METRICS FOR PERFORMANCE EVALUATION OF VIDEO OBJECT SEGMENTATION AND TRACKING WITHOUT GROUND-TRUTH

Çiğdem Eroğlu Erdem, A. Murat Tekalp**, Bülent Sankur**

(*) Department of Electrical and Electronics Engineering
Boğaziçi University, İstanbul 80815, Turkey.
(**) Department of Electrical and Computer Engineering,
University of Rochester, Rochester, NY 14627, USA.

ABSTRACT

We present metrics to evaluate the performance of video object segmentation and tracking methods quantitatively when ground-truth segmentation maps are not available. The proposed metrics are based on the color and motion differences along the boundary of the estimated video object plane and the color histogram differences between the current object plane and its temporal neighbors. These metrics can be used to localize (spatially and/or temporally) regions where segmentation results are good or bad; or combined to yield a single numerical measure to indicate the goodness of the boundary segmentation and tracking results. Experimental results are presented to evaluate the segmentation map of the "Man" object in the "Hall Monitor" sequence both in terms of a single numerical measure, as well as localization of the good and bad segments of the boundary.

1. INTRODUCTION

Although a variety of algorithms have been proposed in the literature for object segmentation and tracking that address many applications, only a few methods for quantitative evaluation of their performances have been proposed [1, 2]. Furthermore, the performance evaluation metrics proposed in [1, 2] are useful if ground-truth segmentation maps are available. The aim of this work is to evaluate the performance of the object tracking and segmentation methods quantitatively, when ground-truth segmentation maps for each frame are not available. To this effect, we present metrics based on intra-frame and inter-frame color and motion features of the segmented video object planes. Often a single numerical measure is not sufficient to evaluate a segmentation/tracking result, since certain parts (spatially or temporally) of the object boundary are better segmented/tracked than others depending on the variation of color and motion features between the object and background regions. Hence, we also propose a localization of the good and bad segments of the object boundary using the proposed metrics.

2. COLOR METRICS

The proposed performance metrics using color features are based on the following assumptions which are true for most video sequences and are also assumed by many segmentation algorithms:

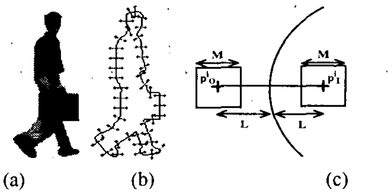

Fig. 1. (a) The video object plane for the 32^{th} frame of the "Hall monitor" sequence. (b) The boundary of the video object plane with the normal lines. (c) A close-up of a normal line drawn to the boundary. The two points 'just inside' and 'just outside' of the boundary are shown with symbols p^i_I and p^i_O, respectively.

1) Object boundaries coincide with color boundaries. 2) The color histogram of the object is stationary from frame to frame. 3) The color histogram of the background is different from the color histogram of the object. Note that the background and its color histogram are not restricted to be stationary from frame to frame. There are also no restrictions on the shape and rigidity of the segmented/tracked object.

Based on the above assumptions, we present two metrics for evaluating the fidelity of the segmented video object plane. The first color metric is based on the intra-frame color differences along the estimated object boundary and is presented in Section 2.1. The second color metric uses the inter-frame color histogram differences and is described in Section 2.2.

2.1. Intra-frame Color Differences Along the Boundary

In order to evaluate the performance of the tracking algorithm based on the first assumption above, the color of the pixels 'just inside' and 'just outside' of the estimated object boundary can be compared. In order to define 'just inside' and 'just outside', we draw short normal lines of length L to the estimated object boundary at equal intervals towards the inside and outside of the object. The points at the ends of these normal lines are marked as illustrated in Figure 1(b). The marked points are shown with plus signs. A closer look at one of these normal lines is given in Figure 1(c). We define the color difference metric calculated along the

This work was supported by Scientific and Technical Research Council of Turkey (TÜBİTAK-BAYG) and Boğaziçi University Research Fund under project 99A203.

boundary of the object in frame t as:

$$0 \leq d_{CB}(t) \;=\; 1 - \frac{1}{K_t}\sum_{i=1}^{K_t} d_{CB}(t;i) \leq 1, \qquad (1)$$

$$d_{CB}(t;i) \;=\; \frac{\|C_O^i(t) - C_I^i(t)\|}{\sqrt{3 \times 255^2}} \qquad (2)$$

where, K_t is the total number of normal lines drawn to the boundary of the object at equal intervals in frame t, and $C_O^i(t)$ is the average color calculated in the $M \times M$ neighborhood of the pixel $p_O^i(x, y; t)$ using YCbCr color space. The average inside color $C_I^i(t)$ is defined similarly. Instead of the averaging operation, the α-trimmed mean can also be used in Eqn. 2, which will be described shortly.

