Performance Measures for Video Object Segmentation and Tracking

Çiğdem Eroğlu Erdem, *Member, IEEE*, Bülent Sankur, and A. Murat Tekalp, *Fellow, IEEE*

***Abstract*—We propose measures to evaluate quantitatively the performance of video object segmentation and tracking methods without ground-truth (GT) segmentation maps. The proposed measures are based on spatial differences of color and motion along the boundary of the estimated video object plane and temporal differences between the color histogram of the current object plane and its predecessors. They can be used to localize (spatially and/or temporally) regions where segmentation results are good or bad; and/or they can be combined to yield a single numerical measure to indicate the goodness of the boundary segmentation and tracking results over a sequence. The validity of the proposed performance measures *without GT* have been demonstrated by canonical correlation analysis with another set of measures *with GT* on a set of sequences (where GT information is available). Experimental results are presented to evaluate the segmentation maps obtained from various sequences using different segmentation approaches.**

***Index Terms*—Canonical correlation analysis, object segmentation, object tracking, performance evaluation without ground truth (GT).**

I. Introduction

OBJECT-BASED video segmentation and object tracking are challenging and active research areas in digital video processing and computer vision. The task of segmenting/tracking a video object emerges in many applications such as object-based video coding (e.g., MPEG-4), content-based video indexing and retrieval (e.g., MPEG-7), video surveillance for security, video editing for post production, and animation for entertainment video.

Comparative assessment of segmentation algorithms is often based upon subjective judgment, which is qualitative and time consuming. Therefore, there is a need for an automatic, objective spatio-temporal methodology, not only for comparison of overall algorithmic performance, but also as a tool to monitor spatio-temporal consistency of the segmentation of individual objects.

Although several measures have been developed for still image segmentation [1]–[7], they are not directly applicable to video object segmentation and tracking. Recently, a number of video segmentation measures have been proposed based on ground-truth (GT) segmentation maps. For example, Senior *et al.* [8] employed trajectories of the centroid of tracked objects and their velocities. Erdem *et al.* [9] utilize a combination of misclassification penalty, shape penalty, and motion penalty to assess the video object segmentation results. Marichal and Villegas [10] also use three measures, which are the spatial accuracy, temporal local stability, and the temporal shift measures. In their work, the spatial accuracy measure favors the segmentation algorithms which overestimate the segmentation masks over the ones which underestimate it. Their temporal coherency measure utilizes the displacement of the gravity centers of the estimated and reference segmentation masks. Correia and Pereira [11] follow a four-step strategy for video segmentation quality evaluation. First, individual segmentation quality for each object is measured. Next, the relevance of each object is computed based on how much visual attention it captures. Then, similarity of reference and estimated segmentations is computed. Finally, an overall segmentation quality evaluation is obtained by combining these three measures. The usefulness of these measures is limited in that they require the presence of GT information, which is not easily available.

Recently, a set of standalone (without GT) performance metrics have been proposed by Correia and Pereira [12], [13]. These metrics are grouped into two classes as *intra-object homogeneity* measures and *inter-object disparity* measures. The *intra-object homogeneity* metrics are composed of the shape regularity (based on compactness, circularity, and elongation properties), spatial uniformity (based on spatial perceptual information and texture variance), temporal stability (based on size, elongation, and texture differences), and motion uniformity (based on variance of motion vectors and criticality). The *inter-object disparity* metrics are composed of local color and motion contrast with neighbors. The authors demonstrate through experiments that the proposed measures are able to estimate segmentation quality in a correlated way with the judgements of a human observer for different types of content.

The main contribution of our work is to develop quantitative performance measures for video object tracking and segmentation, which do not require GT segmentation maps, and then to show statistically that they are indeed correlated with another set of measures based on GT segmentation maps (under certain assumptions). The proposed nonground-truth (NGT)

Manuscript received January 2, 2003; revised November 12, 2003. This work was supported in part by the Scientific and Technical Research Council of Turkey (TÜBİTAK-BAYG) and in part by the Boğaziçi University Research Fund under Project 99A203. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philippe Salembier.

Ç. E. Erdem is with Momentum Digital Media Technology, Inc., İstanbul 80815, Turkey, and also with the Department of Electrical and Electronics Engineering, Boğaziçi University, İstanbul 80815, Turkey (e-mail: cigdem@ieee.org).

B. Sankur is with the Department of Electrical and Electronics Engineering, Boğaziçi University, İstanbul 80815, Turkey (e-mail: sankur@boun.edu.tr).

A. M. Tekalp is with Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA and also with the College of Engineering, Koç University, Sarıyer, İstanbul 80815, Turkey (e-mail: tekalp@ece.rochester.edu; mtekalp@ku.edu.tr).

Digital Object Identifier 10.1109/TIP.2004.828427

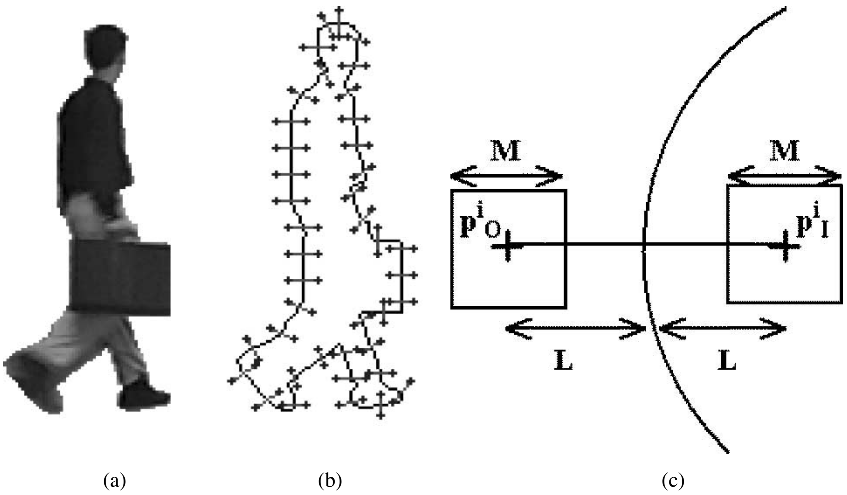


Fig. 1. (a) Video object plane for the 32nd frame of the "Hall monitor" sequence. (b) The boundary of the video object plane with the normal lines. (c) A closeup of a normal line drawn on the boundary. The two points "just inside" and "just outside" of the boundary are shown with symbols p^i_I and p^i_O, respectively.

measures exploit color and motion features in the vicinity of the segmented video object. One of the features is the spatial color contrast along the boundary of each object plane. The second one consists of color histogram differences across video object planes, which aims to evaluate the goodness of segmentation along a spatio-temporal trajectory. The third feature is based on motion vector differences along the video object plane boundary. Often, a single numerical figure does not suffice to evaluate the goodness of a segmentation/tracking for a whole video sequence. Since the spatial segmentation quality can change from frame to frame and/or the temporal segmentation stability may deteriorate over subsequences, we propose additional measures to localize in time or in space the unsuccessful segmentation outcomes.

In Section II, we present color and motion features used in the performance evaluation measures. An overall performance measure for the whole sequence, as well as measures to localize performance spatially and temporally, are developed in Section III. In Section IV, canonical correlation analysis is used to validate NGT measures against GT measures. In Section V, experimental results are provided. Finally, in Section VI, conclusive remarks are given.

II. Features for Video Segmentation Evaluation

The proposed video segmentation performance measures are based on color and motion features. Color features are based on the following assumptions: 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. These assumptions hold true for most video sequences and are also assumed by many segmentation algorithms. Note that the background and its color histogram are not required to be stationary from frame to frame, and there are also no restrictions on the shape and rigidity of the segmented/tracked object.

In addition, we make the following assumptions about the motion of video objects. 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, object boundaries coincide with motion boundaries. 2) The background is either stationary or has global motion, which can be compensated for.

In the following, we present two color features and one motion feature, which will be used under the above assumptions in order to compute the goodness of a segmented video object plane.

A. Spatial Color Contrast Along Object Boundary

Since object boundaries are assumed to coincide with color boundaries, there should be an observable difference between the color of pixels across the estimated object boundary. In order to measure the color difference, we establish a set of probe pixels "just inside" and "just outside" by drawing normal lines of length L astride the estimated object boundary at equal intervals as illustrated in Fig. 1(b). As depicted in the closeup in Fig. 1(c), the color probes are $M \times M$ regions centered at the two ends of the ith normal line, which are marked with plus signs and denoted by p^i_I (inside pixel) and p^i_O (outside pixel).

We define the color difference measure calculated along the boundary of the object in frame t as

$$0 \leq d_{\text{color}}(t) = 1 - \frac{1}{K_t} \sum_{i=1}^{K_t} \delta_{\text{color}}(t;i) \leq 1 \tag{1}$$

$$\delta_{\text{color}}(t;i) = \frac{\|C^i_O(t) - C^i_I(t)\|}{\sqrt{3 \times 255^2}} \tag{2}$$

where K_t is the total number of normal lines drawn on the object boundary in frame t and $C^i_O(t)$ is the average color calculated in the $M \times M$ neighborhood of the pixel $p^i_O(x,y;t)$ using the Y-Cb-Cr color space. The average inside color $C^i_I(t)$ is defined similarly. The worst score is 1 and $d_{\text{color}}(t)$ decreases toward zero as the color contrast along object boundary increase, possibly indicating a good segmentation.

We define the color measure for the whole sequence as

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

where T is the number of frames in the sequence. The function $f(.)$ can be defined in different ways such as the mean function, α-trimmed mean function [14], median or the maximum function.

When the location of the object boundary is estimated correctly in frame t, we expect the color measure $d_{\text{color}}(t)$ to take a small value. However, the converse of this statement is not necessarily true. That is, if the color measure $d_{\text{color}}(t)$ has a small value in frame t, this does not necessarily imply that the object boundary is located correctly. Therefore, this measure should be used carefully depending on the characteristics of the background surrounding the object. This color measure is expected to be reliable when the object and background textures are not cluttered and when the color contrast between the object and the background is high. Note that the color-based measure is applicable to both video and still images.

