Performance Measures for Video Object Segmentation and Tracking

Çiğdem Eroğlu Erdem[a], Bülent Sankur[a], A. Murat Tekalp[b]

[a] Dept. Electrical and Electronics Engineering, Boğaziçi University, İstanbul, Turkey

[b] Dept. Electrical and Computer Engineering, University of Rochester, NY 14627, USA
College of Engineering, Koç University, Rumelifeneri Yolu, Sarıyer, İstanbul, Turkey

ABSTRACT

We propose measures to evaluate the performance of video object segmentation and tracking methods quantitatively without ground-truth 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 neighbors. They can be used to localize (spatially and/or temporally) regions where segmentation results are good or bad; and/or 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 ground truth* have been demonstrated by canonical correlation analysis of the proposed measures with another set of measures *with ground-truth* on a set of sequences (where ground truth information is available). Experimental results are presented to evaluate the segmentation maps obtained from various sequences using different segmentation and tracking algorithms.

Keywords: Object segmentation, object tracking, performance evaluation, canonical correlation analysis

1. INTRODUCTION

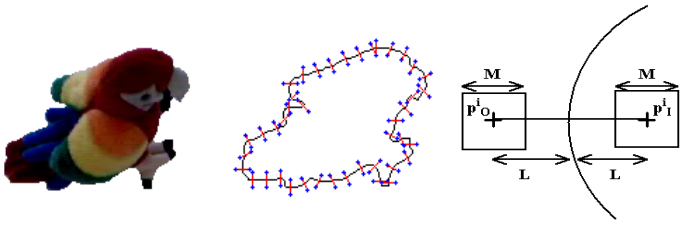
Object-based video segmentation and object tracking are challenging and active research areas in digital video processing and computer vision. The task of segmenting/tracking a video object emerges in many applications such as object-based video coding (e.g., MPEG-4), content-based video indexing and 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 judgement, which is qualitative and time consuming. Therefore, there is need for automatic, objective spatio-temporal measures, not only for comparison of overall algorithmic performance, but also as a tool to monitor spatio-temporal consistency of individual objects. Recently, a number of video segmentation measures have been proposed in the presence of ground-truth.[1–4] The usefulness of these measures is limited in that they require the presence of ground-truth information which is not easily available.

The main contribution of this work is to develop quantitative performance measures for video object tracking and segmentation, which do not require ground-truth segmentation maps. The proposed 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 is color histogram differences across video object planes, which evaluates the goodness of segmentation along a spatio-temporal trajectory. The third feature is based on motion vector differences along the 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, depending upon the scene content 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 2, 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

This work was supported by Scientific and Technical Research Council of Turkey (TÜBİTAK-BAYG) and Boğaziçi University Research Fund (99A203). E-mails: Ç.E.E.: cigdem@ieee.org, B.S.: sankur@boun.edu.tr, A.M.T.: tekalp@ece.rochester.edu.



(a) (b) (c)

Figure 1. (a) A video object plane for the "Parrot" sequence. (b) The boundary of the video object plane with the normal lines. (c) A close-up 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_O^i, respectively.

are developed in Section 3. In Section 4, Canonical correlation analysis is used to validate non-ground-truth (NGT) measures against ground-truth (GT) measures. In Section 5, experimental results are provided. Finally, in Section 6, conclusive remarks are given.

2. 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) Background is either stationary or has global motion which can be compensated for. In the following, we present two color and one motion features, which quantify the above assumptions in order to evaluate the goodness of a segmented video object plane.

2.1. Spatial Color Contrast Along Object Boundary

Since the 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 Figure 1(b). As depicted in the close-up in Figure 1(c), the color probes are $M \times M$ regions centered at the two ends of the i^{th} normal line denoted by p_I^i (inside pixel) and p_O^i (outside pixel), which are marked with plus signs.

