PERFORMANCE EVALUATION METRICS FOR OBJECT-BASED VIDEO SEGMENTATION*

Çigdem Eroglu Erdem and Bülent Sankur
Department of Electrical and Electronics Engineering
Bogaziçi University, 80815, Bebek, Istanbul, Turkey
erogluc@boun.edu.tr, sankur@tsi.enst.fr

ABSTRACT

In this paper we investigate performance metrics for quantitative evaluation of object-based video segmentation algorithms. The metrics address the case when ground-truth video object planes are available. The proposed metrics are used to evaluate three essentially different approaches for video segmentation, i.e., an edge-based [1], a motion clustering based [2], and a total feature vector clustering based [3] algorithm.

1. INTRODUCTION

The emerging standard for multimedia communications is MPEG-4. MPEG-4 provides standardized ways to encode video and audio objects, and the scene description, which indicates how the objects are organized in a scene [4,5]. One of the most important innovations that MPEG-4 brings is the capability of manipulating the individual objects in an image sequence. However, in MPEG-4, the decomposition or spatio-temporal segmentation of a scene into "objects" is not standardized. Therefore, many object-based segmentation algorithms have been proposed in the literature recently [1-3,6]. These algorithms use different sets of techniques and result in different performance. Furthermore, their performance varies with different image sequences depending on the content of the sequence. Their comparative assessment is often based upon subjective judgement of a few frames in known sequences. In this context, an objective spatio-temporal segmentation measure would be a very useful basis of comparison of algorithmic performance. Furthermore, the time and motion consistency of object-based video segmentation algorithms must also be put into evidence. Although there are many approaches, e.g., [7-13], that try to evaluate image segmentation quantitatively, they are for still images, and not for video sequences. Thus, in this work we propose performance metrics to quantitatively evaluate the empirical results of spatio-temporal video segmentation methods.

We comparatively evaluate three different approaches for object-based video segmentation, namely, 1) A region-based parametric motion segmentation algorithm which uses both the estimated dense motion field anthe color segmentation for a region based parametric motion segmentation [2]; 2) An algorithm that finds the best match of the binary model of the object(s) in the edge map of subsequent frames [1]; 3) An interactive algorithm which finds the initial regions using fuzzy c-means clustering of the feature vectors of each pixel [3]. These three approaches are selected specifically as they differ fundamentally in their algorithmic details.

The proposed evaluation metrics are introduced in Section 2. The experimental results using both synthetically generated and real video sequences are presented in Section 3. Finally, in Section 4, concluding remarks are presented.

2. EVALUATION METRICS FOR OBJECT-BASED SEGMENTATION

For still image segmentation, many algorithms have been proposed to evaluate their performance in the absence of ground-truth images [7,8,10-12] and when the ground-truth is available [7,9,12-14]. The ground-truth for a still image defines the support of the object(s), which corresponds to the location, size, and the shape of the object (but not the inner texture). For video segmentation, the ground-truth must be defined differently for intended applications. If we directly extend the definition of ground-truth for still images to video, then the ground-truth for an image sequence consists of the support of the object(s) in the subsequent frames of the sequence. Some MPEG-4 applications require the motion of the objects, which can be in the form of dense motion field, motion along the object boundaries or affine motion parameters for each object. Therefore, in the final analysis the comparison metric should reflect the fidelity of the spatio-temporal evolution of the segmented objects. In the following we introduce four possible quantitative metrics for evaluation of object-based segmentation. We assume that the frames contain N pixels and M objects. When the ground truth (spatial support and motion information of the objects) is available, following distance measures between the ground truth and the segmented video object planes can be considered:

1) **Misclassification Penalty:** The misclassified pixel measure used for still image segmentation is adapted to compare the estimated segmentation map of each frame to

* This work was supported by TÜBITAK-BAYG (Scientific and technical Research council of Turkey) and Bogaziçi University Research Fund under project number 99A203.

its ground-truth version on an object-by-object basis. Misclassified pixels between ground-truth objects and test objects can be penalized as a function of their distance from the ground-truth boundary as discussed in [13] and [14]. Another approach presented here, uses the Chamfer distance as follows:

Let the object set of the ground-truth and segmented image be denoted as $G=\{g_i, i=1,\ldots,M\}$ and $S=\{s_i, i=1,\ldots,M\}$. Let the label functions $L_G^t(k,l)\in\{1,\ldots,M\}$ and $L_S^t(k,l)\in\{1,\ldots,M\}$ denote the objects to which the pixel at location (k,l) belongs to at time *t*, in the ground-truth and segmented frames, respectively. We define the indicator function $I^t(k,l)$ at pixel location (k,l) at frame *t* as:

$$I^t(k,l)=\begin{cases}1 & if \quad L_G^t(k,l)\neq L_S^t(k,l)\\ 0 & if \quad L_G^t(k,l)=L_S^t(k,l)\end{cases}.$$

The discrepancy between two objects g_i and s_i due to misclassified pixels can be calculated as:

$$0\le D_i^t=\frac{\sum_{(k,l)\in(s_i\cup g_i)} I^t(k,l)\,w_{g_i}(k,l)}{\sum_{\mathbf{all}(k,l)} w_{g_i}(k,l)}<1$$

where $w_{g_i}(k,l)$ is the modified Chamfer distance of the pixel from the boundary of g_i. The modified Chamfer distance can be found using Chamfer 3-4 mask [0] and changing the labels of the pixels on the boundary as 1.

The segmentation error averaged over all objects becomes:

$$D^t=\frac{1}{M}\sum_{i=1,\ldots,M} D_i^t\ .$$

Several other variations can be considered such as weighting the object errors proportionally to their size or importance with respect to other criteria before averaging. Finally, the spatio-temporal distortion reflecting the misclassification penalty (DP) over T frames can be calculated as:

$$DP=\frac{1}{T}\sum_{t=1}^{T} f(D^t),$$

where $f(D^t)$ is some error function, e.g, $f(D^t)=(D^t)^2$ or $f(D^t)=\max_t\{D^1,\ldots,D^T\}$.

2) Shape Penalty (DS): A discriminative shape metric is the turning angle function (TAF) of the object boundaries. The distance between the *M* object shapes at time *t* can be calculated from:

$$0\le DS^t=\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\sum_{s=1}^{P}\left|\Theta_{g_i}^t(s)-\Theta_{s_i}^t(s)\right|}{P*2p}\right)\le 1,$$

where $\Theta_{s_i}^t(.)$ and $\Theta_{g_i}^t(.)$ denote the turning angle functions (TAF) of the ground-truth object i at time *t*, and the corresponding segmented video object plane (VOP) respectively. P denotes the number of contour elements. We extend the calculation of the TAF from polygons [15] to any arbitrarily shaped object as follows. If the shapes are convex, we take the boundary pixels of each object and extract the signature of the shape by representing the pixel locations in the (r,q) domain, where r denotes the magnitude of the vector connecting the boundary pixel and the centroid of the boundary, and q denotes the angle of this connecting vector from the x-axis. Then, the signatures of the shapes to be compared are resampled such that equal number of samples is obtained at uniformly spaced and identical q values. Then, we calculate the pointing vectors, which point from one boundary pixel to the next pixel and add up the rotation angle between each successive pointing vector to find the TAF of the object shape. It is also possible to find the rotation angle between the shapes by shifting the TAF of the test object with increasing amounts and taking the angle at which the distance between the two TAF's is minimized. Furthermore, by shifting the signature of one of the objects by the estimated rotation angle and averaging the ratio of the signatures belonging to two objects, we can estimate the scale difference between the shapes. If the shapes are not convex, the resampling of the boundary is done by fitting a continuous curve to the boundary pixels and taking samples at equal intervals on this curve.