B. Temporal Color Histogram Difference

The color histogram of the video object planes vary from frame to frame noticeably if the background is erroneously included into the segmentation map or when a portion of the object is excluded from the segmentation map. A straightforward way to assess the changes in the color histogram of the segmented object is to calculate the pairwise color histogram differences of the video object planes (VOP) at time t and $t-1$. In order to allow small variations due to self occlusions and mild intensity variations within the object, a robust scheme should consider the difference between the color histogram in the present frame (t) and the smoothed color histogram of the video object planes over frames $\{t-i,\ldots,t-1\}$. This frame-by-frame histogram smoothing can be achieved by simple averaging or median filtering of the corresponding histogram bins of VOPs in frames $\{t-i,\ldots,t-1\}$. However, a drawback of this approach is that it may not catch a gradual tracking performance deterioration. Therefore, we can alternatively check the histogram differences between the reference (e.g., first) VOP and current estimated VOP. This method penalizes the cumulative difference effect of the previous approach and is more sensitive.

Let us denote the color histogram of the video object calculated using the Y-Cb-Cr color space at time t as H_t. The reference color histogram with which H_t is going to be compared and calculated using one of the methods discussed in the previous paragraph is denoted by H_{ref}.

In order to estimate the discrepancy between the color histograms H_t and H_{ref}, each with B bins, we studied four different measures as described below [15], [16], namely the L_1, L_2, χ^2 and histogram intersection measures.

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

$$0 \le d_{L_1}(H_t, H_{\text{ref}}) = \frac{\sum_{j=1}^{B} |r_1 H_t(j) - r_2 H_{\text{ref}}(j)|}{2\sqrt{N_{H_t} N_{H_{\text{ref}}}}} \le 1 \quad (4)$$

where the following definitions are used:

$$r_1 = \sqrt{\frac{N_{H_{\text{ref}}}}{N_{H_t}}}, \quad r_2 = \frac{1}{r_1}$$

$$N_{H_t} = \sum_{j=1}^{B} H_t(j), \quad N_{H_{\text{ref}}} = \sum_{j=1}^{B} H_{\text{ref}}(j).$$

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. For gray-scale images, N_{H_t} is equal to the object size, and for color images, it is three times the size of the object.

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

$$0 \le d_{L_2}(H_t, H_{\text{ref}}) = \sqrt{\frac{\sum_{j=1}^{B} [r_1 H_t(j) - r_2 H_{\text{ref}}(j)]^2}{NS_{H_t} + NS_{H_{\text{ref}}}}} \le 1 \quad (5)$$

with the definitions

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

- **The χ^2 Metric** is used to compare two binned data sets, and to determine if they are drawn from the same distribution function [16]. It is defined and normalized to the range [0, 1] as follows:

$$0 \le \chi^2(H_t, H_{\text{ref}}) = \frac{\sum_{j=1}^{B} \frac{[r_1 H_t(j) - r_2 H_{\text{ref}}(j)]^2}{H_t(j) + H_{\text{ref}}(j)}}{N_{H_t} + N_{H_{\text{ref}}}} \le 1. \quad (6)$$

- **Histogram Intersection Measure:** To quantify the difference of the two histograms using the histogram intersection method, we define the histogram intersection measure as

$$0 \le d_{HI}(H_t, H_{\text{ref}}) = 1 - HI(H_t, H_{\text{ref}}) \le 1 \quad (7)$$

where, $HI(H_t, H_{\text{ref}})$ determines the number of pixels that share the same color in the two histograms [17]

$$0 \le HI(H_t, H_{\text{ref}}) = \frac{\sum_{j=1}^{B} \min[H_t(j), H_{\text{ref}}(j)]}{\min(N_{H_t}, N_{H_{\text{ref}}})} \le 1. \quad (8)$$

In order to choose the most sensitive histogram differencing measure, we conducted an experiment in Section V-A, where a number of GT objects were randomly perturbed and the mean and variance of the measures were computed. It was observed that the χ^2 distance was the most sensitive measure. Therefore, we use the χ^2 measure, $d_{\text{hist}}(t) = \chi^2(H_t, H_{\text{ref}})$, in all other experiments. Note that if the two histograms being compared are identical, $d_{\text{hist}}(t) = 0$, and $d_{\text{hist}}(t)$ increases toward 1, as the histograms differ more. We define the histogram difference measure for the whole sequence as

$$0 \le D_{\text{hist}} = f(d_{\text{hist}}(t)) \le 1, \qquad t = 1, \ldots, T \quad (9)$$

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

C. Motion Difference Along Object Boundary

In order to quantify how well the estimated object boundaries coincide with actual motion boundaries, we adopt the geometry of the probes used for color features as in Fig. 1(b) and (c) and consider the difference of the average motion vectors in the neighborhood of the points p_O^i and p_I^i. The motion measure for frame t is estimated as follows:

$$0 \leq d_{\text{motion}}(t) = 1 - \frac{\sum_{i=1}^{K_t} \delta_{\text{motion}}(t;i)}{\sum_{i=1}^{K_t} w_i} \leq 1 \tag{10}$$

$$\delta_{\text{motion}}(t;i) = d(\mathbf{v}_O^i(t), \mathbf{v}_I^i(t)) \cdot w_i \tag{11}$$

$$0 \leq w_i = R(\mathbf{v}_O^i(t)) \cdot R(\mathbf{v}_I^i(t)) \leq 1 \tag{12}$$

where $\mathbf{v}_O^i(t)$ and $\mathbf{v}_I^i(t)$ denote the average motion vectors calculated in the $M \times M$ square around the points $p_O^i(x,y;t)$ and $p_I^i(x,y;t)$, respectively, and $d(\mathbf{v}_O^i(t), \mathbf{v}_I^i(t))$ denotes the distance between the two average motion vectors, which is calculated as

$$0 \leq d(\mathbf{v}_O^i(t), \mathbf{v}_I^i(t)) = 1 - \exp\left(-\frac{\|\mathbf{v}_O^i(t) - \mathbf{v}_I^i(t)\|}{\sigma^2}\right) \leq 1. \tag{13}$$

We observed during the experiments that selecting the parameter $\sigma = 1$ is reasonable and causes the distance of the motion vectors to be approximately 0.63, if the magnitude of their difference is 1. If the sigma is chosen to be larger, then only motion errors of several pixels would be punished. On the other hand, choosing a small sigma signifies that fractional motion errors would be amplified. Since we want to be sensitive to motion dicrepancies of the order of one pixel or more, the choice of $\sigma = 1$ is adequate. In (12), $R(.)$ denotes the reliability of the motion vector $\mathbf{v}^i(t)$ at point p^i [18]

$$R(\mathbf{v}^i(t)) = \exp\left(-\frac{\|\mathbf{v}^i(t) - \mathbf{b}^i(t+1)\|^2}{2\sigma_m^2}\right) \cdot \exp\left(-\frac{\|c(p^i;t) - c(p^i + \mathbf{v}^i(t); t+1)\|^2}{2\sigma_c^2}\right) \tag{14}$$

where $\mathbf{b}^i(t+1)$ denotes the backward motion vector at location $p^i + \mathbf{v}^i(t)$ in frame $t+1$; $c(p^i;t)$ denotes the color intensity and the parameters σ_m, σ_c are chosen similarly as in [18]. According to the above $R(.)$ function, a motion vector at a pixel position is reliable provided that the backward and forward motion predictions agree with each other both in magnitude and in the color of their pixels.

We define the motion measure for the whole sequence as (again choosing a convenient averaging function)

$$0 \leq D_{\text{motion}} = f(d_{\text{motion}}(t)) \leq 1, \qquad t = 1, \ldots, T. \tag{15}$$

There is, however, a caveat for the motion measure. This score can sometimes be large, not because of any wrong segmentation, but as a consequence of the fact that the object is not moving significantly during a subsequence. Hence, there may not exist a clearly definable motion boundary. In such a sequel of frames, we should then rely on the persistence of color boundaries, and the coefficient of the the motion score should be decreased. For example, one can consider a weighting on $d_{\text{motion}}(t)$ as $S(V_{\text{med}}(t))$, where $V_{\text{med}}(t)$ is the median of motion vector magnitudes along the boundary and $S(.)$ is the fuzzy weighting function introduced in (18) and illustrated in Fig. 2. The midpoint of the ramp can be set at what we define as the "small motion threshold," for example, $c_2 = 1$ pixels/frame.

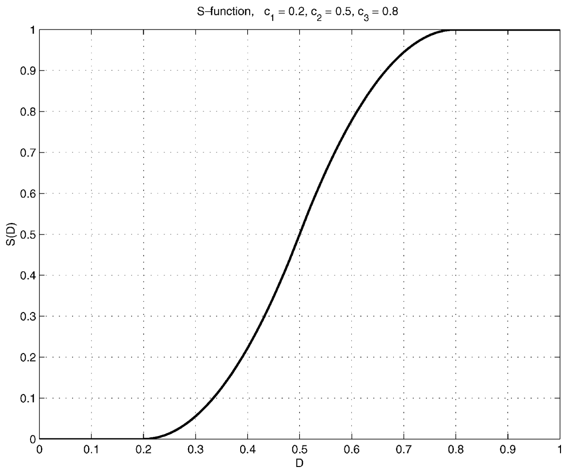


Fig. 2. S function.

III. Performance Measures Without GT

In this section, we combine the color and motion measures to obtain scores that reflect the success of segmentation and tracking of objects for the whole sequence (Section III-A), as well as temporal (Section III-B) and spatial localization (Section III-C) of incorrect boundary segments.

A. Combined Performance Measure for Sequence

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

$$D = \mu D_{\text{color}} + \beta D_{\text{hist}} + \gamma D_{\text{motion}} \tag{16}$$

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 since the sum $\mu + \beta + \gamma$ is restricted to be one, the measure D takes values between [0, 1]. In the absence of any preference indication for color and motion, one can consider the straight arithmetic averaging of the three measures, by simply choosing $\mu = 1/3$, $\beta = 1/3$, $\gamma = 1/3$. Note that, although all the measures are between [0, 1], their numerical scales may be different, which may need prior normalization.