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

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

$$\delta_{color}(t;i) = \frac{\|C_O^i(t) - C_I^i(t)\|}{\sqrt{3 \times 255^2}} \quad (2)$$

where, K_t is the total number of normal lines drawn on the object boundary 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 the Y-Cb-Cr color space. The average inside color $C_I^i(t)$ is defined similarly. The worst score is 1 and $d_{color}(t)$ decreases towards zero as the color contrast along object boundary increase, possibly indicating a good segmentation.

We define the color measure for the whole sequence as:

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

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,[5] median or the maximum function.

When the location of the object boundary is estimated correctly in frame t, we expect the color measure $d_{color}(t)$ to take a small value. However, the converse of this statement is not necessarily true. That is, if the color measure $d_{color}(t)$ 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.

2.2. 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 first (reference) 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,[6,7] namely the L_1, L_2, χ^2 and histogram intersection measures. The χ^2 metric is used to compare two binned data sets, and to determine if they are drawn from the same distribution function.[7] It is defined and normalized to the range $[0, 1]$ as follows:

$$0 \leq \chi^2(H_t, H_{ref}) = \frac{\sum_{j=1}^{B} \frac{[r_1 H_t(j) - r_2 H_{ref}(j)]^2}{H_t(j) + H_{ref}(j)}}{N_{H_t} + N_{H_{ref}}} \leq 1. \tag{4}$$

where the following definitions are used:

$$r_1 = \sqrt{\frac{N_{H_{ref}}}{N_{H_t}}}, \quad r_2 = \frac{1}{r_1}, \quad N_{H_t} = \sum_{j=1}^{B} H_t(j), \quad N_{H_{ref}} = \sum_{j=1}^{B} H_{ref}(j).$$

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.

In order to choose the most sensitive histogram differencing metric, we conducted an experiment,[8] where a number of ground-truth 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 metric.[8] Therefore, we use the χ^2 metric, $d_{hist}(t) = \chi^2(H_t, H_{ref})$, in all other experiments. Note that if the two histograms being compared are identical, $d_{hist}(t) = 0$, and it increases towards 1, as the histograms differ more. We define the histogram difference measure for the whole sequence as:

$$0 \leq D_{hist} = f(d_{hist}(t)) \leq 1, \qquad t = 1, \ldots, T \tag{5}$$

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

2.3. 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 Figure 1(b) and (c) and consider the difference of the average motion vectors in the neighborhood the points p_O^i and p_I^i. The motion measure for frame t is estimated as follows:

$$0 \leq d_{motion}(t) = 1 - \frac{\sum_{i=1}^{K_t} \delta_{motion}(t;i)}{\sum_{i=1}^{K_t} w_i} \leq 1, \quad (6)$$

$$\delta_{motion}(t;i) = d(\mathbf{v}_O^i(t), \mathbf{v}_I^i(t)) \cdot w_i \quad (7)$$

$$0 \leq w_i = R(\mathbf{v}_O^i(t)) \cdot R(\mathbf{v}_I^i(t)) \leq 1, \quad (8)$$

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:

$$0 \leq d(\mathbf{v}_O^i(t), \mathbf{v}_I^i(t)) = 1 - exp\left(-\frac{\|\mathbf{v}_O^i(t) - \mathbf{v}_I^i(t)\|}{\sigma^2}\right) \leq 1. \quad (9)$$

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. In (8), $R(.)$ denotes the reliability[9] of the motion vector $\mathbf{v}^i(t)$ at point p^i:

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

where $\mathbf{b}^i(t+1)$ denotes the backward motion vector at location $\mathbf{p}^i + \mathbf{v}^i$ in frame $t+1$; $c(p^i;t)$ denotes the color intensity and the parameters σ_m, σ_c are chosen similarly as in.[9] 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):

$$0 \leq D_{motion} = f(d_{motion}(t)) \leq 1, \qquad t = 1, \ldots, T. \quad (10)$$

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 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 $d_{motion}(t)$ as $S(V_{med}(t))$, where $V_{med}(t)$ is the median of motion vector magnitudes along the boundary and $S(.)$ is the fuzzy weighting function introduced in (13). The midpoint of the ramp can be set at what we define as the "small motion threshold", for example $c_2 = 1$ pixels/frame.