The discrepancy between two objects can be weighted by object area before averaging. The shape distortion of a sequence can be computed as in Case 1, that is, some weighted averaging over the interval *T*:

$$DS=\frac{1}{T}\sum_{t=1}^{T} f(DS^t).$$

3) Motion Penalty: Let $M_{g_i}^t$ and $M_{s_i}^t$ denote the motion information at frame *t* of the ground-truth and segmented object, respectively. M itself can be parametric, such as the affine motion parameters of the object, or it may represent the motion field vectors at each pixel or a neighbourhood of pixels. For video objects, it is more significant to compare the consistency of motion trajectories over time. Thus, one can have:

$$D_i=D(M_{g_i}^t, M_{s_i}^t),$$

where $D(.,.)$ is some distance function between two time evolution surfaces described by $M_{g_i}^t$ and $M_{s_i}^t$. For example, one can consider:

$$0\le D_i=\frac{1}{T}\sum_{t=1}^{T}\frac{\left\|M_{g_i}^t-M_{s_i}^t\right\|}{\left\|M_{g_i}^t\right\|+\left\|M_{s_i}^t\right\|}\le 1.$$

Finally, the object trajectory scores can be averaged over all objects:

$$DM=\frac{1}{M}\sum_{i=1}^{M} f(D_i).$$

4) Combined Penalty (CP): A weighted averaging of more than one criterion can be used to attain a more comprehensive quality measure of the spatio-temporal segmentation with respect to its ground-truth:

$$CP = \alpha\, DP + \beta\, DS + \gamma DM + C_{\max},$$

where the parameters α, β, and γ can be adjusted to give different weights to misclassification penalty, shape penalty and motion penalty, respectively, and $C_{\max}$ is a penalty term for the missing objects.

3. QUANTITATIVE EVALUATION OF THE VIDEO SEGMENTATION METHODS

The metrics proposed in Section 2 are tested on a synthetically generated sequence and a real sequence, named 'taxi in the garden' and 'Hamburg Taxi,' respectively. A sample frame from each sequence is shown in Figure 1. In order to generate the synthetic sequence, the white car extracted from the 'Hamburg Taxi' sequence is mounted on the first frame of the 'flower garden' sequence and the car is moved with a velocity $(V_x, V_y) = (3, -1)$. The segmentation results for the second frame of the 'taxi in the garden' sequence using the algorithms discussed in Section 3 are given in Figure 2, while the performance metrics proposed in Section 2 are evaluated and plotted for 15 frames in Figure 3. The segmentation results for the fourth frame of the 'Hamburg Taxi' sequence are given in Figure 4. The results for both sequences are summarized in Table 1. For the 'taxi in the garden' sequence, Altunbasak-Eren-Tekalp method performs best in terms of the misclassification penalty and the motion penalty. However, its shape penalty score is high. This is due to the busy background of the flower garden sequence. For the same reason, the shape penalty and misclassification penalty scores of Castagno-Ebrahimi-Kunt method are also high. In terms of motion penalty, the Meier-Ngan method performs the worst since it finds the motion of the edge-model of the car by searching in the edge map of the next frame, which is quite crowded because of the flowers. In the 'Hamburg Taxi' sequence which contains multiple objects, the relative perfomance results do not change much. Although the quantitative performances can change with the parameters chosen in the individual algorithms and the test sequences, Altunbasak-Eren-Tekalp performed better than the other algorithms in our experiments and the other two algorithms closely follow it.

Figure 1. (a) The second frame of the synthetically generated sequence 'Taxi in the garden'. (b)The fourth frame of the 'Hamburg Taxi' sequence.

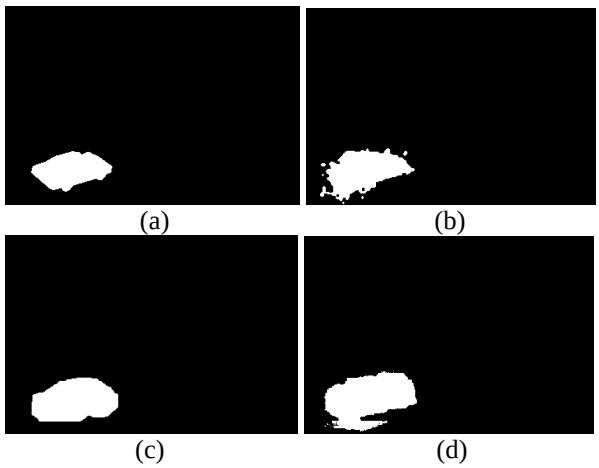

Figure 2. (a) The ground truth VOP for frame 2 of the 'Taxi in the garden' sequence. (b) The result of Altunbasak-Eren-Tekalp method (M1). (c) The result of Meier-Ngan method (M2). (d) The result of Castagno-Ebrahimi-Kunt method (M3).