Alternative combinations of the three measures may be desirable. For example, the given sequence can be judged by $D = \max\{D_{\text{color}}, D_{\text{hist}}, D_{\text{motion}}\}$. A more lenient penalty function would be

$$D = \frac{\mu(D_{\text{color}})D_{\text{color}} + \mu(D_{\text{hist}})D_{\text{hist}} + \mu(D_{\text{motion}})D_{\text{motion}}}{\mu(D_{\text{color}}) + \mu(D_{\text{hist}}) + \mu(D_{\text{motion}})} \tag{17}$$

where $\mu(.)$ is a fuzzy weighting function given by the $S(D)$ curve defined as

$$S(D) = \begin{cases} 0, & D \leq c_1 \\ \frac{(D-c_1)^2}{(c_2-c_1)(c_3-c_2)}, & c_1 \leq D < c_2 \\ 1 - \frac{(D-c_3)^2}{(c_3-c_2)(c_3-c_1)}, & c_2 \leq D < c_3 \\ 1, & D \geq c_3. \end{cases} \tag{18}$$

A sample $S(.)$ function with parameters $c_1 = 0.2$, $c_2 = 0.5$ and $c_3 = 0.8$ is shown in Fig. 2. The combination strategy in (17), gives more weight to large errors. As one of the three penalties gets larger, its weight also becomes saturated and vice versa.

For multiple object segmentation, we propose to consider the overall measure as

$$D_{\text{overall}} = \max\{D_{O_i}\} \tag{19}$$

where $D_{O_i}, i = 1, \ldots, N_O$ is defined in (16) for the object O_i, and N_O is the number of objects in the video sequence. The maximum operation has been used because a badly segmented object may attract enough negative attention causing an unsatisfactory subjective judgment, no matter how well the other objects in a scene are segmented. This approach is based on the assumption that all objects in the scene are equally important. If this is not the case, the performance scores of each can be weighted by an "importance coefficient" and an average of the object scores can be calculated. A way of estimating such an "importance coefficient" has been proposed by [13], which discusses a "contextual relevance metric" of each object based on its motion, texture, and shape properties. This contextual relevance metric can be adopted as an importance coefficient in our framework.

Using similar combinations of measures, it is possible to trace the performance of segmentation over time or in space and, thus, localize, for example, incorrect boundary segments.

B. Temporal Localization

The temporal performance localization can be achieved by checking, per frame, color, and motion measures, as a function of time, that is

$$d(t) = \mu_1 d_{\text{color}}(t) + \mu_2 d_{\text{hist}}(t) + \mu_3 d_{\text{motion}}(t) \tag{20}$$

where μ_1, μ_2, μ_3 could be determined with a method as in (16) or (17). In a sequence, any set of frames for which the $d(t)$ score exceeds a threshold T_d is judged to be "poorly segmented."

C. Spatial Localization

We can further identify incorrectly tracked boundary portions within any frame whose $d(t)$ score is above the threshold, using only the color and motion measures. Thus, rather than summing the measures along the object boundary, we consider pixel-individual discrepancies using (2) and (11)

$$\mu_4 \delta_{\text{color}}(t; i) + \mu_5 \delta_{\text{motion}}(t; i) > T_\delta \tag{21}$$

where T_δ is a threshold value. If the threshold T_δ is exceeded, we then mark that segment between points $i-1$ and $i+1$ of the estimated object boundary as incorrect. This threshold may be set at k sigma point, that is a boundary pixel is considered as badly segmented if its score in (21) is k standard deviations above the mean of the measure over the object.

IV. Statistical Validation of Proposed Measures

In order to check the validity of the proposed NGT performance measures, we introduce a canonical correlation analysis of the proposed measures against measures using GT maps. To this effect, we first review performance measures using GT information. The canonical correlation analysis framework will be discussed next. Experimental results of this correlation analysis on different sequences with different object segmentation/tracking approaches will be provided in Section V.

A. Measures With GT

The measures using GT segmentation maps [9] are based on the pixel misclassification penalty, shape difference penalty, and the motion penalty. These GT measures are also all normalized to the range [0, 1] and they are marked with a superscript "g" to distinguish them from the NGT measures. In order to calculate the **misclassification penalty** (d^g_{pixel}), the misclassified pixels in the estimated segmentation map that are farther from the actual object boundary are penalized more than the misclassified pixels that are closer to the actual object boundary

$$d^g_{\text{pixel}}(t) = \frac{\sum\limits_{(x,y)} I(x,y;t)\text{Cham}_g(x,y;t)}{\sum\limits_{(x,y)} \text{Cham}_g(x,y;t)} \tag{22}$$

where $I(x,y;t)$ denotes an indicator function which takes the value 1 if reference and estimated segmentation masks of the object differ, $\text{Cham}_g()$ denotes the Chamfer distance transform of the boundary of GT the object. Chamfer distance is a technique to obtain the distance transform of a binary image which approximates the Euclidean distance [19].

The **shape penalty** (d^g_{shape}) between the GT and the estimated segmentation maps are calculated by looking at the difference between the turning angle functions (TAF) [9] of the segment boundaries

$$d^g_{\text{shape}}(t) = \frac{\sum\limits_{j=1}^{K} |\Theta^t_g(j) - \Theta^t_s(j)|}{2\pi K} \tag{23}$$

where $\Theta^t_g(j)$ and $\Theta^t_g(j)$ denote the turning angle function of the GT and estimated object boundary and K is the total number of points in the TAF. Starting from a point on the boundary, the turning angle function [20] increases by the amount of the rotation angle if we turn left and decreases if we turn right. The total amount of turning angle for any closed shape is 360°.

Finally, the **motion penalty** (d^g_{motion}) is calculated by computing the motion vectors on the GT and the estimated segmentation maps

$$d^g_{\text{motion}}(t) = \frac{\|\mathbf{v}_g(t) - \mathbf{v}_s(t)\|}{\|\mathbf{v}_g(t)\| + \|\mathbf{v}_s(t)\|} \tag{24}$$

where $\mathbf{v}_g(t)$ is any parametric motion representation for the GT object.

The measures for the whole sequence or video shot can be found by averaging or taking the maximum of the values for each frame. The measures for the whole sequence will be denoted by D^g_{pixel}, D^g_{shape}, and D^g_{motion}. The extension to multiple objects can also be carried out as discussed in Section III-A.

B. Representation of Data

Let the **GT** performance scores obtained for the tth frame ($t = 1, \ldots, n$) video shot or a video sequence denoted by

$$\mathbf{x}_t = [x_{t1} \ \ldots \ x_{tp}]^T = [d^g_{\text{pixel}}(t) \ \ d^g_{\text{shape}}(t) \ \ d^g_{\text{motion}}(t)]^T \tag{25}$$

consisting of the pixel misclassification penalty, shape penalty and the motion penalty and hence $p = 3$. The superscript g denotes that these measures use GT segmentation maps. Similarly, let the **NGT** performance scores obtained for the tth frame be denoted by

$$\mathbf{y}_t = [y_{t1} \ \ldots \ y_{tq}]^T = [d_{\text{hist}}(t) \ \ d_{\text{color}}(t) \ \ d_{\text{motion}}(t)]^T \tag{26}$$

where $q = 3$ and the first variable is the inter-frame color histogram difference measure calculated using the χ^2 measure, the second parameter ($y_{t2} = d_{\text{color}}(t)$) is the measure calculated from color differences along the estimated object boundary, and the third parameter ($y_{t3} = d_{\text{motion}}(t)$) is the measure of motion differences along the estimated object boundary.

Using the above vectors, the following data matrix can be constructed:

$$\mathbf{Z} = [\,\mathbf{z}_1 \quad \ldots \quad \mathbf{z}_n\,] = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ \mathbf{y}_1 & \ldots & \mathbf{y}_n \end{bmatrix} \tag{27}$$

where the performance measure vector in a column for $t = 1, \ldots, n$ reads as (28), shown at the bottom of the page, and n is the total number of observations. For example, n can be the number of frames of a sequence for which the performance measures are computed. If separate tracking results are collected for the same frame with different algorithms, the number of observations increases accordingly. For scale independence, the data matrix has been standardized by subtracting from each row its mean and by dividing by its standard deviation to yield $\bar{z}_t$.

Using the normalized variates, we calculate the matrix of sample covariances where the GT and NGT partitions are indicated as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{XX}}(p \times p) & \mathbf{\Sigma}_{\mathbf{XY}}(p \times q) \\ \mathbf{\Sigma}_{\mathbf{YX}}(q \times p) & \mathbf{\Sigma}_{\mathbf{YY}}(q \times q) \end{bmatrix} \tag{29}$$

where $\text{Cov}(\mathbf{X}) = \mathbf{\Sigma}_{\mathbf{XX}}$, $\text{Cov}(\mathbf{Y}) = \mathbf{\Sigma}_{\mathbf{YY}}$, and $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{\Sigma}_{\mathbf{XY}}$.

We will assume that $\text{Cov}(\mathbf{X})$ has full rank and we will take $p \leq q$, without loss of generality.

C. Canonical Correlation Analysis

The association between the two data sets, which, in our case, consist of GT and NGT measures, can be quantified by using canonical correlation analysis. The space defined by the measurement vectors is transformed in such a way that linear combination of one set is maximally correlated with the linear combination of the other set, while being mutually uncorrelated with the remaining $p-1$ eigensolutions [21]. Let us define these linear combinations as $(\mathbf{a}_i^T\mathbf{X}, \mathbf{b}_i^T\mathbf{Y}), i = 1, \ldots, p$, where $\mathbf{a}_1$ is (p× 1) and $\mathbf{b}_1$ is (q× 1). These pairs of linear combinations are referred to as *canonical variables* and their correlations are referred to as *canonical correlations*. Furthermore, the combinations in a set must be orthogonal to each other.