We define the color metric for the whole sequence as:

$$0 \leq D_{CB} = f(d_{CB}(t), t = 1, \ldots, T) \leq 1, \qquad (3)$$

where T is the number of frames in the sequence. The function $f(.)$ can be defined in different ways such as the mean function, or the α-trimmed mean function [3]. The α-trimmed mean is defined as:

$$D_{CB,\alpha} = \frac{1}{T - 2[\alpha T]} \sum_{i=[\alpha T]+1}^{T-[\alpha T]} d_{CB}(i), \qquad (4)$$

where [.] denotes the integer ceiling function and $d_{CB}(i)$ denotes the i^{th} element of the sorted array $d_{CB}(.)$. When α is zero, the above expression is identical to the sample mean. When T is odd and α is 0.5, the α-trimmed mean is the same as the median of the sample set.

2.2. Inter-frame Color Histogram Differencing

In order to test whether the object is segmented/tracked correctly in each frame, we make use of the second assumption stated above, i.e. the color histogram of the object is assumed to be stationary from frame to frame. If a part of the background is included into the segmentation map by mistake, the color histogram of the segmented object is expected to follow the background histogram changes.

We can evaluate the stationarity of the color histogram of the segmented object by calculating the pairwise color histogram difference of the video object planes at time t and $t - 1$. Another approach to make the histogram differencing more robust to self-occlusions and mild intensity variations is to look at the difference between the color histogram of the video object plane at frame t and the smoothed color histogram of the video object planes for frames $\{t - i, \ldots, t + i\}$. This smoothing can be achieved by simple averaging or median filtering of the corresponding bins in the histograms of object planes in frames $\{t - i, \ldots, t + i\}$.

Let us denote the color histogram of the video object calculated using the YCbCr color space at time t as H_t. The locally smoothed color histogram is calculated using the formula:

$$H_{t,av}(j) = Med\{H_{t-i}(j), \ldots, H_{t+i}(j)\}, \quad j = 1, \ldots, B, \qquad (5)$$

where B denotes the total number of bins in the color histogram. The color histogram is represented as a 1-D vector obtained by concatenating the histograms for Y, Cb and Cr components.

The discrepancy between the color histograms H_t and $H_{t,av}$ is estimated using four different metrics as described below [4, 5], namely the L_1, L_2, χ^2 and histogram intersection metrics. In the following formulae, the scaling parameters R_1 and R_2 are used to normalize the data when the total number of elements in the two histograms are different:

$$R_1 = \sqrt{\frac{N_{H_t}}{N_{H_{t,av}}}}, \quad R_2 = \frac{1}{R_1},$$

$$N_{H_t} = \sum_{j=1}^{B} H_t(j), \quad N_{H_{t,av}} = \sum_{j=1}^{B} H_{t,av}(j),$$

$$NS_{H_t} = \sum_{j=1}^{B} H_t^2(j), \quad NS_{H_{t,av}} = \sum_{j=1}^{B} H_{t,av}^2(j).$$

1. **The L_1 Metric:** The L_1 distance between the two histograms is calculated and normalized to the range [0, 1] as follows,

$$0 \leq d_{L_1}(H_t, H_{t,av}) = \frac{\sum_{j=1}^{B} |R_1 H_t(j) - R_2 H_{t,av}(j)|}{2\sqrt{N_{H_t} N_{H_{t,av}}}} \leq 1. \qquad (6)$$

2. **The L_2 Metric:** The L_2 distance between the two histograms is calculated and normalized to the range [0, 1] as follows,

$$0 \leq d_{L_2}(H_t, H_{t,av}) = \sqrt{\frac{\sum_{j=1}^{B} [R_1 H_t(j) - R_2 H_{t,av}(j)]^2}{NS_{H_t} + NS_{H_{t,av}}}} \leq 1. \qquad (7)$$