3. PERFORMANCE MEASURES WITHOUT GROUND-TRUTH

In this section, we combine the color and motion measures to obtain scores that reflects the success of segmentation and tracking of objects for the whole sequence (Section 3.1), as well as temporal (Section 3.2) and spatial localization (Section 3.3) of incorrect boundary segments.

3.1. 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:

$$D = \mu D_{color} + \beta D_{hist} + \gamma D_{motion}, \quad (11)$$

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 $D = \max\{D_{color}, D_{hist}, D_{motion}\}$. A more lenient penalty function would be

$$D = \frac{\mu(D_{color})D_{color} + \mu(D_{hist})D_{hist} + \mu(D_{motion})D_{motion}}{\mu(D_{color}) + \mu(D_{hist}) + \mu(D_{motion})}, \tag{12}$$

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

$$S(x) = \begin{cases} 0, & x \leq c_1 \\ \frac{(x-c_1)^2}{(c_2-c_1)(c_3-c_2)}, & c_1 \leq x < c_2 \\ 1 - \frac{(x-c_3)^2}{(c_3-c_2)(c_3-c_1)}, & c_2 \leq x < c_3 \\ 1, & x \geq c_3 \end{cases} \tag{13}$$

For example, the parameters can be chosen as $c_1 = 0.2, c_2 = 0.5$ and $c_3 = 0.8$. The combination strategy in (12), gives more weight to large measures. As one of the three measures gets larger, its weight also becomes large and vice versa.

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

$$D = \max\{D_{O_i}\} \tag{14}$$

where $D_{O_i}, i = 1, \ldots, N_O$ is D measure as defined in (11) for the object O_i, and N_O is the number of objects in the video sequence.

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.

3.2. Temporal Localization

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

$$d(t) = \mu_1 d_{color}(t) + \mu_2 d_{hist}(t) + \mu_3 d_{motion}(t), \tag{15}$$

where μ_1, μ_2, μ_3 could be determined with a method as in (11) or (12). In a sequence, any set of frames for which the $d(t)$ score exceeds a threshold T_d is judged to be "poorly segmented".

3.3. 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 (7):

$$\mu_4 \delta_{color}(t; i) + \mu_5 \delta_{motion}(t; i) > T_\delta, \tag{16}$$

where T_δ is a threshold value. If the threshold 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 (16) is k-standard deviations above the mean of the measure over the object.

4. STATISTICAL VALIDATION OF PROPOSED MEASURES

In order to check the validity of the proposed non-ground-truth performance measures, we introduce a canonical correlation analysis of the proposed measures with measures using ground truth maps. To this effect, we first review performance measures using ground truth information. The canonical correlation analysis framework will be discussed next. Experimental results of this correlation analysis on two different sequences (Bream, Parrot) with five different segmentation/tracking methods (ETRI, open-loop, edges-only, equal-weight, adaptive-weight) will be provided in Section 5.

4.1. Measures with Ground-Truth

The measures using ground-truth segmentation maps[2] 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:

$$d^g_{pixel}(t) = \frac{\sum_{(x,y)} I(x,y;t)\mathrm{Cham}_g(x,y;t)}{\sum_{(x,y)} \mathrm{Cham}_g(x,y;t)}, \tag{17}$$

where $I(x, y; t)$ denotes an indicator function which takes the value 1 if reference and estimated segmentation masks of the object differ, $Cham_g()$ denotes the Chamfer distance transform of the boundary of ground-truth the object. For multiple objects one can simply average the $d_{pixel}(t)$ score over all segmented objects.

The **shape penalty** (d^g_{shape}) between the ground-truth and the estimated segmentation maps are calculated by looking at the difference between the turning angle functions (TAF)[2] of the segment boundaries:

$$d^g_{shape}(t) \quad = \quad \frac{\sum_{j=1}^{K} |\Theta^t_g(j) - \Theta^t_s(j)|}{2\pi K}, \tag{18}$$

where $\Theta^t_g(j)$ and $\Theta^t_g(j)$ denote the turning angle function of the ground-truth 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,[10] 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 degrees.

