A hierarchical semantic video object tracking algorithm using mathematical morphology

Jaeyoun Yi, Hyun Sang Park, and Jong Beom Ra

Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
373-1 Kusongdong, Yusonggu, Taejon, 305-701, Republic of Korea

Abstract

In this paper, we propose a hierarchical segmentation method for tracking a semantic video object using a watershed algorithm based on morphological filtering. In the proposed method, each hierarchy consists of three steps: First, markers are extracted on the simplified current frame. Second, region growing by a modified watershed algorithm is performed for segmentation. Finally, the segmented regions are classified into 3 categories, i.e., inside, outside, and uncertain regions according to region probability values, which are acquired by the probability map calculated from a estimated motion field. Then, for the remaining uncertain regions, the above three steps are repeated at lower hierarchies with less simplified frames until every region is decided to a certain region. The proposed algorithm provides prospective results in video sequences such as *Miss America*, *Clair*, and *Akiyo*.

1. Introduction

The main feature of MPEG4, which distinguishes this coming video coding standard from its predecessors MPEG1 and MPEG2, is the support for content-based scalability and content-based multimedia data access. These functionalities enable separate coding of scene contents and will be very useful for future multimedia environments, e.g., when making synthetic sequences from various video sources. Images are not considered solely rectangular compositions of pixels, but compositions of different objects residing in their own layers. The MPEG4 denotes these layers as video object planes (VOP's). As the MPEG4 standardization becomes finalized, the method to extract the VOP's from a real video sequence is emerging as an interesting research issue.

Semantic video object represents a meaningful entity in a digital video clip. Although there are many conventional methods to segment an image into some homogeneous regions for the purpose of so-called 2nd generation coding, they can not be applied directly to the object segmentation problem. This is because a semantic video object does not compose of a single homogeneous region, but of many various non-homogeneous regions. Therefore, this problem has been dealt with two separate steps; human-aided segmentation for the first frame and automatic tracking for the remaining frames [1].

There have been several approaches to this object tracking problem recently. The method using a change detection mask is based on the evaluation of the frame difference [2]. Starting from the current frame, the change detection is carried out with respect to the preceding frame or the motion-compensated frame. But this approach

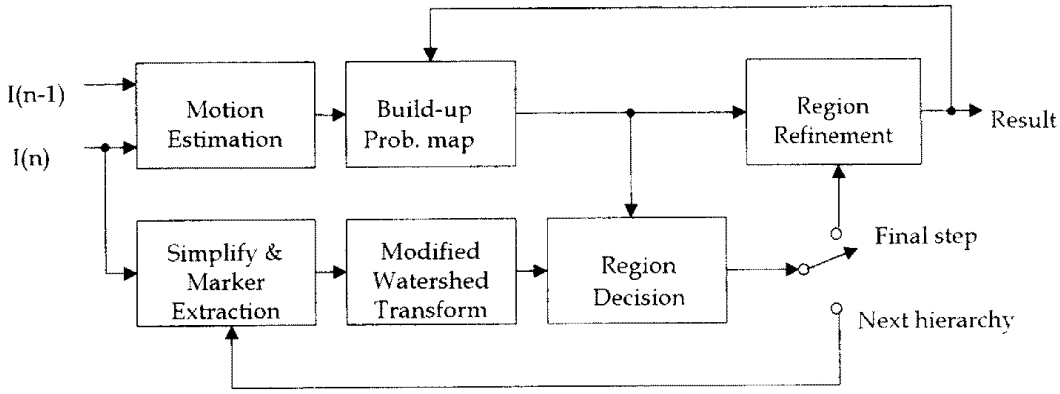


Fig. 1. Block diagram of the proposed tracking algorithm.

suffers from inherent inaccuracies due to mask operation and an imperfect moving area detection. On the other hand, active contour model based approaches have been proposed [3, 4]. These approaches refine motion-compensated contour pixels pel-by-pel by adjusting them to high gradient locations repeatedly. But they are improper for real-world images because high-gradient points may not be on the true boundary between a semantic object and its background.

For those reasons, a region-based approach is more appropriate for the real-world semantic object-tracking problem because regions can provide better boundaries. In this paper, we propose a hierarchical segmentation scheme to track a video object under the assumption that the intra-frame is already segmented by other methods, or by hand. In section 2, our proposed algorithm will described, and some experimental simulation results will be shown in section 3.

