VIDEO OBJECT SEGMENTATION BASED ON OBJECT ENHANCEMENT AND REGION MERGING

Ken Ryan and Aishy Amer

Concordia University,
Electrical & Computer Engineering
Montreal, Quebec, Canada
{kenne_ry, amer}@ece.concordia.ca

Langis Gagnon

Computer Research Institute of Montreal (CRIM)
Montreal, Quebec, Canada
langis.gagnon@crim.ca

ABSTRACT

This paper proposes a number of improvements to existing work in off line video object segmentation. Object color and motion variance, and histogram-based merging are used to improve the initial segmentation. Segmentation quality measures taken from throughout the clip are used to enhance video objects. Cumulative histogram-based merging, occlusion handling, and island detection are used to help group regions into meaningful objects. Objective and subjective tests were performed on a set of standard video test sequences which demonstrate improved accuracy and greater success in identifying the real objects in a video clip compared to the reference method.

1. INTRODUCTION

Content-based representation of video sequences for applications such as MPEG-4 and MPEG-7 coding is an area of growing interest in video processing [1, 2, 3]. One of the key steps to content-based representation is segmenting the video into a meaningful set of objects.

Video object segmentation requires a consistency of object labeling throughout a clip. For this reason, many video segmentation approaches involve segmenting the first frame and tracking the segments through the rest of the clip [4, 5].

In this paper, we focus on off line, unsupervised segmentation methods which use multiple features. In [4], a maximum a posteriori (MAP) framework is proposed. They assign weights to color and motion terms, which are adjusted at every pixel. They also model the spatial pdf of each region in order to impose temporal consistency. A major drawback of [4] is that the number of objects must be known beforehand.

A slightly different approach is employed by [5]. Here, initial segmentation with the K-Means with Connectivity Constraint (KMCC) algorithm combines color, motion and spatial information to estimate the number of regions and cluster pixels into their best fit region. The first frame segmentation is then enhanced using a histogram-based Bayesian edge reclassification. Tracking is also performed using a Bayesian approach, where disputed pixels (chosen based on color difference) are re-classified using the histograms of objects in the previous frame. After tracking is complete, regions are merged into real objects based on their trajectories. Bilinear motion parameters are estimated for each region, and regions that are spatio-temporal neighbors, and whose motion can be well modeled by the same set of bilinear parameters, are merged. Merged regions are then labeled as background or foreground based on their consistency with global motion. The main drawback of [5] is that objects with little motion or complex motion cannot be well segmented.

Other authors use color and motion to segment objects in the first frame, which are then tracked by using their estimated motion to predict their location in the next frame [6].

The remainder of this paper is organized as follows. Section 2 proposes our improvements to existing work. Results are presented in section 3, and a conclusion in section 4.

2. PROPOSED APPROACH

Our approach to better segmentation uses the following steps:
1) Initial segmentation: We include motion and color variances in the distance function of the KMCC algorithm, and add histogram distance-based merging (Sec. 2.1).
2) Histogram-based object enhancement: We take a set of segmentation measures while tracking objects to improve the accuracy of object boundaries (Sec. 2.2).
3) Post-tracking merging: Regions are merged based on cumulative histograms gathered over the entire clip (Sec. 2.3).
4) Trajectory-based merging: We handle partial occlusion and deal with isolated regions (Sec. 2.4).

2.1. Initial Segmentation

We propose to include variance information about each region when classifying pixels with the KMCC algorithm. After the

This work was supported, in part, by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

initial centers are estimated, the feature variance of each region is calculated, and pixels are classified according to their distance from the center of each feature divided by the variance. So we propose the distance function

$$D_{KMCC} = \frac{\left\|\mathbf{C}(p)-\overline{\mathbf{C}}_{R_i}\right\|}{\sigma^2_{R_i,C}} + \lambda_1 \frac{\left\|\mathbf{M}(p)-\overline{\mathbf{M}}_{R_i}\right\|}{\sigma^2_{R_i,M}} + \lambda_2 \frac{A_{R_i}}{\overline{A}} \left\|\mathbf{p}-\overline{\mathbf{S}}_{R_i}\right\| \quad (1)$$