Finally, The **motion penalty** (d^g_{motion}) is calculated by computing the motion vectors on the ground-truth and the estimated segmentation maps:

$$d^g_{motion}(t) \quad = \quad \frac{\|\mathbf{v}_g(t) - \mathbf{v}_s(t)\|}{\|\mathbf{v}_g(t)\| + \|\mathbf{v}_s(t)\|}, \tag{19}$$

where $\mathbf{v}_g(t)$ is any parametric motion representation for the ground-truth 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 $D^g_{pixel}, D^g_{shape}, D^g_{motion}$.

4.2. Representation of Data

Let the **ground-truth** performance scores obtained for the t^{th} frame ($t = 1, \ldots, n$) video shot or a video sequence denoted by

$$\mathbf{x}_t = \left[\begin{array}{ccc} x_{t1} & \ldots & x_{tp} \end{array} \right] = \left[\begin{array}{ccc} d^g_{pixel}(t) & d^g_{shape}(t) & d^g_{motion}(t) \end{array} \right] \tag{20}$$

consisting of the pixel misclassification penalty, shape penalty and the motion penalty and hence $p = 3$. The superscript g denotes that these measures use ground-truth segmentation maps. Similarly let the **non-ground-truth** performance scores obtained for the i^{th} frame be denoted by

$$\mathbf{y}_t = \left[\begin{array}{ccc} y_{t1} & \ldots & y_{tq} \end{array} \right] = \left[\begin{array}{ccc} d_{hist}(t) & d_{color}(t) & d_{motion}(t) \end{array} \right], \tag{21}$$

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

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

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_n \end{bmatrix} = [\mathbf{X}|\mathbf{Y}] = \left[\begin{array}{c|c} \mathbf{x}_1 & \mathbf{y}_1 \\ \vdots & \vdots \\ \mathbf{x}_n & \mathbf{y}_n \end{array}\right] \tag{22}$$

where the performance measure vector in a row for $t = 1, \ldots, n$ reads as:

$$\mathbf{z}_t = \left[\begin{array}{ccc|cc} d^g_{pixel}(t) & d^g_{shape}(t) & d^g_{motion}(t) & d_{hist}(t) & d_{color}(t) \quad d_{motion}(t) \end{array}\right] \tag{23}$$

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 ground-truth and non-ground-truth partitions are indicated as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma_{XX}}(p \times p) & \mathbf{\Sigma_{XY}}(p \times q) \\ \mathbf{\Sigma_{YX}}(q \times p) & \mathbf{\Sigma_{YY}}(q \times q) \end{bmatrix}, \tag{24}$$

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

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

4.3. Canonical Correlation Analysis

The association between the two data sets, which in our case consist of ground-truth and non-ground-truth 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 $p-1$ eigen-solutions.[11] Let's 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 $(p \times 1)$ and $\mathbf{b}_1$ is $(q \times 1)$. 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.

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

$$\max_{\mathbf{a}_1, \mathbf{b}_1} \quad Corr(\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y}) = \rho_1. \tag{25}$$

It can be shown that the parameters $\rho_1^2 \geq \rho_2^2 \geq \ldots \rho_p^2$ are the joint eigenvalues of the matrices

$$\mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1/2}, \qquad or \qquad \mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}. \tag{26}$$

These matrices have, respectively the eigenvector set $\mathbf{e}_1, \ldots, \mathbf{e}_p$ $(p \times 1)$ and $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_q$. Finally, the linear combiner weights $\mathbf{a}_i$ and $\mathbf{b}_i$ in (25) result from $\mathbf{e}_i^T\mathbf{\Sigma}_{XX}^{-1/2}$ and $\mathbf{f}_i^T\mathbf{\Sigma}_{YY}^{-1/2}$. The readers are referred to[11] 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 researches argue that[12] the degree of association 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 in the following Section.