2. The proposed algorithm

The boundary of a segmented region is clearer at the strongly simplified image because the strong morphological filtering can alleviate noise effects. But strong filtering can also remove some true object boundaries with small gradient values. So we propose a hierarchical algorithm depending on simplification degree. In this algorithm, most of regions are decided into either inside or outside regions with clear boundaries at the high hierarchy and the remaining undecided regions are examined at the lower hierarchies until all regions are categorized into certain regions. The algorithm consists of three hierarchies.

Fig. 1 shows the whole diagram of the proposed object-tracking algorithm. At each hierarchy, the current frame is simplified and segmented, and every segmented region is classified into one of 3 categories, i.e., "inside", "outside" or "uncertain" region. This region classification is based on the probability map produced by motion compensation of the previous segmentation result. The details will be explained below.

2.1. Simplification

To simplify the current frame, morphological filters

are used at each hierarchy as follows.

$$\text{Level 1:}\quad \varphi^{(rec)}\{\delta_6(\gamma^{(rec)}(\varepsilon_6(f),\gamma_1(f)),\ \varphi_6(\gamma^{(rec)}(\varepsilon_6(f),\gamma_1(f))\}$$

$$\text{Level 2:}\quad \varphi^{(rec)}\{\delta_4(\gamma^{(rec)}(\varepsilon_4(f),f),\varphi_4(\gamma^{(rec)}(\varepsilon_4(f),f))\}$$

$$\text{Level 3:}\quad \varphi^{(rec)}\{\delta_1(\gamma^{(rec)}(\varepsilon_1(f),f),\varphi_1(\gamma^{(rec)}(\varepsilon_1(f),f))\}$$

Markers are then extracted from uncertain regions in the simplified current frame. Even though many algorithms to extract good markers have been proposed, in this paper, we adopt the method given in [6].

2.2. Modified watershed algorithm

In morphological segmentation, a watershed algorithm is used as a region-growing tool after marker extraction. The watershed algorithm is derived from topographic works where catchment basins and their dividing lines, called watershed lines, have been extensively studied. In image processing, this notion has been introduced by considering gray-level values of a picture as the altitude of an imaginary relief. Most efficient implementations are based on immersion simulations and rely on hierarchical queues [5].

We adopt the modified watershed algorithm for region growing, and set the priority measure to the absolute difference between the gray value of the pixel to be merged and the mean gray value of the catchment pixels.

2.3. Building-up the probability map

Fig. 2 shows the procedure for building up the probability map. With the motion field estimated in a forward direction, we can build up the probability map by counting the number of region pixels inside a diagonal window centered at the motion-compensated location. Here, the diagonal window has weighting parameters with the maximum value at the center pixel. Let us assume the

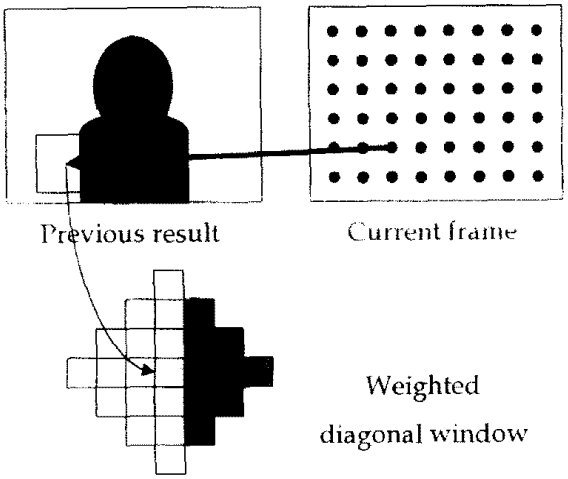


Fig. 2. Procedure for building-up the probability map.

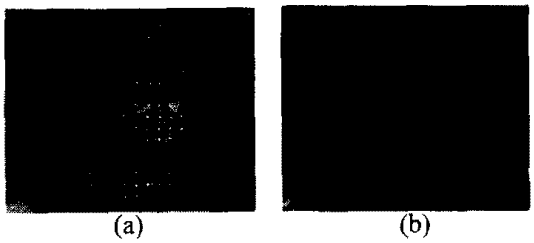
motion vector at (x, y) is denoted as (mv_x, mv_y). Then the probability value at (x, y) is defined as follows.

$$p(x,y)=\frac{\sum_{k,l} w_{k,l}\cdot M_{n-1}(x+mv_x+k,\,y+mv_y+l)}{\sum_{k,l} w_{k,l}} \qquad (1)$$

and

$$w_{k,l}=Max(5-|k|-|l|,0),\quad -5\le k,l\le 5 \qquad (2)$$

where $\{w_{k,l}\}$ are weighting parameters assigned to the diagonal window, and M_{n-1} designates the segmentation result of the previous frame, i.e., $M_{n-1}(x, y) = 1$ if (x, y) is inside the object, $M_{n-1}(x, y)=0$, otherwise. Then $p(x, y)$ represents the probability that a motion-compensated pixel



(a) (b)

Fig. 3. Miss America #90; (a) motion vectors and (b) probability map (Uncertain area where the probability value is between 0 and 1 is highlightened.)