3. **The χ^2 Metric:** is used to compare two binned data sets, and to determine if they are drawn from the same distribution function [5]. It is defined and normalized to the range [0, 1] as follows:

$$0 \leq \chi^2(H_t, H_{t,av}) = \frac{\sum_{j=1}^{B} \frac{[R_1 H_t(j) - R_2 H_{t,av}(j)]^2}{H_t(j) + H_{t,av}(j)}}{N_{H_t} + N_{H_{t,av}}} \leq 1. \qquad (8)$$

4. **Histogram Intersection Metric:** To quantify the difference of the two histograms using the histogram intersection method, we define the histogram intersection metric as:

$$0 \leq d_{HI}(H_t, H_{t,av}) = 1 - HI(H_t, H_{t,av}) \leq 1, \qquad (9)$$

where, $HI(H_t, H_{t,av})$ determines the number of pixels that share the same color in the two histograms [6]:

$$0 \leq HI(H_t, H_{t,av}) = \frac{\sum_{j=1}^{B} \min[H_t(j), H_{t,av}(j)]}{\min(N_{H_t}, N_{H_{t-1}})} \leq 1. \qquad (10)$$

Note that when $N_{H_t} = N_{H_{t-1}}$, i.e. the number of pixels histograms are equal, histogram intersection metric is equivalent to the L_1 metric [6]. The experimental results for local color histogram differences are given in Section 5.2 for the above metrics. The sensitivity of these metrics to the deviations in the location of the segmentation map are also analyzed by randomly shifting the correct segmentation maps.

Let $d_{CH}(t)$ denote the histogram difference metric calculated using one of the four metrics defined above. We define the histogram difference metric for the whole sequence as:

$$0 \leq D_{CH} = f(d_{CH}(t), t = 1, \ldots, T) \leq 1, \qquad (11)$$

where the function $f(.)$ can be chosen as discussed in the previous section.

3. MOTION METRIC

The assumptions that we make about the motion of the segmented object are as follows: 1)The motion vectors of the object that are 'just inside' of the object boundary and the background motion vectors that are 'just outside' of the object boundary are different. In other words, motion boundaries coincide with the object boundaries. 2) Background is either stationary or has global motion which shall be compensated for.

In order to quantify how well the estimated object boundaries coincide with actual motion boundaries, we use an approach similar to the one used for color. We draw small normal lines to the boundary at regular intervals as shown in Figure 1(b), and we look at the difference of the motion vectors around the points $p_O^i(x, y; t)$ and $p_I^i(x, y; t)$. The motion metric estimated following this approach for frame t can be expressed as follows:

$$
\begin{aligned}
0 \leq d_M(t) &= 1 - \frac{\sum_{i=1}^{K_t} d_M(t; i)}{\sum_{i=1}^{K_t} w_i} \leq 1, && (12) \\
d_M(t; i) &= d(v_O^i(t), v_I^i(t)) \cdot w_i && (13) \\
0 \leq w_i &= R(v_O^i(t)) \cdot R(v_I^i(t)) \leq 1, && (14)
\end{aligned}
$$

where $v_O^i(t)$ and $v_I^i(t)$ denote the average motion vectors calculated in a $M \times M$ square around the points $p_O^i(x, y; t)$ and $p_I^i(x, y; t)$, respectively, and $d(v_O^i(t), v_I^i(t))$ denotes the distance between the two average motion vectors which is calculated as:

$$
0 \leq d(v_O^i(t), v_I^i(t)) = \frac{\| v_O^i(t) - v_I^i(t) \|}{\| v_O^i(t) \| + \| v_I^i(t) \|} \leq 1. \qquad (15)
$$

In Eq. (14), $R(.)$ denotes the reliability of the motion vector $f^i(t)$ at point p^i [7]:

$$
R(v^i(t)) = e^{\left(- \frac{\| v^i(t) - b^i(t+1) \|^2}{2\sigma_m^2} \right)} \cdot e^{\left(- \frac{\| c(p^i; t) - c(p^i + v^i; t+1) \|^2}{2\sigma_c^2} \right)},
$$

where $b^i(t+1)$ denotes the backward motion vector at location $p^i + v^i$ in frame $t+1$; $c(p^i; t)$ denotes the color intensity and the parameters σ_m, σ_c are chosen freely.

We define the motion metric for the whole sequence as:

$$
0 \leq D_M = f(d_M(t), t = 1, \ldots, T) \leq 1. \qquad (16)
$$

4. PERFORMANCE EVALUATION

In this section, we derive a single numerical measure to evaluate the performance of object segmentation and tracking results, as well as spatial and temporal localization of incorrect boundary segments.

4.1. Combining Color and Motion Metrics

A single numerical measure can be obtained to evaluate the performance of spatio-temporal segmentation of a video object by combining the color and motion metrics defined above as follows:

$$
D = \mu D_{CB} + \beta D_{CH} + \gamma D_M, \qquad (17)
$$

where the parameters μ, β, and γ can be adjusted according to the characteristics of the video sequence and the relative importance and accuracy of color and motion features. Note that if the summation $\mu + \beta + \gamma$ is restricted to be one, the the metric D takes values between $[0, 1]$. If this measure is above a certain threshold, it is possible to localize incorrect boundary segments in time and space as described next.

4.2. Temporal Localization

The temporal localization can be achieved by checking the color and motion components of the measure

$$
d(t) = \mu_t d_{CB}(t) + \beta_t d_{CH}(t) + \gamma_t d_M(t), \qquad (18)
$$

at each frame against a threshold.

4.3. Spatial Localization

In frames t for which $d(t)$ is above the threshold, we can identify the segments of the boundary that have been tracked incorrectly using the color and motion scores that are obtained from 'inside' and 'outside' point pairs. Using Eqns. (2) and (13), if

$$
d_{CB}(t; i) + d_M(t; i) < \gamma, \qquad (19)
$$

where γ is a threshold value, we mark that segment between points $i - 1$ and $i + 1$ of the estimated object boundary as incorrect.

5. EXPERIMENTAL RESULTS

In order to test the effectiveness of the above proposed metrics, we quantitatively evaluate the ground-truth segmentation maps of the Hall monitor sequence which are obtained by hand for frames 32-230. A sample video object plane is shown in Figure 1(a).

5.1. Experiments with intra-frame color differences along the boundary

The color differences along the boundary of the Hall monitor sequence are calculated in the YCbCr color space using the Euclidean distance.

In the first column of Table 1, the color metric $D_{CB,\alpha}$ is given which is calculated using the expression given in Eqn.3 for frames 32-230, with $\alpha = 0.1$. The second column shows the variance of the values $d_{CB}(t)$ for different values of L. In order to observe the sensitivity of the metric $d_{CB}(t)$ to shifts in the correct segmentation masks, we randomly shifted the ground-truth segmentation masks with ±10 pixels, in an attempt to simulate incorrect segmentation. The percentage increase in $D_{CB,\alpha}$, and the variance of $d_{CB}(t)$ are given in Table 1, which shows that the maximum increase occurs when L = 2 and L = 3, respectively. Although the variance of $d_{CB}(t)$ depend on the characteristics of the background and object color, we expect it to be small if the object is correctly segmented and the background surrounding the object is not cluttered.

5.2. Experiments with inter-frame color histogram differences

The results for the metric based on histogram differences are summarized in Table 2. We can observe that χ^2 is the most sensitive metric to the shifts in the segmentation map since the percentage increase in $D_{CH,\alpha}$, and the variance of $d_{CH}(t)$ are largest for the χ^2 metric when the segmentation map is shifted. This indicates that it is able to discriminate the imperfections in the segmentation/tracking results better then the other three metrics. The variance of $d_{CH}(t)$ is ideally expected to be low when the segmentation masks are correctly located since the color histogram of the object is not expected to change much between frames. In Figure 2, a plot of the χ^2 metric is given calculated using unshifted segmentation masks up to frame 100 and with masks shifted by ±10 pixels for frames 101-230. As seen in the figure, the histogram difference metric based on χ^2 distance calculation signals the incorrectness of the the segmentation mask successfully.

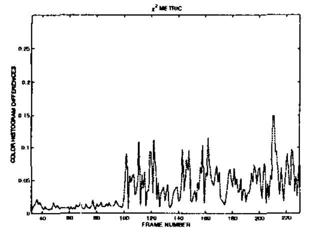

Fig. 2. The color histogram differences between H_t and $H_{t,av}$, calculated with χ^2 metric, using segmentation maps shifted by ±10 pixels, starting from frame 100.