4.4. Canonical Loadings and Redundancy

We would like to show that the ground-truth (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*[12] 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), 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).

There is a straightforward relation between the canonical weights ($\mathbf{a}_i$ or $\mathbf{b}_i$) and the loadings.[12] Using the symbol $\mathbf{s}_{XX_1}$ to represent the vector of intraset loadings for the $\mathbf{X}$ set on its first canonical variate, we obtain:

$$\mathbf{s}_{XX_1} = \mathbf{\Sigma}_{XX}\mathbf{a}_1, \tag{27}$$

where $\mathbf{a}_1$ is the weight vector and $\mathbf{\Sigma}_{XX}$ 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:

$$\mathbf{s}_{YX_1} = \mathbf{\Sigma}_{YX}\mathbf{a_1} \tag{28}$$

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 j^{th} component of set $\mathbf{Y}$ is

$$Red_{\mathbf{X}Y_j} = \frac{1}{p}\mathbf{s}^T_{\mathbf{X}Y_j}\mathbf{s}_{\mathbf{X}Y_j} = \frac{1}{p}\sum_{i=1}^{p} s^2_{X_iY_j}, \tag{29}$$

where $\mathbf{s}_{\mathbf{X}Y_j}$ is the vector of interset loadings of the measures in the set $\mathbf{X}$ with the j^{th} 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 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.

More explicitly, a variable may receive a small weight simply because it is highly correlated with another variable in its set, even though both variables have high correlations with the canonical variate.[12] 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.

5. EXPERIMENTAL RESULTS

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

The second sequence is called the "Parrot" sequence. 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.

5.0.1. "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[13,14] and the last technique was developed at ETRI (Electronics and Telecommunications Research Institute) in Korea.[15,16]

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

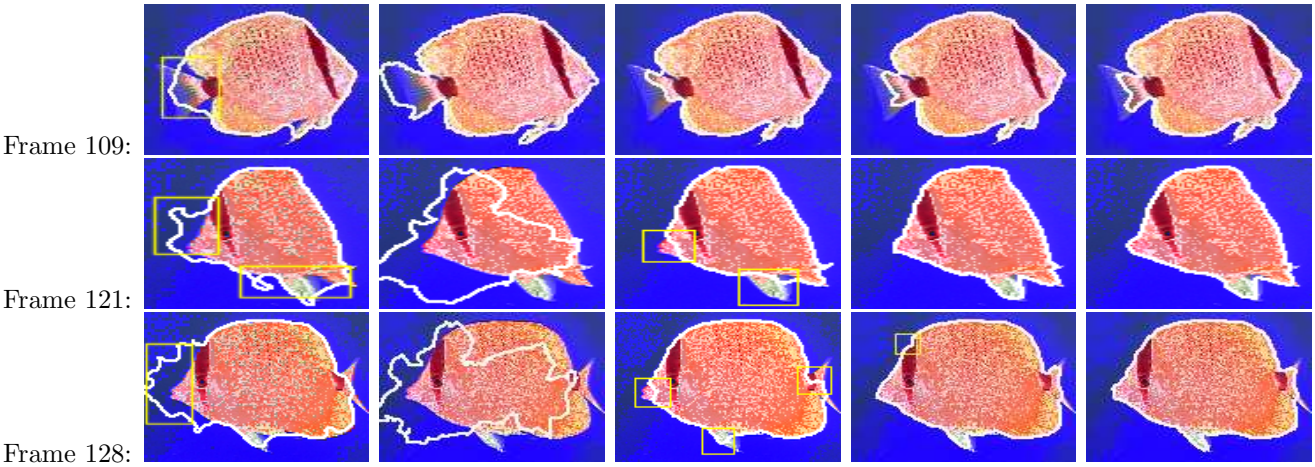


Figure 2. Tracking results for frames 109, 121, and 128 (with some zoom in). *Column 1:* ETRI results. *Column 2:* Open-loop method. *Column 3:* Closed-loop with edge energy only. *Column 4:* Closed-loop with equal weighting. *Column 5:* Closed-loop with adaptive weighting methods.