(x, y) locates inside the object.

Fig. 3 demonstrates motion vectors and the resultant probability map for a frame of the *Miss America* sequence. For these simulation results, the conventional block match algorithm is used for motion estimation.

2.3. Region decision

Fig. 4 shows the overall flow diagram of the region decision step. Let us define the region probability value as follows.

$$reg\_prob(R_i) = \frac{\sum_{(x_k, y_k) \in R_i} p(x_k, y_k)}{\sum_{(x_k, y_k) \in R_i} 1}, \quad (3)$$

where R_i denotes the i-th segmented region. Then this *reg_prob* represents the probability that the corresponding region locates inside the object. After $reg\_prob(R_i)$ are calculated for all i, regions are classified into 3 categories, i.e., "inside", "outside", and "uncertain" regions, by comparing region probability values with the predetermined threshold value. After region classification, only "uncertain" regions are repeated at the lower hierarchies for further spatial segmentation.

At the final hierarchy, if there still exist any uncertain regions, they are mapped using the motion vector similarity. To do this, first, we perform morphological dilation of a 7x7 structural element (SE) on unmapped areas. In a dilated area, there are three kind of regions, i.e., inside, outside and unmapped regions. To represent the motion vector for each region, mean motion vectors, mv_i, mv_o and mv_u, are calculated. Then, if the inner product of mean motion vectors, $mv_i \cdot mv_u$ is larger than $mv_o \cdot mv_u$, the unmapped region is classified into an inside region, otherwise, it is classified into an outside region. By doing this, every region is categorized to one of two certain regions, i.e., inside and outside regions.

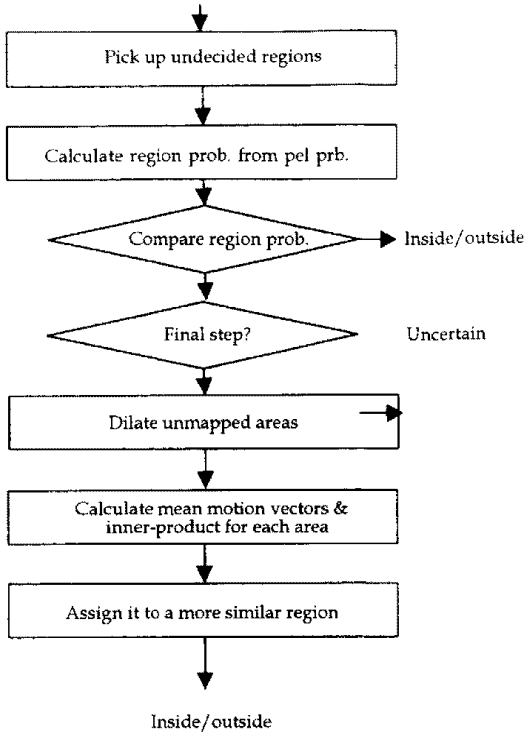


Fig. 4. Flow diagram of the region decision step.

2.4. Region refinement

A segmentation region can include some falsely merged regions. In tracking an object, falsely merged regions must be handled very carefully because they can introduce very severe performance degradation. Therefore, we perform a kind of post-processing in this stage. By using the probability map and the change detection mask, we force an outside pixel to inside if its probability indicates surely inside and its gray level is not changed much.

3. Simulation results

Some simulation results are depicted in Figs. 5, 6, and

7 for head-and-shoulder sequences; *Miss America*, *Clair*, and *Akiyo*. Here, the goal is to tracking the human body by assuming that the segmentation result for the intra-frame (0th frame) is given. All sequences have been tested at a frame rate of 10Hz in the QCIF (176x144) format. The resulting object masks are subjectively evaluated because the true masks are unknown. The simulation results demonstrate that the proposed algorithm fulfils a object-tracking task well.

4. Conclusions

We have dealt with a semantic video object-tracking problem. Because of non-homogeneity of a video object, we propose a hierarchical region-based tracking algorithm by building up the probability map from a forwardly estimated motion field. Consequently, this tracking algorithm can be used to achieve an automatic or semi-automatic creation of VOP.

The proposed segmentation scheme performs well for the scenes with moderate complexity. However, more complex scenes with shrinking or zooming objects can not be handled by the current approach, which is based on a block-matching algorithm for motion estimation. To solve this deficiency, further work is required.

5. References

[1] C. Gu and M. C. Lee, "Semantic video object segmentation and tracking using mathematical morphology and perspective motion model," *International Conf. on Image Processing*, pp. 514-517, 1997.

[2] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences," *International Conf. on Acoustics, Speech, and Signal Processing*, pp. 2657-2660, 1997.

[3] A. Steudel and M. Glesner, "Object tracking in video sequences with fuzzy contour refinement," *Workshop on Image Analysis for Multimedia Interactive Services*, pp. 33-38, June 1997.

[4] S. Kruse, "A snake-based tool for VOP-creation," *Picture Coding Symposium*, pp. 597-601, 1997.

[5] P. Salembier and M. Pardas, "Hierarchical morphological segmentation for image sequence coding," *IEEE Trans. Image Proc.*, vol. 3, no. 5, pp. 639-651, Sept. 1994.

[6] S. W. Lee and S. D. Kim, "Scene segmentation using a combined criterion of motion and intensity," *Optical Engineering*, vol. 36, no. 8, pp. 2346-2352, Aug. 1997.

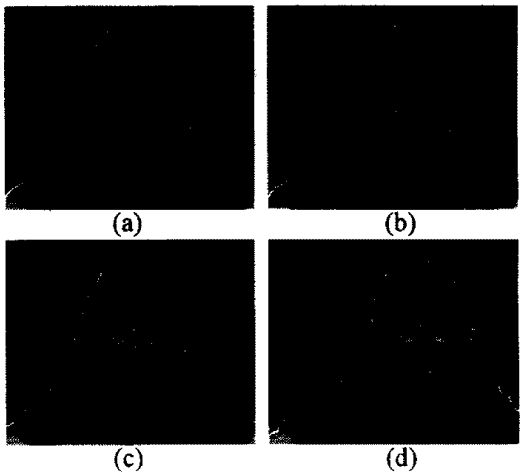


Fig. 5. Results for the *Miss America* sequence: (a) intra, (b) 3rd, (c) 30th, and (d) 90th frames.

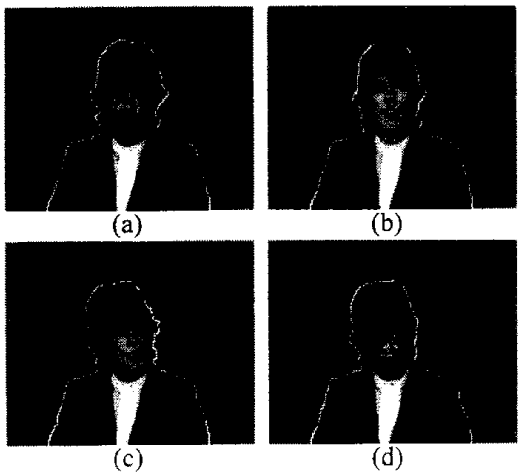


Fig. 6. Results for the *Clair* sequence; (a) intra, (b) 3rd, (c) 30th, and (d) 90th frames.

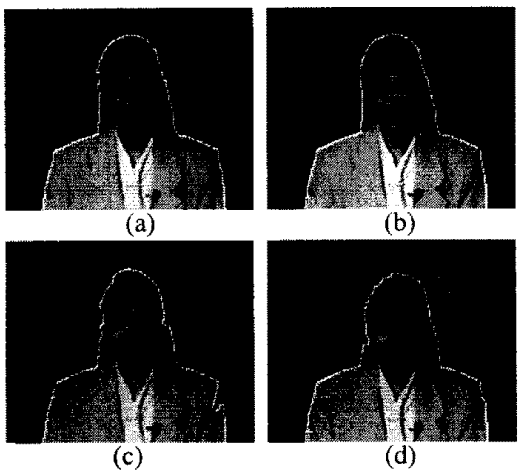


Fig. 7. Results for the *Akiyo* sequence; (a) intra, (b) 3rd, (c) 30th, and (d) 90th frames.