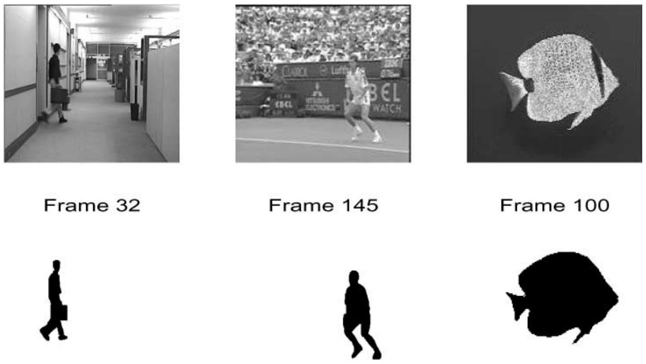


Fig. 3. (Top row) Sample frames and (bottom row) corresponding GT segmentation maps of the (left) Hall monitor, (middle) Stefan, and (right) Bream sequences.

The first set of canonical variates $\mathbf{a}_1^T\mathbf{X}$ and $\mathbf{b}_1^T\mathbf{Y}$ can be generated by

$$\max_{\mathbf{a}_1, \mathbf{b}_1} \ \text{Corr}(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y}) = \rho_1. \tag{30}$$

It can be shown that the parameters $\rho_1^2 \geq \rho_2^2 \geq \cdots \rho_p^2$ are the joint eigenvalues of the matrices $\mathbf{\Sigma}_{\mathbf{XX}}^{-1/2}\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1/2}$ or of $\mathbf{\Sigma}_{\mathbf{YY}}^{-1/2}\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_{\mathbf{YY}}^{-1/2}$. These matrices have, respectively, the eigenvector set $\mathbf{e}_1, \ldots, \mathbf{e}_p$ and $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_q$. Finally, the linear combiner weights $\mathbf{a}_i$ and $\mathbf{b}_i$ in (30) result from $\mathbf{e}_i^T\mathbf{\Sigma}_{\mathbf{XX}}^{-1/2}$ and $\mathbf{f}_i^T\mathbf{\Sigma}_{\mathbf{YY}}^{-1/2}$. The readers are referred to [21], [22] for a complete discussion of canonical correlation analysis.

The square of the canonical correlation ρ_1^2 gives us the proportion of variance in each canonical variate ($\mathbf{a}_1^T\mathbf{X}$ or $\mathbf{b}_1^T\mathbf{Y}$) that is related to the other canonical variate of the pair. Some researchers argue that [22] the degree of association (shared variance) between the two sets of variables ($\mathbf{X}$ and $\mathbf{Y}$) cannot be represented by ρ_i^2 and prefer the "canonical loadings" approach as explained in the following section.

D. Canonical Loadings and Redundancy

We would like to show that the GT performance measures are mostly redundant given the information about the NGT measures. Redundancy in this context corresponds to the amount of the variance of GT measures accounted for by the NGT measures. This redundancy can be computed via the *canonical loadings* [22] which will be described below.

With this goal in mind, we look at the relations of the performance measures in one set (say, GT) with the canonical variates of its own GT set and of the other set (NGT), called *intraset loadings* and *interset loadings*, respectively. In other words, we can use the correlations of the NGT performance measures in set $\mathbf{X}$ with the canonical variates of set $\mathbf{X}$ (intraset loadings),

$$\mathbf{z}_t = \left[d^g_{\text{pixel}}(t) \quad d^g_{\text{shape}}(t) \quad d^g_{\text{motion}}(t) \quad \middle| \quad d_{\text{hist}}(t) \quad d_{\text{color}}(t) \quad d_{\text{motion}}(t) \right]^T \tag{28}$$

TABLE I
SCORES FOR COLOR HISTOGRAM DIFFERENCE MEASURE

	Correct Segmentation		Incorrect Segmentation		Ratio (Incorrect / Correct)		
	Hall Monitor		Hall Monitor		Hall	Bream	Stefan
	D_{hist}	$\sigma_{d_{hist}}$	D_{hist}	$\sigma_{d_{hist}}$	D_{hist}	D_{hist}	D_{hist}
χ^2	0.013	0.005	0.041	0.026	3.13	9.30	2.16
L_2	0.117	0.036	0.190	0.064	1.62	3.21	1.62
L_1	0.080	0.019	0.139	0.047	1.74	3.24	1.60
HI	0.071	0.020	0.131	0.048	1.84	5.70	1.87

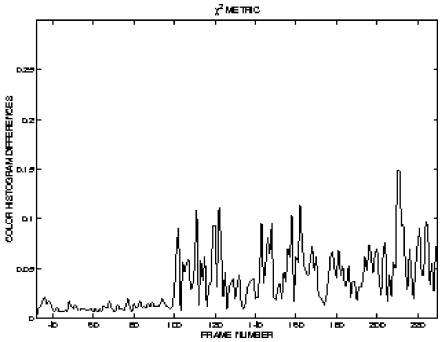


Fig. 4. Color histogram differences between H_t and H_{ref}, calculated with χ^2 measure, using segmentation maps shifted by ±10 pixels, starting from frame 100.

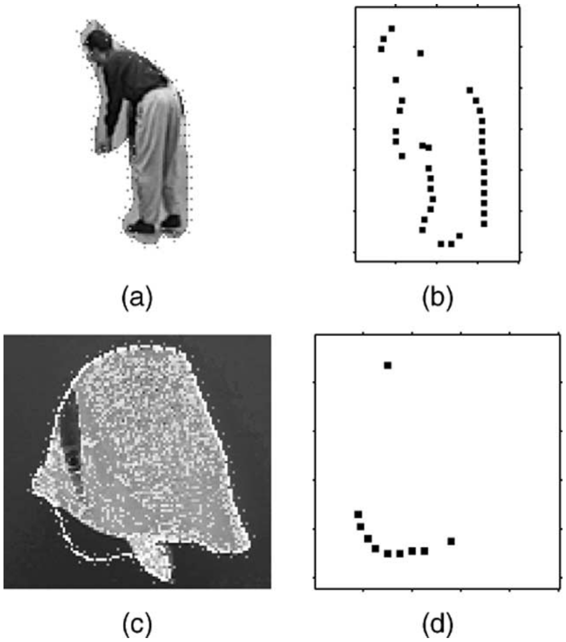


Fig. 5. (a) Video object plane for the 134th frame of the Hall monitor sequence which is downloaded from the web page of COST 211 group. The center probe blocks are marked with dots. (b) Incorrectly segmented regions of the boundary are marked with black boxes. (c) A segmentation of the 123rd frame of the Bream sequence. (d) Incorrectly segmented regions are marked with black squares.

or the correlations of the NGT performance measures in set $\mathbf{X}$ on the canonical variates of set $\mathbf{Y}$ of NGT measures (interset loadings).

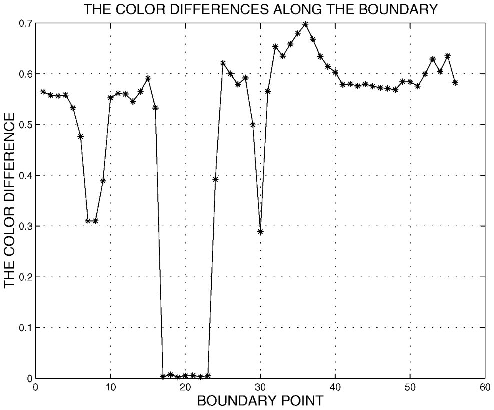


Fig. 6. Color differences $\delta_{\text{color}}(t,i)$ along the boundary of the segmented Bream object in Fig. 5(c).

There is a straightforward relation between the canonical weights ($\mathbf{a}_i$ or $\mathbf{b}_i$) and the loadings [22]. Using the symbol $\mathbf{s}_{\mathbf{X}X_1}$ to represent the vector of intraset loadings for the $\mathbf{X}$ set on its first canonical variate, we obtain

$$\mathbf{s}_{\mathbf{X}X_1} = \mathbf{\Sigma}_{\mathbf{XX}}\mathbf{a}_1 \tag{31}$$

where $\mathbf{a}_1$ is the weight vector and $\mathbf{\Sigma}_{\mathbf{XX}}$ is the matrix of correlations between variables of that set. The interset loadings can be computed similarly using cross-correlation matrices. For example, the vector of correlations of the $\mathbf{Y}$ set measures with the first canonical variate of the $\mathbf{X}$ set is

$$\mathbf{s}_{\mathbf{Y}X_1} = \mathbf{\Sigma}_{\mathbf{YX}}\mathbf{a}_1. \tag{32}$$

Once the loadings have been computed, it is easy to obtain a measure of the association between the two sets of measures. The squared interset loadings give the proportion of each measure's variance that is accounted for by a canonical variate of the other set. Therefore, the mean of square interset loadings for a given component is its redundancy. That is, the proportion of variance in set $\mathbf{X}$ that is related to the jth component of set $\mathbf{Y}$ is

$$Red_{\mathbf{X}Y_j} = \mathbf{s}_{\mathbf{X}Y_j}^T \mathbf{s}_{\mathbf{X}Y_j} = \frac{1}{p}\sum_{i=1}^{p} s_{X_iY_j}^2 \tag{33}$$

where $\mathbf{s}_{\mathbf{X}Y_j}$ is the vector of interset loadings of the measures in the set $\mathbf{X}$ with the jth measure of the $\mathbf{Y}$ set. The total redundancy of one set given the other is the sum of the redundancies of the individual components. In the context of our problem, we

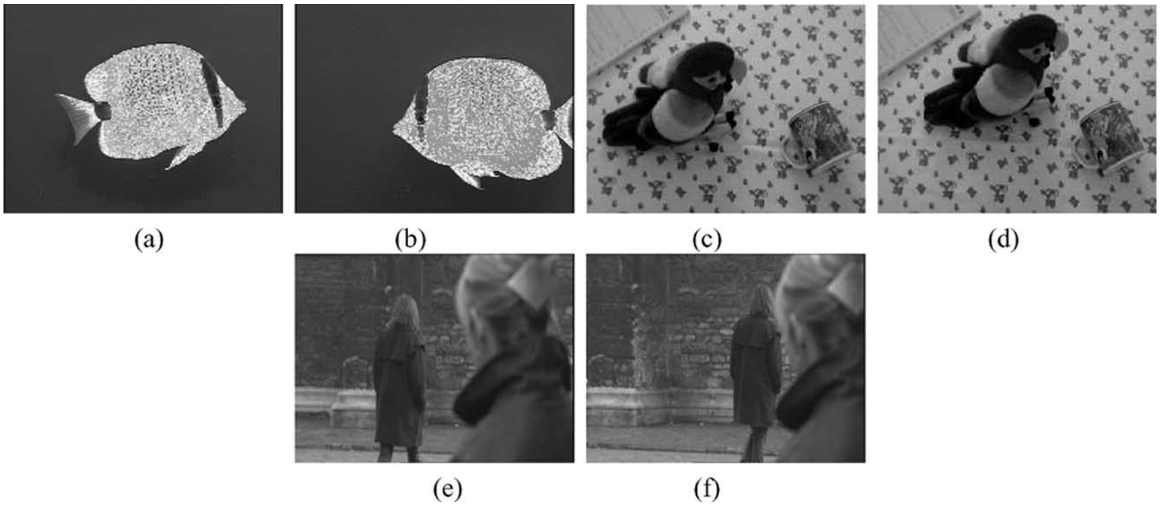


Fig. 7. (a), (b) Frames 100 and 130 of the "Bream" sequence. (c), (d) Frames 1 and 18 of the "Parrot" sequence. (e), (f) Frames 453 and 483 of the "Flikken" sequence.