In Figure 4, we give the performance evaluation measures using the ground-truth and non-ground-truth measures, respectively. Table 1 summarizes the performance evaluation measures by providing the mean values over all frames. We can observe in Table 1 that, the ETRI and Open Loop methods have worst scores in all ground-truth and non-ground-truth 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 ground-truth measures deteriorate, there is a commensurate increase in the non-ground-truth 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 Figure 4(a) and (d)).

Table 1. The 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
	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combined	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combined
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

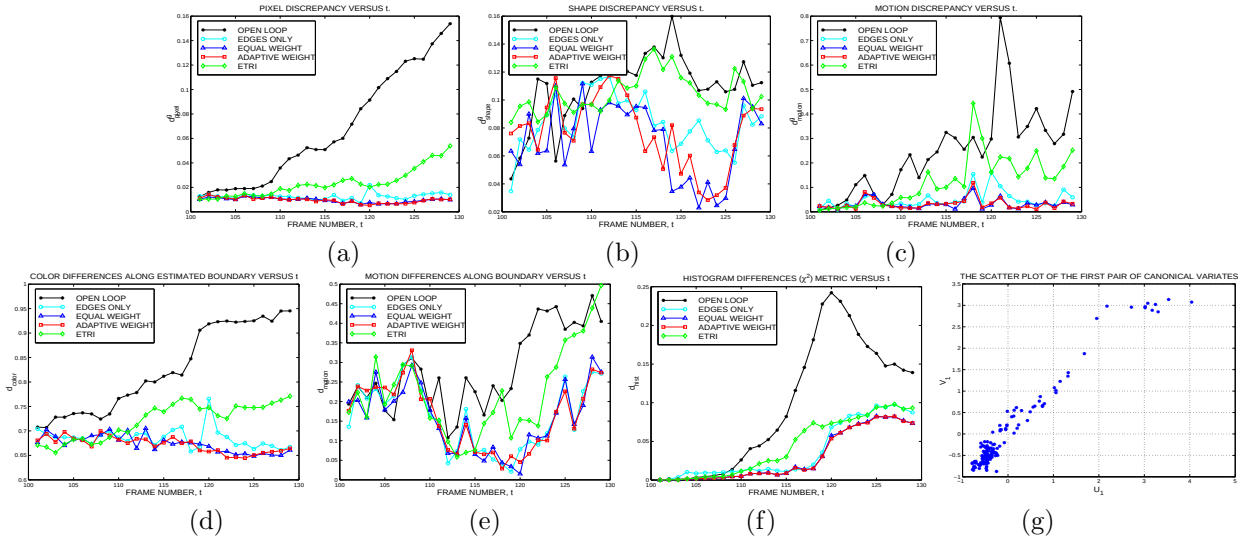


Figure 3. Measures with ground-truth for "Bream" sequence: (a) The misclassification penalty $DP(t)$ for each frame (b) The shape penalty $DS(t)$ (c) The motion penalty $DM(t)$ Measures without ground-truth: (d) Color differences along boundary (e) Motion differences along boundary (f) Inter-frame histogram differences using χ^2 measure. (g) The scatter plot of the first canonical variate pair.

In order to quantify the correlation, we have calculated the correlation matrix which was defined in (24):

$$\mathbf{\Sigma} = \left[\begin{array}{ccc|ccc} d^g_{pixel} & d^g_{shape} & d^g_{motion} & d_{hist} & d_{color} & d_{motion} \\ \hline 1.00 & 0.47 & 0.82 & 0.74 & 0.95 & 0.61 \\ 0.47 & 1.00 & 0.48 & 0.27 & 0.56 & 0.26 \\ 0.82 & 0.48 & 1.00 & 0.73 & 0.86 & 0.46 \\ \hline 0.74 & 0.27 & 0.73 & 1.00 & 0.70 & 0.45 \\ 0.95 & 0.56 & 0.86 & 0.70 & 1.00 & 0.55 \\ 0.62 & 0.26 & 0.46 & 0.45 & 0.55 & 1.00 \end{array}\right]. \tag{30}$$