5.3. Experiments with the motion differences

Forward and backward motion estimation between successive frames of the Hall monitor sequence is performed using a hierarchical version of the Lucas-Kanade motion estimation algorithm [8]. In the first two columns of Table 3, the values of $D_{M,\alpha}$ and the variance of $d_M(t)$ are given for different L values. The last two columns show the percentage increase in $D_{M,\alpha}$ and variance of $d_M(t)$ for the shifted segmentation maps. When the background is stationary or moving uniformly, the variance of $d_M(t)$ is expected to be small if the object is segmented correctly. When the segmentation map shifts from its correct location, the variance is expected to increase, which is the case for the Hall monitor sequence as shown in Table 3.

5.4. Localization of incorrect segmentation

In Figure 3, we show the video object plane for the 134^{th} frame of the Hall monitor sequence (downloaded from the web page of COST 211 group). As observed, the boundary of the object is located incorrectly except for a short segment around the shirt. The correctly located boundary segments are marked with solid lines and the incorrectly located segments are marked with dashed lines. The metric (19) is able to support the subjective observations quantitatively.

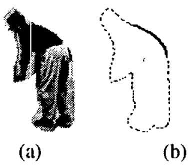

Fig. 3. (a) The video object plane for the 134^{th} frame of the Hall monitor sequence which is downloaded from the web page of COST 211 group. (b) Correctly segmented regions of the boundary are marked with solid lines and incorrectly segmented regions are marked with dashed lines.

	No Shift		±10 pixel shift	
L	$D_{CB,\alpha}$ $\times 10^{-1}$	var($d_{CB}(t)$) $\times 10^{-4}$	Perc. Incr. $D_{CB,\alpha}$	Perc. Inc. var($d_{CB}(t)$)
5	8.66	6.33	5.6	21.1
4	8.62	5.86	6.4	22.7
3	8.58	4.40	8.5	55.8
2	8.65	4.81	9	16.6

Table 1. The scores for color difference metric along the object boundary.

	No Shift		±10 pixel shift	
	$D_{CH,\alpha}$ $\times 10^{-2}$	var($d_{CH}(t)$) $\times 10^{-5}$	Perc. Incr. $D_{CH,\alpha}$	Perc. Inc. var($d_{CH}(t)$)
χ^2	0.54	0.649	238.5	2392.96
L_2	4.54	71.49	80.1	243.17
L_1	4.54	16.74	80.1	454.5
HI	3.73	19.65	90.5	376.64

Table 2. The scores for color histogram difference metric.

	No Shift		±10 pixel shift	
L	$D_{M,\alpha}$ $\times 10^{-2}$	var($d_M(t)$) $\times 10^{-4}$	Perc. Incr. $D_{M,\alpha}$	Perc. Inc. var($d_M(t)$)
5	4.3	7.75	28	59.8
7	6.35	11.24	24.4	55.29

Table 3. The scores for motion difference metric along the object boundary.

6. CONCLUSIONS

We presented three different performance evaluation metrics for quantitative evaluation of video object segmentation and tracking algorithms and we tested the sensitivity of the proposed metrics to shifts in the segmentation map. Using the proposed metrics, it is possible to locate the regions of the boundary where the segmentation is not correct. An object tracking scheme that optimizes its performance based on the proposed metrics has been developed and will be presented elsewhere.

7. REFERENCES

[1] Ç. E. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X European Signal Processing Conference*, September 2000, vol. 2, pp. 917–920.

[2] X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. X European Signal Processing Conference*, September 2000, vol. 4.

[3] J. Bednar and T. L. Watt, "Alpha-trimmed means and their relationship to median filters," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 32, no. 1, pp. 145–153, 1984.

[4] A. M. Ferman, A. M. Tekalp, and R. Mehrotra, "Robust histogram descriptors for video segment retrieval and identification," *Submitted to IEEE Trans. on Image Processing*, 2000.

[5] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flanney, *Numerical Recipes in C*, Cambridge Univeristy Press, 1992.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *Int. Journal of Computer Vision*, vol. 7, no. 11, pp. 11–32, 1991.

[7] Y. Fu, A. T. Erdem, and A. M. Tekalp,, "Tracking visible boundary of objects using occlusion adaptive motion snake," *IEEE Trans. Image Processing*, vol. 9, no. 12, pp. 2051–2060, December 2000.

[8] A. M. Tekalp, *Digital Video Processing*, Prentice-Hall, New Jersey, 1995.