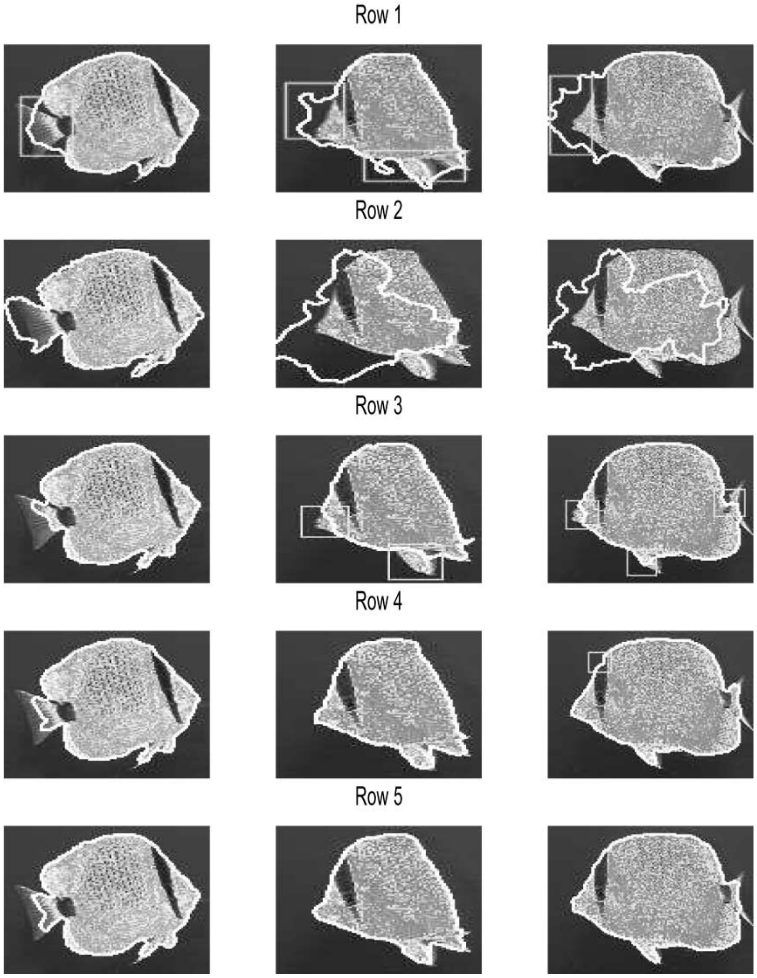
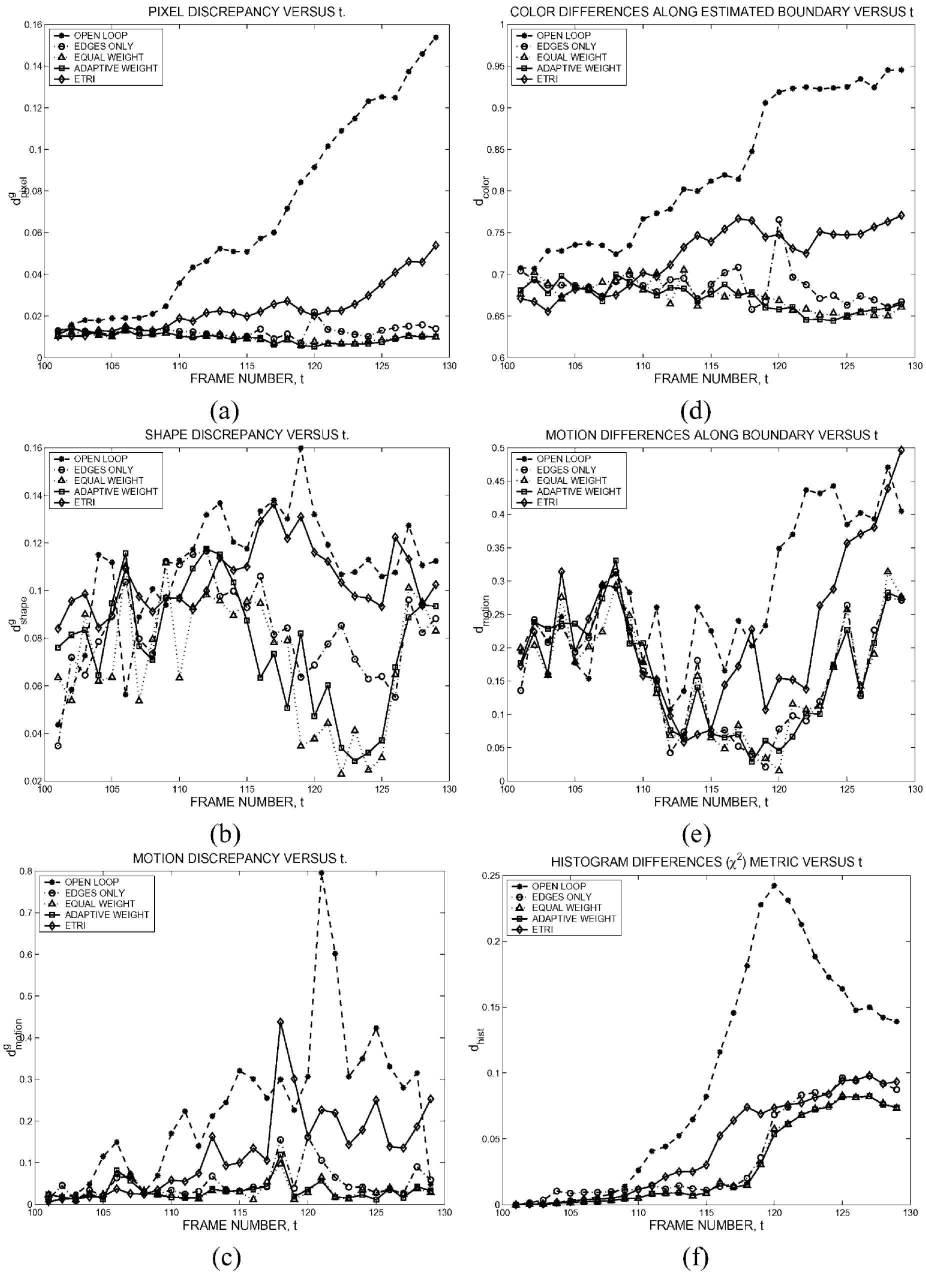


Fig. 8. Tracking results for frames 109, 121, and 128 (with some zoom in). Row 1: ETRI results. Row 2: Open-loop method. Row 3: Closed loop with edge energy only. Row 4: Closed loop with equal weighting. Row 5: Closed loop with adaptive weighting methods.

want to determine the "redundancy" of the GT data given the NGT measure set.

There are two major reasons to turn our attention to the loadings. First, the loadings are bounded by plus and minus 1 and are standardized across canonical variates. Neither is the case for the canonical weights. Second, the loadings appear to be less affected by the correlations among the variables as compared to the canonical weights [22].

More explicitly, a variable may receive a small weight simply because it is highly correlated with another variable in its set,



(a) (d)

(b) (e)

(c) (f)

Fig. 9. Measures with GT for "Bream" sequence: (a) the misclassification penalty $DP(t)$ for each frame; (b) the shape penalty $DS(t)$; (c) the motion penalty $DM(t)$. Measures without GT: (d) color differences along boundary; (e) motion differences along boundary; (f) inter-frame histogram differences using χ^2 measure.

even though both variables have high correlations with the canonical variate [22]. Hence, when another variable is added or removed from the set, the weight for a particular variable may change drastically. However, the canonical loadings are expected to be more stable.

V. Experimental Results

A. Sensitivity Analysis and Parameter Selection

In order to understand which histogram difference calculation method discussed in Section II-B is more sensitive to segmentation errors and also to tune the parameters, e.g., (L) of the performance measures, we performed some experiments. We used the GT segmentation maps of the well-known Hall Monitor (frames 32–230), Bream (frames 100–130), and Stefan (frames 145–175) sequences. Sample frames of these sequences together with their GT segmentation maps are shown in Fig. 3. In order to simulate incorrect segmentation, the GT segmentation maps have been randomly shifted in each frame by a maximum of ± 10 pixels in horizontal and vertical directions [10].