where $\overline{\mathbf{C}}_{R_i}$, $\overline{\mathbf{M}}_{R_i}$, and $\overline{\mathbf{S}}_{R_i}$ are the color, motion and spatial centers of region R_i, respectively. $\mathbf{C}(\mathbf{p})$ and $\mathbf{M}(\mathbf{p})$ are the color and motion vector values for image point $\mathbf{p}$. A_{R_i} is the area of region R_i in pixels, and $\overline{A}$ is the average region area. $\sigma^2_{R_i,C}$ and $\sigma^2_{R_i,M}$ are the color and motion variances of region R_i. λ_1 and λ_2 are regularization parameters [5].

Classifying pixels in this way is more accurate than using only distances from region centers as in [5], since more information about the distribution of each region is being utilized. Also, this method divides the image into a smaller number of more complex regions, which reduces the over-segmentation normally associated with the KMCC algorithm. Reducing the over-segmentation of the first frame decreases the chances for error in later stages of the algorithm.

To improve the robustness of the initial segmentation, we examine the regions at the end of each iteration of the KMCC. If the algorithm converges to less than two regions, R_i, R_j, that meet Eq. 2, indicating under segmentation, the entire process resets and the original KMCC is used.

$$A_{R_i} > \alpha \times X \times Y \quad (2)$$

where X and Y are the image width and height, and α set experimentally to 0.02.

Furthermore, we propose a merging stage based on color-histogram distance and motion distance between region centers. First, a color histogram is calculated for each region, and the χ^2 histogram distance between each pair of neighboring regions is calculated as follows:

$$\forall R_i, R_j \in P_1, \quad \chi^2(H_{R_i}, H_{R_j}) = \sum_b \frac{(H_{R_i}(b) - H_{R_j}(b))}{(H_{R_i}(b) + H_{R_j}(b))} \quad (3)$$

where P_1 is the set of all pairs of neighboring regions (R_i, R_j) in the first frame, H_{R_i} and H_{R_j} are the histograms of R_i and R_j, and b is the histogram bin. After the distances have been calculated, all neighboring regions as in Eqs. 4 and 5 are merged.

$$\chi^2(H_{R_i}, H_{R_j}) < \beta \times S_{hist} \quad (4)$$

$$\left\|\overline{\mathbf{M}}_{R_i} - \overline{\mathbf{M}}_{R_j}\right\| < 2 \times max(\sigma^2_{R_i,M}, \sigma^2_{R_j,M}) \quad (5)$$

where S_{hist} is the histogram size and β experimentally set to 1.3. Second we re-evaluate the region motion centers and histograms and re-determine neighbor relationships. The merging continues until no more regions meet Eqs. 4 and 5.

By reducing over-segmentation compared with [5], we identify and merge regions in the first frame that better represent the true video objects.

2.2. Histogram-Based Object Enhancement

During object tracking, we measure the segmentation quality as 1) color homogeneity of the region [5] defined as the average of the MAP probabilities of every pixel in the region, 2) color contrast across the object boundary [7], and 3) motion contrast across the object boundary [7] for each region in each frame. We then examine these segmentation measures and object movements to determine for which frames we will enhance which objects after the objects have been tracked through the entire clip.

For a given object, most variation in object segmentation quality between frames is due to movement. Therefore, we are here mainly interested in moving objects. To this end, we examine the trajectories of all objects in the entire video clip and choose which ones to enhance as follows.

The (x, y) coordinates of each object's center in each frame are used to calculate the maximum displacement of every object in the clip. The displacement is taken with respect to the first frame. Objects whose maximum displacement is above a certain threshold are considered to have undergone significant motion and are candidates for enhancement (Eq. 6).

$$\begin{aligned} &\forall R_i \in I \quad \text{and} \quad t = \frac{\sqrt{\overline{A}_{R_i}/\pi}}{2} \\ &\mathbf{\Delta D}_{R_i,max} > t : \text{enhance} \quad R_i \\ &\mathbf{\Delta D}_{R_i,max} \leq t : \text{keep} \quad R_i \end{aligned} \quad (6)$$

where $\mathbf{\Delta D_{R_i,max}}$ is the maximum displacement of R_i over the entire clip I and $\overline{A}_{R_i}$ is the size of R_i averaged over I.

Once we have chosen which objects to enhance, we examine their segmentation quality measures for each frame and enhance objects according to the following rules:
1) If an object's color homogeneity in a given frame is below that same object's average color homogeneity for all frames, this indicates that pixels belonging outside the object have been classified inside the object in this frame. In this case, pixels within the object and close to the boundary will be marked as disputed and re-classified.
2) High color homogeneity with below average color contrast indicates that pixels belonging inside the object have been classified outside. In this case, pixels close to the boundary but outside the object will be re-classified.
3) High color homogeneity with high color contrast indicates a good segmentation. Nothing will be done.

We re-classify pixels through a Bayesian approach using histograms from key frames of the clip to determine the MAP probability of each disputed pixel. Out of every five frames, the frame with the highest homogeneity and contrast is a key frame. The disputed pixels in each frame are re-assigned based on each object's nearest key frame histogram.

After re-assigning pixels, we perform an error check based on the assumption that object enhancements should not result in drastic changes in object size. We measure the size of the object, and if it has increased in size by more than 200% or

decreased by more than 70%, the test fails. If the object's motion contrast has decreased, the error check fails as well. Due to the use of block-based motion estimation, motion contrast is not effective for locating small inaccuracies in object boundaries, so it was not used in selecting the frames needing improvement or the key frames. However, a decrease in motion contrast does indicate a significant reduction in boundary accuracy, making motion contrast an effective measure for error checking. If the enhanced object fails either of the error checks, the enhancement is rejected, otherwise it is accepted.

This enhancement stage improves the boundaries of tracked objects over that of [5]. This also allows more accurate motion parameters to be estimated for each object, improving the performance of the trajectory-based merging stage (Sec. 2.4).

2.3. Post-Tracking Region Merging

Post-tracking region merging simplifies the trajectory-based merging stage (Sec. 2.4). This is desirable, because trajectory-based merging can fail when an object's motion is too complicated (deformation or articulated motion), or when accurate motion vectors are not available (e.g., when objects are highly uniform in color).

Color histograms are used to merge regions which are spatio-temporal neighbors. Here we use cumulative histograms calculated from an object's pixels taken over all frames in the clip. Compared with histograms computed for an object in a single frame, cumulative histograms are less sensitive to noise, inaccurate object boundaries for particular frames, changing illumination, and occlusion. For example, an object with lighting that varies across it's surface in the first frame could be segmented into two regions, but as the object moves these illumination differences could even out, and the two halves of the object can be merged. As with the first frame histogram-based merging (Sec. 2.1), the χ^2 histogram distance (Eq. 3) is used to select regions to merge. This stage improves the segmentation of objects with complex motion that present problems for [5].

2.4. Trajectory-Based Merging

We propose a trajectory-based merging that accounts for high occlusion of the background. The trajectory-based merging stage of [5] only examines regions which are spatio-temporal neighbors. However, since region connectivity is enforced during the initial segmentation with the KMCC algorithm, it is possible for the background to be initially segmented into multiple regions that are not spatio-temporal neighbors. One example is when there is a large object, extending from top to bottom in the middle of a frame. In these cases, the video cannot be segmented correctly without merging these non-neighboring background regions. To account for this, any region that contains a corner point, (0,0), (X-1,0), (0, Y-1), (X-1,Y-1), of a frame is considered to be a potential background region, and will be treated as a spatio-temporal neighbor of all other potential background regions in the clip for the purposes of trajectory-based merging. Note that the trajectories are still used to decide if to merge these potential background regions. Thus foreground objects with corner points can still be correctly identified (e.g., Fig. 1). With this change of the spatio-temporal neighbor criteria, we are able to correctly segment the disconnected pieces of the background, while still enforcing connectivity of all other objects. Furthermore, after the trajectory-based merging is finished, any island regions (those with only one spatio-temporal neighbor which is not a potential background region) are merged into their surrounding object

3. RESULTS

We applied our method on the videos Suzie, Miss, Harbor, Mobile, Tennis, Road, Quizshow, and Basketball (both are from the MPEG-7 Content Set). Compared to the reference method, we obtained significantly improved results for Suzie, Miss, Harbor, Mobile, Quizshow, and Basketball, and similar results for Tennis and Road. Here we present results for Suzie (object deformation, Fig. 1), Harbor (global motion, high background occlusion, Fig. 2), Mobile (global motion, highly textured, Fig. 3), and Basketball (fast global motion, Fig. 4). As can be seen, the reference [6] has difficulties with these video characteristics. The proposed combination of reduced over-segmentation of the first frame, object enhancement, and histogram-based merging, significantly improve the segmentation.

Fig. 1. Frames 1, 50, 100 and 150 of the Suzie clip. Proposed (top) and reference (bottom) method.

Fig. 2. Frames 1, 10, 20 and 30 of the Harbor clip (with global motion). Proposed (top) and reference (bottom) method.

Objective measures [7] confirm the subjective quality. See Fig. 5 (cf. Figs. 1, 3, 4) where lower normalized values of

Fig. 3. Frames 1, 35, 70 and 100 of Mobile (global motion). Proposed (top) and reference (bottom) method.

Fig. 4. Frames 1, 5, 15 and 20 of Basketball (global motion). Proposed (top) and reference (bottom) method.

the color and motion contrast measures indicate more accurate segmentation and lower values of the histogram distance means the object histogram is more stable over the clip, indicating better object tracking.

4. CONCLUSION

A number of innovations for video object segmentation have been proposed. These include reducing over-segmentation of the first frame, using segmentation quality measures to enhance object accuracy, merging tracked regions based on histograms, and accounting for occlusion. Experimental results have been presented which demonstrate improved performance over the reference method. Future work includes changes to the segmentation and tracking to improve handling of events such as occlusion and object splitting.

5. REFERENCES

[1] B. Furht and O. Marques, *Handbook of Video Databases: Design and Applications*, CRC Press, 2004.

[2] L. Gagnon, "R&D status of ERIC-7 and MADIS: two systems for MPEG-7 indexing/search of audio-visual content," in *Proc. SPIE Conference on Multimedia Systems and Applications VIII (SPIE #6015)*, October 2005, pp. 341–352.

[3] A. Amer and C. Regazzoni, "Introduction to the special issue on video object processing for surveillance applications," *Real-Time Imaging*, vol. 11, pp. 1–5, 2005.

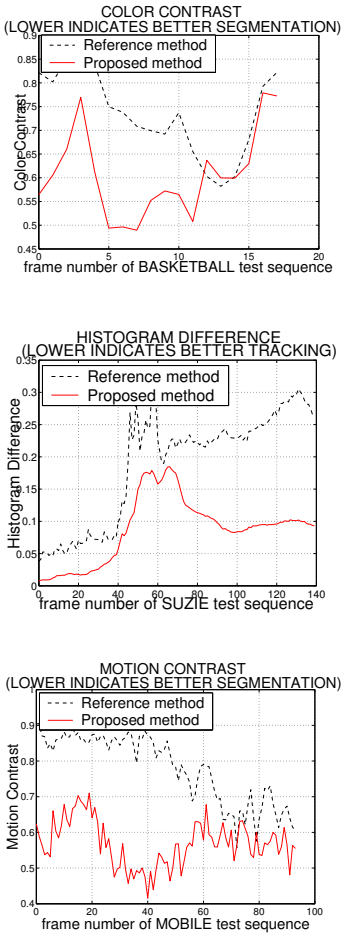

Fig. 5. Sample Objective Results (see Figs. 1, 3, 4).

[4] S. Khan and M. Shah, "Object based segmentation of video using color, motion and spatial information," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, December 2001, vol. 2, pp. II–746 – II–751.

[5] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, "Video object segmentation using bayes-based temporal tracking and trajectory-based region merging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 782 – 795, June 2004.

[6] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 539 – 546, September 1998.

[7] C. Erdem, B. Sankur, and A. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Transactions on Image Processing*, vol. 13, no. 7, July 2004.