the simplification step. Therefore, the ball and some parts of background are merged into one marker in Fig. 3 (a). But, we can separate the ball from the intensity marker by using the joint marker extraction. Fig. 4 (a) and (b) show the results of spatio-temporal segmentation and motion-based region fusion, respectively. As shown in Fig. 4 (b), the "Table tennis" image is segmented into 3 regions: the ball; racket and arm; background and table.

Fig. 5 shows the segmentation results for the image sequence "Claire". Fig. 5 (a) shows the original image. Fig. 5 (a) - (c) show the results of the joint markers extraction, the spatio-temporal segmentation and the motion-based region fusion, respectively. We can see that our joint spatio-temporal segmentation gives visually meaningful results. Furthermore, this spatio-temporal approach is well suited to coding applications such as object- and region-based coding.

5. CONCLUSION

An efficient spatio-temporal segmentation algorithm is presented in this paper. The proposed segmentation algorithm incorporates motion and luminance information simultaneously, and uses morphological tools such as morphological filters and the watershed algorithm. The algorithm gives visually meaningful segmentation results which makes region- or object-based coding efficient. Also this unsupervised algorithm makes the automatic segmentation possible.

6. REFERENCES

[1] M. Kunt, A. Ikonomopoulos and M. Kocher, "Second generation image coding techniques," *Proc. IEEE*, vol. 74, no. 4, pp. 549-574, 1985.

[2] E. A. Adelson and J. Y. A. Wang, "Representing moving images with layers," *IEEE Trans. on Image Processing*, vol. 3, pp. 625-638, Sept. 1994.

[3] P. Salembier, L. Torres, F. Meyer and C. Gu, "Region-based video coding using mathematical morphology," *Proc. IEEE*, vol. 83, no. 6, pp. 843-857, June 1995.

[4] D. Cortez, P. Nunes, M. Sequeira, and F. Pereira, "Image segmentation towards mew image representation methods," *Signal Processing: Image Communication*, vol. 6, no. 6, pp. 485-498, Feb. 1995.

[5] P. Salembier, "Morphological multiscale segmentation for image coding", *Signal Processing*, vol.38, pp. 359-386, 1994.

[6] C. Gu, T. Ebrahimi and M. Kunt, "Morphological spatio-temporal segmentation for content-based

video coding", *International workshop on coding techniques for very low bit-rate video*, Tokyo, Nov. 8-10, 1995.

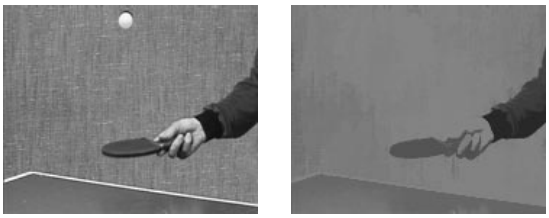
[7] N. T. Watsuji, H. Katata and T. Aono, "Morphological segmentation with motion based feature extraction", *International workshop on coding techniques for very low bit-rate video*, Tokyo, Nov. 8-10, 1995.

[8] F. Dufaux, F. Moscheni and A. Lippman, "Spatio-temporal segmentation based on motion and static segmentation", *IEEE Proc. ICIP'95*, Volume 1, Washington, DC, Oct. 1995, pp. 306-309.

[9] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: a review and a new contribution," *Proc. IEEE*, vol. 83, no. 6, pp. 858-876, June 1995.

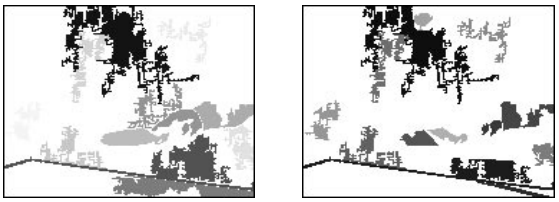
[10] W. H. Hong, N. C. Kim, and S. M. Lee, "Video segmentation using spatial proximity, color, and motion information for region-based coding," *Proc. Visual Communications and Image Processing '94*, vol. 2308, pp. 1627-1633, 1994.

[11] Jae Gark Choi and Seong-Dae Kim, "Multi-stage segmentation of optical flow field," *Signal Processing*, vol. 54, pp. 109-118, Oct. 1996.



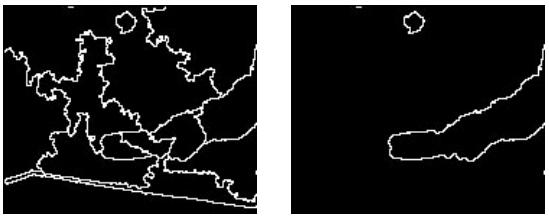
(a) (b)

Figure 2: Original image (a) and its simplified image (b) of "Table tennis"



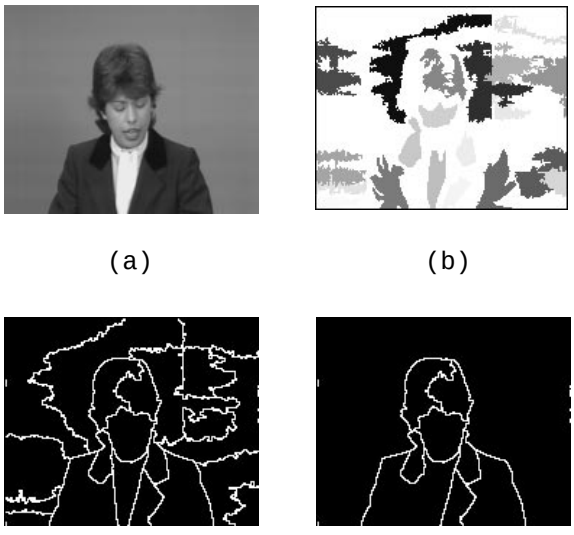
(a) (b)

Figure 3: Intensity markers (a) and joint markers (b) of "Table tennis"



(a) (b)

Figure 4: Spatio-temporal segmentation (a) and motion-based region fusion (b) of "Table tennis"



(a) (b) (c) (d)

Figure 5: The segmentation results of "Claire" image: (a) original image; (b) simplified image; (c) spatio-temporal segmentation; (d) motion-based region fusion