The results for the performance measure based on histogram differences for the three sequences are summarized in Table I. We can observe that χ^2 is the most sensitive metric to the perturbations in the segmentation map since the percentage increase in D_{hist}, and in the variance of $d_{\text{hist}}(t)$ ($\sigma_{d_{\text{hist}}}$) are largest for the χ^2 measure. Note that both the mean and the variance increase

TABLE II
MEAN OF PERFORMANCE EVALUATION SCORES FOR THE "BREAM" SEQUENCE

	Ground-truth				No Ground Truth			
	D^g_{pixel}	D^g_{shape}	D^g_{motion}	D^g	D_{hist}	D_{color}	D_{motion}	D
	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combi.	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combi.
Open Loop	6.71	1.09	2.46	1.00	9.68	8.25	2.82	1.00
ETRI	2.32	1.04	1.24	0.61	4.39	7.21	2.21	0.70
Edges Only	1.26	0.83	0.51	0.38	3.74	6.85	1.61	0.60
Equal Weight	0.96	0.77	0.32	0.31	2.99	6.73	1.57	0.56
Adapt. Weight	0.94	0.70	0.31	0.32	2.97	6.71	1.60	0.56

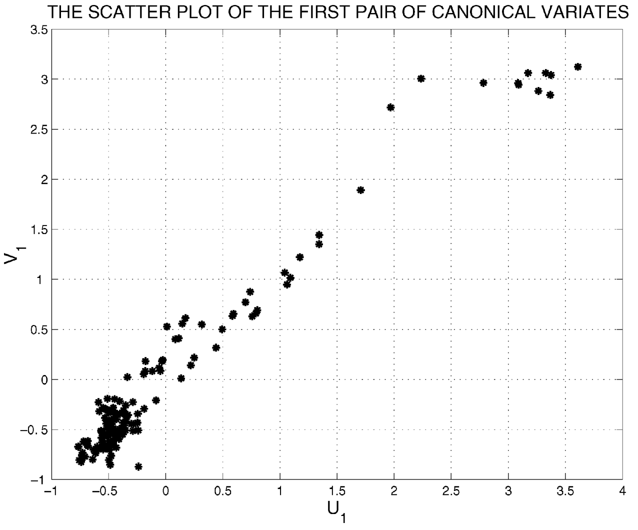


Fig. 10. Scatter plot of the first canonical variate pair.

significantly. Normally, the variance of $d_{\text{hist}}(t)$ is 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.

Another way of simulating incorrect segmentation is to introduce flicker to the video object planes by deleting a rectangular block randomly [10]. The same ratio tests for this type of distortion also indicated that the χ^2 histogram measure is more sensitive than the other measures.

In order to determine optimal values for the L (probe length) which was shown in Fig. 1(c), we calculated the D_{color} and D_{motion} measures defined in (3) and (15) for different values of L ($L = 2, \ldots, 7$). We observed the maximum ratio of the D_{color} and D_{motion} values (similar to Table I) for the Hall monitor, Stefan and Bream sequences. The experiments indicated that a value of $L = 3$, gave the highest ratio for the D_{color} measure, and $L = 5$, gave the highest ratio for D_{motion} measure. A longer probe length is needed for the D_{motion} measures, since motion estimation is not perfect especially around object boundaries. Therefore, during the experiments we use a probe length of $L = 3$, for calculation of the D_{color} measure, and a probe length of $L = 5$, for calculation of the D_{motion} measure. A reasonable value for M was found to be 3.

B. Localization of Incorrect Segmentation

In order to demonstrate the *temporal localization* capability of the proposed measures, we chose $\mu_1 = 0$, $\mu_2 = 1$, and $\mu_3 = 0$ in (20). In Fig. 4, a plot of the χ^2 measure for the Hall monitor sequence is given, which is calculated using the GT segmentation masks up to frame 100 and with perturbed segmentation masks for frames 101–230. As seen in Fig. 4, the histogram difference measure based on χ^2 distance calculation can signal the onset of the perturbed segmentation masks accurately.

Temporal localization of incorrect segmentation is signaled by a big jump in the plot of $d(t)$ can be seen in Fig. 4. Therefore, for the determination of the threshold T_d [introduced in (20)], a mean value of $d(t)$ up to time t can be calculated, and an incoming $d(t+1)$ that exceeds this mean by a certain magnitude can be marked as a badly segmented frame. However, a threshold that is more correlated with human judgements (a just noticeable value that creates the jitter sensation) can only be determined by extensive perceptual experiments, which requires further research.

The performance measures are also capable of *spatial localization* by utilizing (21). In Fig. 5(a) and (b), we show the video object plane for the 134th 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 incorrectly segmented boundary segments are marked with black boxes. The measure (21) is able to support the subjective observations quantitatively. Another example is given for the Bream sequence in Fig. 5(c) and (d). In order to achieve these results, we obtained two binary images by choosing the parameters in (21) as $\mu_4 = 1$, $\mu_5 = 0$ and $\mu_4 = 0$, $\mu_5 = 1$ with $k = 1$, that is, the threshold set at one standard deviation of the scores. Then, we AND these two binary images to obtain the final localization. A plot of the color differences along the object boundary is given in Fig. 6.

C. Canonical Correlation Analysis

In this section, we present the experimental results for the proposed performance evaluation measures, based on three video sequences. The first video sequence is the well-known "Bream" sequence, two frames of which are shown in Fig. 7. The object to be tracked is the fish swimming toward right and then turning toward left, causing a lot of self occlusion. However, the background is not cluttered.

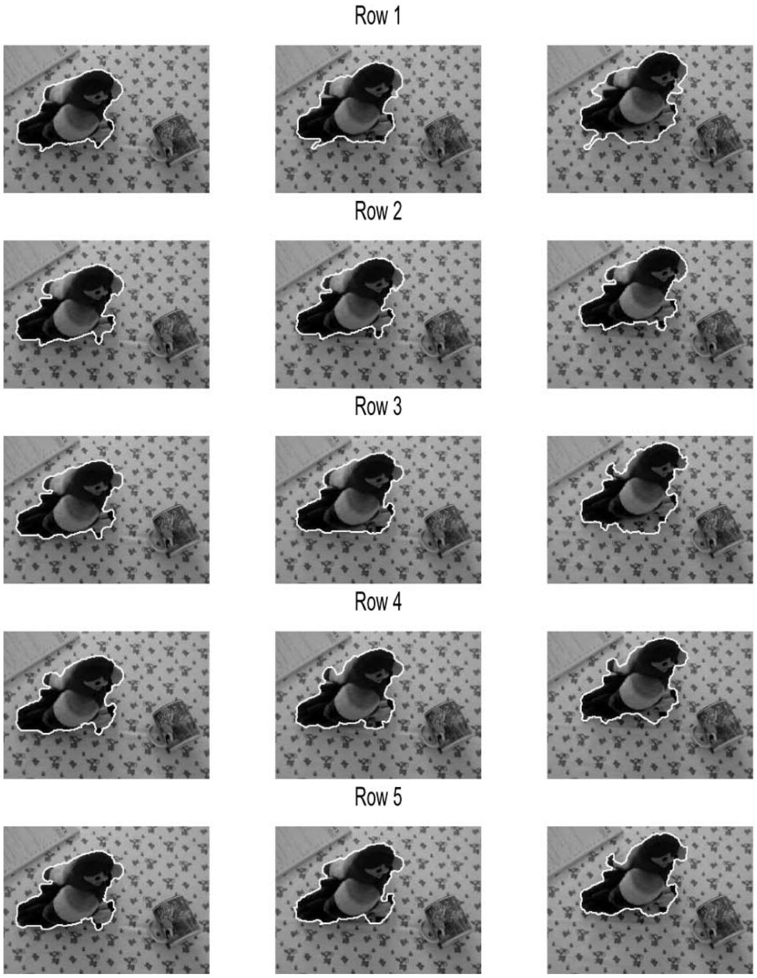


Fig. 11. Tracking results for frames 2, 8, and 18. Row 1: ETRI results. Row 2: Open-loop method. Row 3: Closed loop with edge energy only. Row 4: Closed loop with equal weighting. Row 5: Closed loop with adaptive weighting methods.

The second sequence is called the “Parrot” sequence, sample frames of which are shown in Fig. 7. In this sequence, the rigid parrot object translates a total of (26, -20) pixels over 18 frames. The background, on the other hand, is very cluttered in this sequence.

The third sequence is a 30 frame section of a TV series called “Flikken”[1], the first and the last frames of which can be seen in Fig. 7. There are two foreground objects in this scene. The lady on the left will be named as object 1, and the lady on the right will be named as object 2.

1) “Bream” Results: The tracking results for frames 100–130 are obtained using five different object tracking techniques (open loop, edges only, equal weight, adaptive weight, and ETRI). The first four techniques are obtained from intermediate steps of the video object tracking algorithm described in [23], [24] and the last technique was developed at Electronics and Telecommunications Research Institute (ETRI), Korea [25], [26]. The ETRI method is a semiautomatic object segmentation tool which consists of two steps: intra-frame segmentation and inter-frame segmentation. In the intra-frame segmentation step, the user draws the object boundary manually. In the inter-frame segmentation, the regions defined by the user are tracked by using multiple affine motion models.

Sample tracking results are given in Fig. 8, where incorrect boundary segmentations are pointed out using a square marker. Visual inspection of the results reveals that closed-loop results (row 5) with adaptive weighting are the best, closely followed by the equal weighting results (row 4). The method using edge energy only is third in the rank (row 3) and the worst results are obtained by ETRI (row 1) and open-loop methods (row 2).

In Fig. 9, we give the performance evaluation measures using the GT and NGT measures, respectively. Table II summarizes the performance evaluation measures by providing the mean values over all frames. We can observe in Table II that the ETRI and open loop methods have the worst scores in all GT and NGT measures, which is in agreement with our subjective evaluations. The combined performance scores in the last column are obtained by equal weight averaging after normalizing each column by its maximum value. As the GT measures deteriorate, there is a commensurate increase in the NGT measures. This correlation is especially strong between misclassification penalty d^{g}_{pixel} and the measure of color differences along the object boundary d_{color} [compare Fig. 9(a) and (d)].

[1] A TV series produced by MMGNV for VRT.