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 ground-truth and non-ground-truth measures, we get the following pair of transformation coefficients:

$$\mathbf{a}_1 = \left[\begin{array}{ccc} 0.77 & 0.08 & 0.21 \end{array}\right]^T, \mathbf{b}_1 = \left[\begin{array}{ccc} 0.14 & 0.84 & 0.09 \end{array}\right]^T$$

The first canonical variate of the set of ground-truth 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 non-ground-truth measure set, the canonical variate is constructed with the color difference measure (0.84) and the histogram measure (0.14) with motion information discarded (0.09). The less conclusive evidence from shape discrepancy in the first set and the motion measure in the second set is also obvious in Figure 4(b) and (e), respectively. The reason that the shape and motion measures have small weights can be due to their correlations with the other measures in their sets.

The most important result of the canonical analysis is the fact that, the maximum correlation between the first canonical variate pair (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. The scatter plot of the first canonical variate pair is given in Figure 4(g). This high canonical correlation implies that non-ground-truth measures reflect the information contained in the ground-truth measures.

We also carried out the redundancy computation through canonical loadings as was discussed in Section 4.4. As a result, we observed that 66% of the variance in the set of ground-truth measures is accounted for by the set of non-ground-truth measures. On the other hand, 65% of the variance in the set of non-ground-truth measures was found to be accounted for by the set of ground-truth measures.

5.0.2. "Parrot" Results

Several tracking results for the Parrot sequence are shown in Figure 5. If we analyze the tracking results in Figure 5, we can say that ETRI results (column 1) are the worst and Open Loop results (column 2) are the best. The quantitative results of the "Parrot" sequence are summarized in Table 2 together with the combined measure. Using the quantitative evaluation in Table 2, we can also see that ETRI results get the highest (worst) scores and the Open Loop results get the lowest (best) scores.

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

$$\mathbf{a}_1 = \begin{bmatrix} -1.36 & 0.09 & 0.41 \end{bmatrix}^T, \mathbf{b}_1 = \begin{bmatrix} 0.84 & 0.38 & -0.02 \end{bmatrix}^T.$$