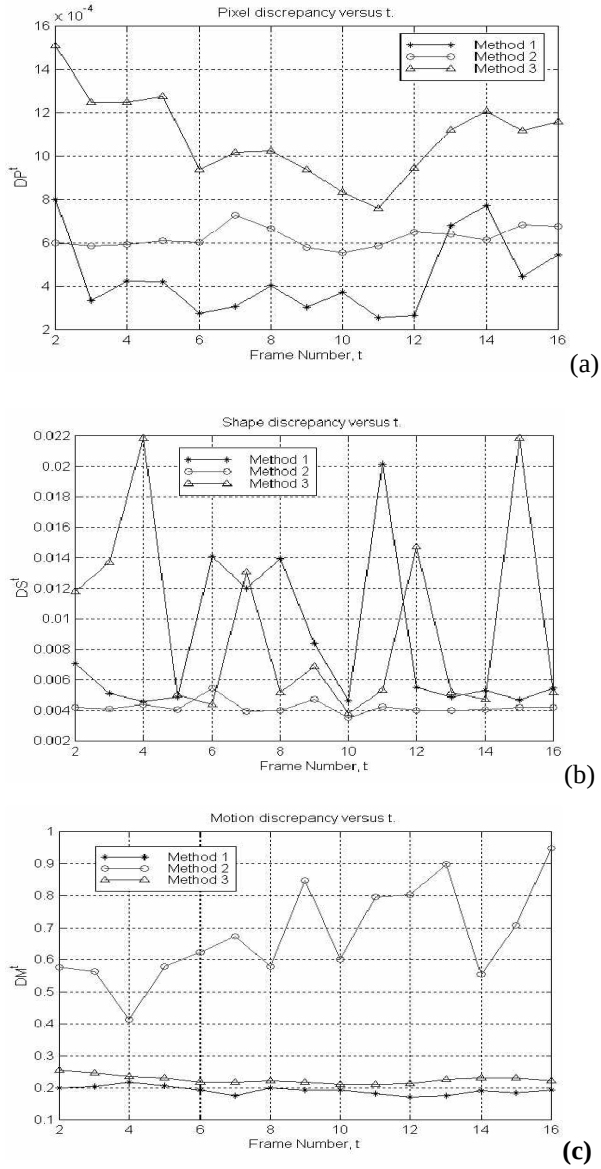

Figure 3. (a) The weighted misclassification penalty(DP), (b) The shape penalty (DS), (c) The motion penalty (DM) versus frame number plots for the 'Taxi in the garden' sequence. The ground truth segmented sequence is found by marking the support of the three moving vehicles by hand.

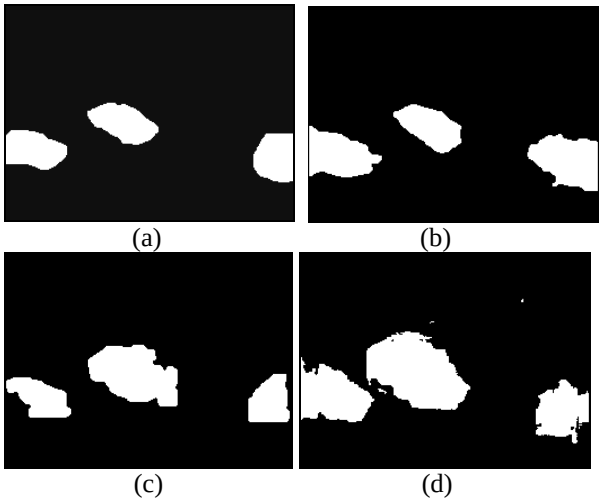

Figure 4. (a) The ground truth for the fourth frame of the 'Hamburg Taxi' sequence. (b) The result of Altunbasak-Eren-Tekalp method. (c) The result of Meier-Ngan method. (d) The result of Castagno-Ebrahimi-Kunt method.