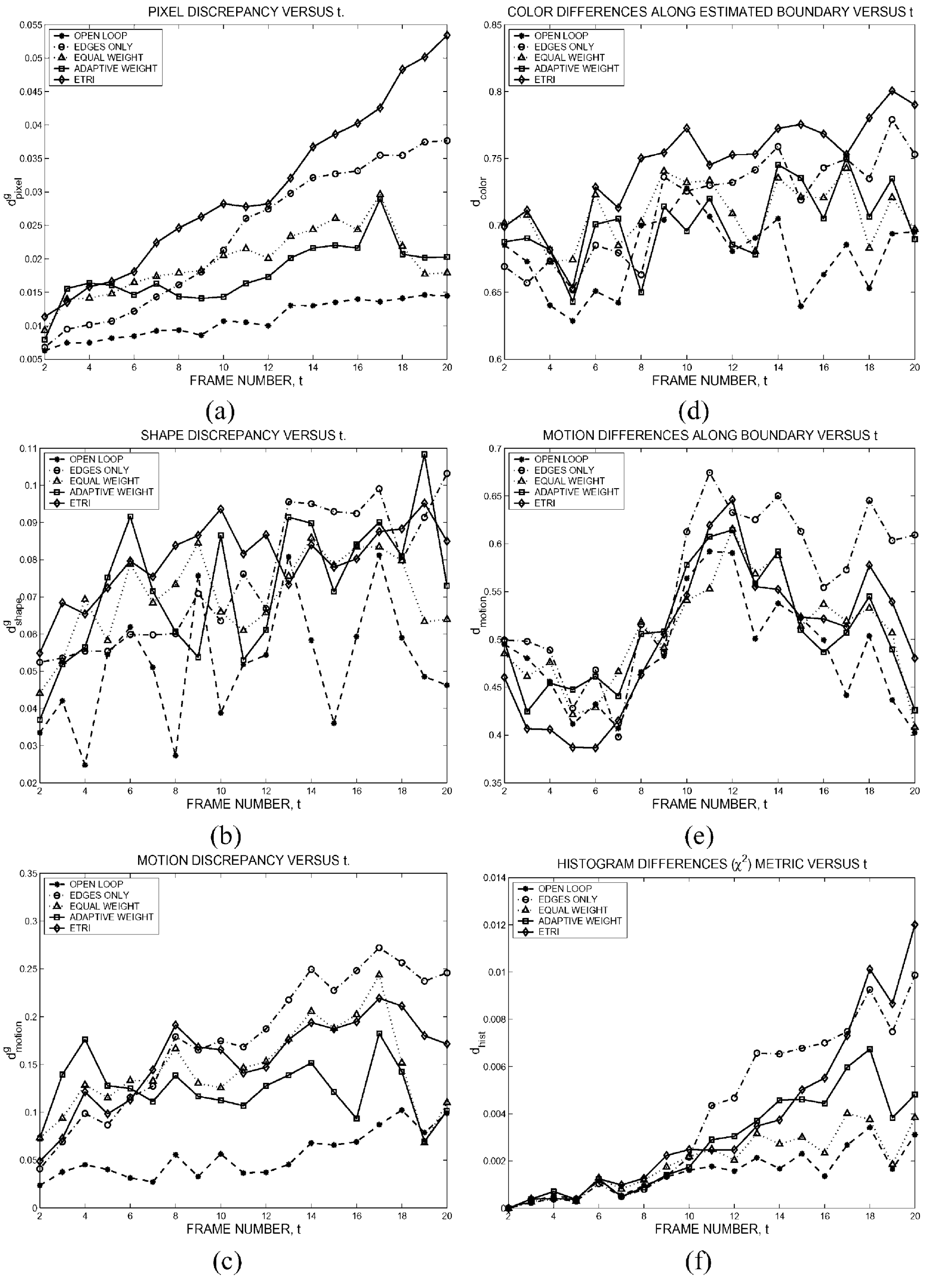


Fig. 12. Measures with GT for "Parrot" sequence: (a) the misclassification penalty $d^g_{\text{pixel}}(t)$ for each frame; (b) the shape penalty $d^g_{\text{shape}}(t)$; (c) the motion penalty $d^g_{\text{motion}}(t)$. Measures without GT: (d) color differences along boundary $d_{\text{color}}(t)$ in YCbCr color space; (e) motion differences along boundary $d_{\text{motion}}(t)$; (f) inter-frame histogram differences using χ^2 measure $d_{\text{hist}}(t)$.

In order to quantify the correlation, we have calculated the correlation matrix which was defined in (29)

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc|ccc} d^g_{\text{pixel}} & d^g_{\text{shape}} & d^g_{\text{motion}} & d_{\text{hist}} & d_{\text{color}} & d_{\text{motion}} \\ \hline 1.00 & 0.47 & 0.82 & 0.74 & 0.95 & 0.61 \\ 0.47 & 1.00 & 0.48 & 0.27 & 0.56 & 0.26 \\ 0.82 & 0.48 & 1.00 & 0.73 & 0.86 & 0.46 \\ \hline 0.74 & 0.27 & 0.73 & 1.00 & 0.70 & 0.45 \\ 0.95 & 0.56 & 0.86 & 0.70 & 1.00 & 0.55 \\ 0.62 & 0.26 & 0.46 & 0.45 & 0.55 & 1.00 \end{array}\right]. \tag{34}$$

We can observe that there are significant correlations (larger than 0.5) between d^g_{pixel} and d_{hist} (0.74); d^g_{pixel} and d_{color} (0.95); d^g_{pixel} and d_{motion} (0.61); d^g_{motion} and d_{hist} (0.73); and d^g_{motion} and d_{color} (0.86). If we carry out the canonical correlation analysis as discussed before to find the pair of linear transformations that maximize the correlation between the GT and NGT measures, we get the following pair of transformation coefficients:

$$\mathbf{a}_1 = [0.77 \quad 0.08 \quad 0.21]^T, \mathbf{b}_1 = [0.14 \quad 0.84 \quad 0.09]^T.$$

The first canonical variate (composite) of the set of GT measures assigns the largest weight (0.77) to the misclassification penalty and about one fourth of it (0.21) to the motion discrepancy. Shape discrepancy is discarded (0.08). From the NGT measure set, the canonical variate (composite) is constructed with

TABLE III
MEAN OF PERFORMANCE EVALUATION SCORES FOR THE "PARROT" SEQUENCE

	Ground-truth				No Ground Truth			
	D^g_{pixel}	D^g_{shape}	D^g_{motion}	D^g	D_{hist}	D_{color}	D_{motion}	D
	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-1}$	Combi.	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combi.
Open Loop	1.09	5.18	0.54	0.44	0.15	6.77	4.85	0.71
ETRI	3.03	8.00	1.55	0.96	0.37	7.45	5.00	0.94
Edges Only	2.35	7.49	1.77	0.90	0.40	7.15	5.56	0.99
Equal Weight	1.95	7.03	1.44	0.78	0.20	7.09	5.06	0.78
Adapt. Weight	1.78	7.30	1.24	0.73	0.27	7.01	5.08	0.84

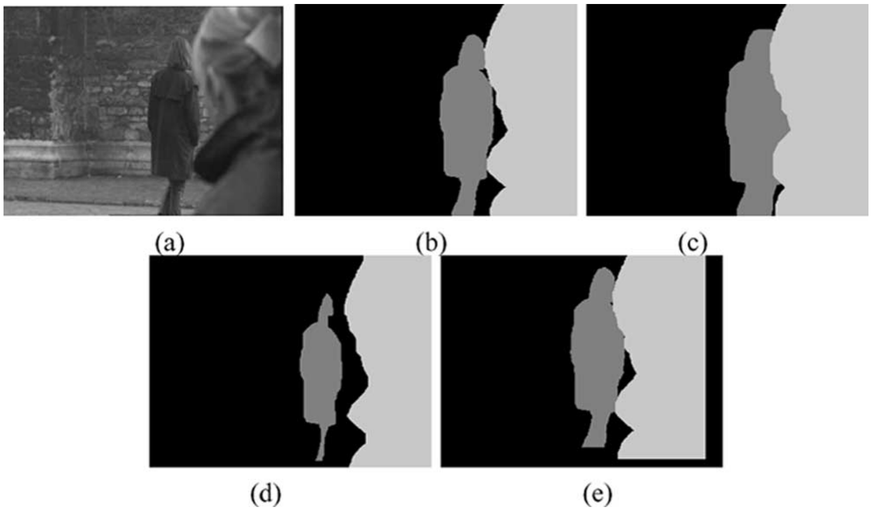


Fig. 13. (a) Frame 483 of the "Flikken" sequence. (b) The GT segmentation map. (c), (d), (e) The distorted segmentation obtained through dilation, erosion, and random shifting, respectively.

the color difference measure (0.84) and the histogram measure (0.14) with motion information discarded (0.09). The less conclusive evidences from shape discrepancy in the first set and the motion measure in the second set are also obvious in Fig. 9(b) and (e), respectively. The reason that shape and motion measures have small weights can be due to their small correlations with the other measures in the set. For example, the correlation of shape measure with the histogram measure is small (0.27) as seen in the correlation matrix (34).

The most important result of the canonical analysis is the fact that, the maximum correlation between the first canonical variate pair, consisting of linear combinations of GT and NGT measures (ρ_1) is 0.98. The square of it $\rho_1^2 = 0.96$ expresses the proportion of variance in each composite (canonical variate) that is related to the other composite (variate) of the pair consisting of the linear combinations of GT and NGT measures. The scatter plot of the first canonical variate pair is given in Fig. 10. This high canonical correlation implies that the set of NGT measures reflect the information contained in the set of GT measures.

We also carried out the redundancy computation through canonical loadings as was discussed in Section IV-D. As a result, we observed that 66% of the variance in the set of GT measures is accounted for by the set of NGT measures. On the other hand, 65% of the variance in the set of NGT measures was found to be accounted for by the set of GT measures.

2) "Parrot" Results: Several tracking results for the Parrot sequence are shown in Fig. 11. If we analyze the tracking results in Fig. 11, we can say that the results of ETRI (row 1) and edges only (row 3) are the worst and open loop results (row 2) are the best. The plots for the GT measures and the NGT measures are given in Fig. 12. The quantitative results of the "Parrot" sequence are summarized in Table III together with the combined measure. The quantitative evaluation results show that ETRI and edges only results get the highest (worst) scores and the open loop results get the lowest (best) scores, although the exact ordering of ETRI and edges only methods is different for GT and NGT measures.