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

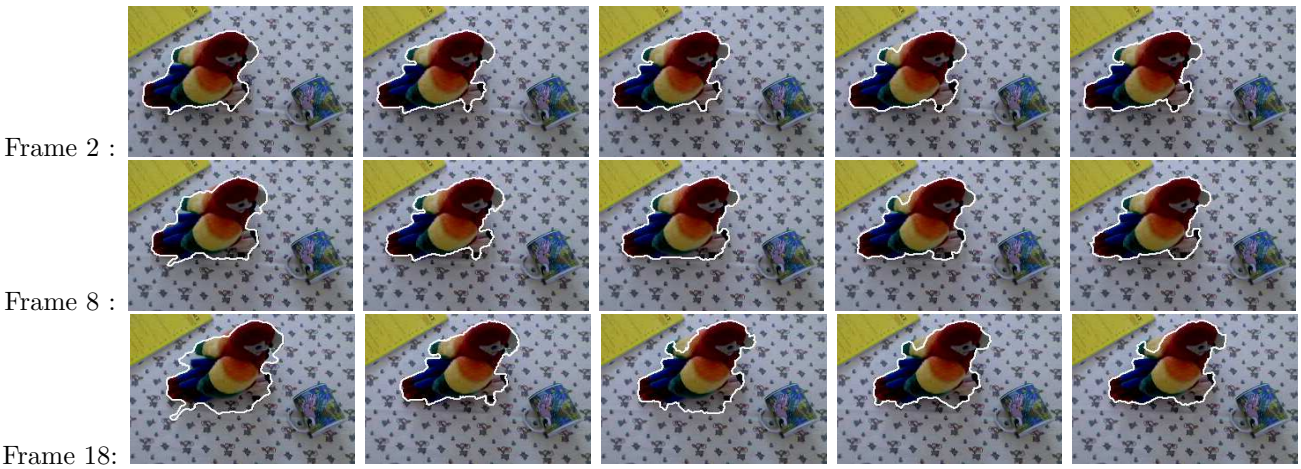


Figure 4. Tracking results for frames 2, 8, and 18. **Column 1**: ETRI results. **Column 2**: Open-loop method. **Column 3**: Closed-loop with edge energy only. **Column 4**: Closed-loop with equal weighting. **Column 5**: Closed-loop with adaptive weighting methods.

Table 2. The 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} (HSV)	D_{motion}	D
	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-1}$	Combined	$\times 10^{-2}$	$\times 10^{-1}$	$\times 10^{-1}$	Combined
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

6. CONCLUSIONS

We presented three non-ground-truth measures for quantitative performance evaluation of video object segmentation and tracking algorithms. The proposed measures yield a figure of merit for the whole segmented sequence or in turn can give more local results. Thus it is possible to identify the frames being badly segmented or even parts of object boundary within a frame. We tested the sensitivity of the measures to distortions in the segmentation map such as random shifts. We have also analyzed the correlation between the three proposed non-ground-truth measures and a set of ground-truth measures. We have found that they are reasonably correlated, implying that non-ground-truth measures can be reliably used for performance monitoring in lieu of ground-truth measures. Thus the extremely tedious and time-consuming task of ground-truth extraction can be avoided.

REFERENCES

1. A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. Second IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2000.
2. C. E. Erdem and B. Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X European Signal Processing Conference*, **2**, pp. 917–920, September 2000.
3. X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. X European Signal Processing Conference*, **4**, September 2000.
4. P. Correira and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2000.
5. J. Bednar and T. L. Watt, "Alpha-trimmed means and their relationship to median filters," *IEEE Transactions Acoustics, Speech, and Signal Processing* **32**(1), pp. 145–153, 1984.
6. A. M. Ferman, A. M. Tekalp, and R. Mehrotra, "Robust color histogram descriptors for video segment retrieval and identification," *IEEE Transactions on Image Processing* **11**, pp. 497–508, May 2002.
7. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flanney, *Numerical Recipes in C*, Cambridge Univeristy Press, 1992.
8. C. E. Erdem, A. M. Tekalp, and B. Sankur, "Metrics for performance evaluation of video object segmentation and tracking without ground-truth," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, **2**, pp. 69–72, 7-10 Oct., Greece, 2001.
9. Y. Fu, A. T. Erdem, and A. M. Tekalp, "Tracking visible boundary of objects using occlusion adaptive motion snake," *IEEE Transactions on Image Processing* **9**, pp. 2051–2060, December, 2000.
10. 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 Analysis and Machine Intelligence* **13**, pp. 209–215, 1991.
11. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 1998.
12. H. E. A. Tinsley and S. D. Brown, *Handbook of Applied Multivarite Analysis and Mathematical Modeling*, Academic Press, 2000.
13. C. E. Erdem, A. M. Tekalp, and B. Sankur, "Non-rigid object tracking with feedback of performance evaluation measures," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 9-14 Dec., Hawaii, USA, 2001.
14. C. E. Erdem, A. M. Tekalp, and B. Sankur, "Video object tracking with feedback of performance measures," *IEEE Transactions Circuits and Systems for Video Technology* **13**(4), 2003.
15. J. G. Choi, "A user assisted segmentation method for video object plane segmentation," in *The Int. Conf. on Circuits/Systems, Computers and Communications*, pp. 7–10, July 1998, Korea.
16. M. Kim, J. G. Choi, M. H. Lee, and C. Ahn, "User's guide for a user-assisted video object segmentation tool," in *ISO/IEC JTC1/SC29/WG11 MPEG98/m3935*, October 1998.