	Taxi in the garden sequence			Hamburg Taxi sequence		
Criteria	M1	M2	M3	M1	M2	M3
DP	0.16	0.33	1	0.72	0.41	1
DS	0.69	0.15	1	0.83	0.98	1
DM	0.08	1	0.10	0.15	1	0.24
CP	0.31	0.49	0.7	0.57	0.79	0.74

Table 1. The performances of the three object-based video segmentation methods averaged over all frames. The scores are normalized such that the worst score for a metric among the three methods is assigned to 1. CP is found by giving equal weights to DP, DS, and DM.

4. CONCLUSIONS

We proposed three performance evaluation metrics, based on weighted misclassification penalty, shape penalty and motion penalty, for evaluating the performances of object-based video segmentation algorithms in the presence of ground-truth segmentation. The proposed metrics have been tested on real and synthetic image sequences and have been shown to agree with the subjective quality assesments of the three leading segmentation algorithms found in the literature.

References

[1] T. Meier and K. N. Ngan, "Automatic Segmentation of Moving Objects for Video Object Plane Generation", *IEEE Trans. On Circuits and Systems for Video Technology*, Vol. 8, No. 5, pp. 525-538, 1998.

[2] Y. Altunbasak, P. Erhan Eren, and A. Murat Tekalp, "Region-Based Parametric Motion Segmentation Using Color Information", *Graphical Models and Image Processing*, Vol. 60, No. 1, January, pp.13-23, 1998.

[3] R. Castagno, T. Ebrahimi, and M. Kunt, "Video Segmentation Based on Multiple Features for Interactive Multimedia Applications", *IEEE Trans. On Circuits and Systems for Video Technology*, Vol.8, No.5, pp. 562-571, 1998.

[4] L. Chiariglione, "The MPEG-4 Standard", *Journal of China Institute of Communications*, pp.54-67, September 1998.

[5] R. Koenen, F. Pereira, and L. Chiariglione, "MPEG-4: Context and Objectives", *Signal Processing: Image Communication*, Vol.9, pp. 295-304, 1997.

[6] F. Moscheni, S. Bhattacharjee, and Murat Kunt, "Spatiotemporal Segmentation Based on Region Merging", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No.9, pp.897-915, 1998.

[7] Y. J. Zhang, "A Survey on Evaluation Methods for Image Segmentation", *Pattern Recognition*, Vol.29, No.8, pp.1335-1346, 1996.

[8] M. Borsotti, P. Campdelli, and P. Schettini, "Quantitative Evaluation of Color Image Segmentation Results", *Pattern Recognition Letters*, Vol.19, pp.741-747, 1998.

[9] Y. J. Zhang, "Evaluation and Comparison of Different Segmentation Algorithms", *Pattern Recognition Letters*, Vol.18, pp.963-974, 1997.

[10] M. D. Levine, and A. M. Nazif, "Dynamic Measurement of Computer Generated Image Segmentations", *IEEE Trans. On Pattern Recognition and Machine Intelligence*, Vol.7, No.2, pp.155-164, 1985.

[11] G. Goehrig, L. Ledford, "Analysis of Image Segmentation Approaches with Emphasis on Performance Evaluation Criteria", *SPIE* Vol.252, pp.124- 129, 1980.

[12] G. Rees, P. Greenway, and D. Morray, "Metrics for Image Segmentation", *SPIE Conf. On Visual Information Processing VII*, Vol.3387, 1998.

[13] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, "Error Measures for Scene Segmentation", *Pattern Recognition*, Vol.9, pp. 217-231,1977.

[14] K. C. Strasters, "Three-dimensional Image Segmentation Using a Split, Merge and Group Approach", *Pattern Recognition Letters*, Vol.12, pp. 307-325, 1991.

[15] E. M. Arkin, L. P. Chew, D. P. Huttenlocker, K. Kedem, and J. Mitchell, "An efficient Computable Metric for Comparing Polygonal Shapes", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 13, No. 3, pp.209-216, 1991.