The canonical correlation analysis for this sequence yields a maximum correlation of $\rho_1 = 0.93$ with the following transformation parameters:

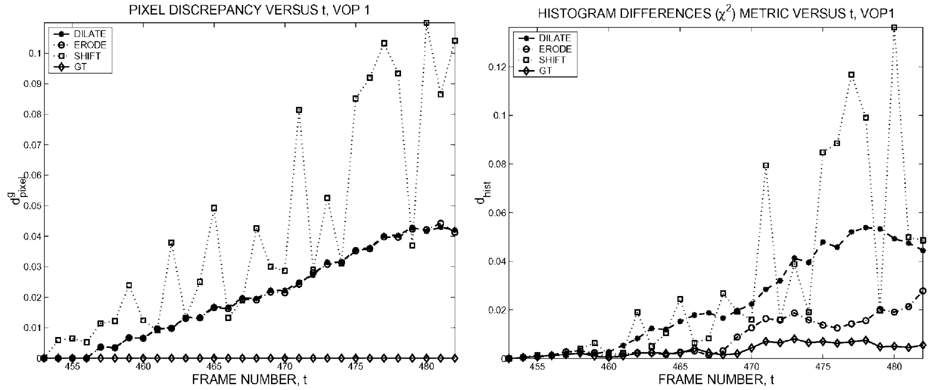
$$\mathbf{a}_1 = [1.36 \quad 0.09 \quad 0.41]^T, \mathbf{b}_1 = [0.84 \quad 0.38 \quad 0.02]^T.$$

The computation of redundancy through canonical loadings revealed that 62% of the variance in the set of GT measures is accounted for by the set of NGT measures. However, 56% of the variance in the set of NGT measures is accounted for by the set of GT measures.

3) "Flikken" Results: The GT segmentation map of the last frame of the "Flikken" sequence together with the evaluated segmentation results are shown in Fig. 13. In order to test the

TABLE IV
COMBINED PERFORMANCE EVALUATION SCORES FOR THE "FLIKKEN" SEQUENCE

	Ground-truth			No Ground Truth		
	Object 1	Object 2	Overall D^g	Object 1	Object 2	Overall D
Dilate	0.59	0.50	0.55	0.70	0.74	0.72
Erode	0.57	0.39	0.48	0.66	0.58	0.62
Shift	0.81	0.96	0.89	0.99	0.98	0.99
Ground Truth	0	0	0	0.44	0.50	0.47



(a) (b)

Fig. 14. Measures for "Flikken" sequence: (a) the misclassification penalty $d^g_{\text{pixel}}(t)$ for each frame for object 1; (b) inter-frame histogram differences using χ^2 measure $d_{\text{hist}}(t)$ for object 1.

performance measures under distortions different from the ones that have been introduced by the tracking algorithms we considered, we distorted the GT segmentation maps with increasing amounts of dilation, erosion and random shifts to obtain the segmentation results to be evaluated. The GT and NGT scores for object 1 (lady on the left) and object 2 (lady on the right) are summarized in Table IV. The combined scores are obtained by averaging the scores of object 1 and object 2, assuming that they are of equal importance. It can be seen from the table that GT and NGT measures give the same ordering of the segmentation results. In Fig. 14, the plots of pixel misclassification penalty ($d^g_{\text{pixel}}(t)$) and the inter-frame histogram difference measure ($d_{\text{hist}}(t)$) for object 1 are given, which have the highest correlation value in the correlation matrix (0.85) as can also be observed in Fig. 14.

The canonical correlation analysis for object 1 yields a maximum correlation of $\rho_1 = 0.95$ with the following transformation parameters:

$$\mathbf{a}_1 = [0.95 \quad 0.05 \quad 0.08]^T, \mathbf{b}_1 = [0.54 \quad 0.08 \quad 0.48]^T.$$

The computation of redundancy through canonical loadings revealed that 60% of the variance in the set of GT measures is accounted for by the set of NGT measures, and 72% of the variance in the set of NGT measures is accounted for by the set of GT measures.

The canonical correlation analysis for object 2 gives a maximum correlation of $\rho_1 = 0.89$ with the following transformation parameters:

$$\mathbf{a}_1 = [0.88 \quad 0.04 \quad 0.13]^T, \mathbf{b}_1 = [0.52 \quad 0.08 \quad 0.58]^T.$$

The computation of redundancy through canonical loadings revealed that 49% of the variance in the set of GT measures is accounted for by the set of NGT measures, and 51% of the variance in the set of NGT measures is accounted for by the set of GT measures.

VI. CONCLUSION

We presented three NGT measures for quantitative performance evaluation of video object segmentation and tracking algorithms. The proposed measures yield a figure of merit for the whole segmented video sequence or, in turn, can give more local results, such as per frame scores or per object scores. Thus it is possible to identify within a given video sequence the frames that are poorly segmented, or even parts of an object within a frame with poorly defined boundary. The proposed measures for segmentation quality can be easily extended to the case of multiple foreground objects. We have analyzed the correlation between the three proposed NGT measures and a set of three GT measures. We have shown experimentally that they are significantly correlated, implying that NGT measures can be reliably used for performance monitoring in lieu of GT measures under certain assumptions. Thus the extremely tedious and time-consuming task of GT extraction can be avoided.

REFERENCES

[1] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[2] M. Borsotti, P. Campdelli, and P. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognit. Lett.*, vol. 19, pp. 741–747, 1998.

[3] Y. J. Zhang, "Evaluation and comparison of different segmentation algorithms," *Pattern Recognit. Lett.*, vol. 18, pp. 963–974, 1997.

[4] M. D. Levine and A. M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, pp. 155–164, Feb. 1985.

[5] G. Goehrig and L. Ledford, "Analysis of image segmentation approaches with emphasis on performance evaluation criteria," *Proc. SPIE*, vol. 252, pp. 124–129, 1980.

[6] G. Rees, P. Greenway, and D. Morray, "Metrics for image segmentation," in *Proc. SPIE Conf. Visual Information Processing VII*, vol. 3387, 1998.

[7] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, "Error measures for scene segmentation," *Pattern Recognit.*, vol. 9, pp. 217–231, 1977.

[8] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. 2nd IEEE Workshop Performance Evaluation of Tracking and Surveillance*, 2000.

[9] C. E. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X Eur. Signal Processing Conf.*, vol. 2, Sept. 2000, pp. 917–920.

[10] X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. X Eur. Signal Processing Conf.*, vol. 4, Sept. 2000.

[11] P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Sept. 2000, pp. 308–311.

[12] P. L. Correia and F. P. Pereira, "Stand-alone objective segmentation quality evaluation," *EURASIP J. Appl. Signal Processing*, no. 4, pp. 390–402, 2002.

[13] ——, "Objective evaluation of video segmentation quality," *IEEE Trans. Image Processing*, vol. 12, pp. 186–200, Feb. 2003.

[14] J. Bednar and T. L. Watt, "Alpha-trimmed means and their relationship to median filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 145–153, Feb. 1984.

[15] A. M. Ferman, A. M. Tekalp, and R. Mehrotra, "Robust color histogram descriptors for video segment retrieval and identification," *IEEE Trans. Image Processing*, vol. 11, pp. 497–508, May 2002.

[16] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flanney, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[17] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 11, pp. 11–32, 1991.

[18] 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, pp. 2051–2060, Dec. 2000.

[19] G. Borgefors, "Distance transformations in digital images," *Comput. Vis., Graph. Image Processing*, vol. 34, pp. 344–371, 1986.

[20] E. M. Arkin, L. P. Chew, D. P. Huttenlocker, K. Kedem, and J. S. B. Mitchell, "An efficient computable metric for comparing polygonal shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 209–215, Jan. 1991.

[21] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[22] H. E. A. Tinsley and S. D. Brown, *Handbook of Applied Multivarite Analysis and Mathematical Modeling*. New York: Academic, 2000.

[23] C. E. Erdem, A. M. Tekalp, and B. Sankur, "Non-rigid object tracking with feedback of performance evaluation measures," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, vol. 2, dec. 2001, pp. 323–330.

[24] ——, "Video object tracking with feedback of performance measures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 310–324, June 2003.

[25] J. G. Choi, "A user assisted segmentation method for video object plane segmentation," in *Proc. Int. Conf. Circuits/Systems, Computers, Communications*, Korea, 1998, pp. 7–10.

[26] M. Kim, J. G. Choi, M. H. Lee, and C. Ahn, "User's Guide for a User-Assisted Video Object Segmentation Tool," ISO/IEC JTC1/SC29/WG11 MPEG98/m3935, 1998.

Çiğdem Eroğlu Erdem (S'93–M'03) received the Ph.D. degree in electrical engineering from Boğaziçi University, İstanbul, Turkey, in 2002.

In 2001, she was a Visiting Researcher at the University of Rochester, Rochester, NY. From 2003 to 2004, she was a Postdoctoral Fellow, Delft University of Technology, The Netherlands, where she was also affiliated with Philips Research Laboratories, Eindhoven, The Netherlands. She is currently with Momentum Digital Media Technologies, Inc., İstanbul. Her research interests are in the area of digital images, video and speech processing, including motion estimation, video segmentation, object tracking, and speech processing for natural character animation.

Bülent Sankur received the M.Sc. and Ph.D. degrees from Rensselaer Polytechnic Institute, Troy, NY.

He has been active in the Department of Electric and Electronic Engineering, Boğaziçi University, İstanbul, Turkey, establishing curricula and laboratories and guiding research in the areas of digital signal processing, image and video compression, and biometry and multimedia systems. He has held visiting positions at the University of Ottawa, Ottawa, ON, Canada, İstanbul Technical University, the Technical University of Delft, Delft, The Netherlands, and the Ecole Nationale Supérieure des Telecommunications, France.

Dr. Sankur was the Chairman of the 1996 International Telecommunications Conference and the technical Co-Chairman of ICASSP 2000.

A. Murat Tekalp (S'80–M'84–SM'91–F'03) received the Ph.D. degree in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1984.

From December 1984 to August 1987, he was with Eastman Kodak Company, Rochester, NY. He joined the Electrical and Computer Engineering Department, University of Rochester, in September 1987, where he is currently a Distinguished Professor. Since June 2001, he has been with Koç University, İstanbul, Turkey. His research interests are in the areas of digital image and video processing, including video compression and streaming, video filtering for high resolution, and video segmentation. He is the author of the book *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall).