ON EVALUATING METRICS FOR VIDEO SEGMENTATION ALGORITHMS

Elisa Drelie Gelasca, Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland. {elisa.drelie,touradj.ebrahimi}@epfl.ch.

ABSTRACT

Evaluation is a central issue in the design, implementation, and performance assessment of all systems. Recently, a number of metrics have been proposed to assess the performance of segmentation algorithms for image and video data. This paper provides an overview of state of the art metrics proposed so-far, and introduces a new and efficient such metric. Doing so, subjective experiments are carried out to derive a perceptual metric. As a result, it also provides a comparison of performance of segmentation assessment metrics for different video object segmentation techniques.

1. INTRODUCTION AND SURVEY

The performance of algorithms for subsequent image or video processing often depends on a prior efficient segmentation.

Many researchers prefer to rely on qualitative human judgment for evaluation. In fact, Pal and Pal [1] say that a "human being is the best judge to evaluate the output of any segmentation algorithm". However, subjective evaluation asks for a large panel of human observers, thus resulting in a time-consuming and expensive process.

To avoid systematic subjective evaluation of segmentation, an automatic procedure is preferred. *Empirical methods* [2] are used to evaluate the segmentation algorithms indirectly, through their results. Empirical methods are divided into *empirical discrepancy*, metrics when the segmentation result is compared to an ideally segmented reference map (ground truth), and *empirical goodness* metrics, when the quality of the segmentation result is based on intuitive measures of goodness such as gray-level or color uniformity, shape regularity or contrast between regions.

Although goodness evaluation methods can be very useful for on-line evaluation, their results do not necessarily coincide with human perception of the goodness of segmentation. For this reason, when a reference mask is available or can be generated, discrepancy evaluation methods are preferred.

Despite several quality metrics proposed for still image segmentation [3, 4, 5, 6], they are not directly applicable to video object segmentation. In this section we will start by presenting the state of the art evaluation metrics for video object segmentation [7, 8, 9, 10, 11, 12, 13, 14]. In particular, we will provide the details of three methods [12, 13, 15] that will be used when assessing the performance of a new metric which will be described in the next section.

Empirical goodness methods have been defined not only for still image [3, 4] but also for video in [8, 9]. In [8], goodness metrics are developed and grouped into two classes: *intra-object homogeneity* (shape regularity, spatial uniformity, temporal stability and motion uniformity) and *inter-object disparity* (local color and motion contrast with neighbors). The goodness metrics are all combined in a composite metric with weights differentiated according to the type of content (stable or moving content). Erdem *et al.* [9] utilized a *spatial color contrast measure*, *color histograms differences* along the temporal axis and *motion vector differences* along the boundaries of the segmented objects, all combined in a single performance measure. Piroddi *et al.* [14] improved Erdem's goodness method in terms of sensitivity as well as immunity to noise.

To evaluate a video scene with segmented moving objects by means of discrepancy methods, Erdem and Sankur [10] combined three empirical discrepancy measures into an overall quality segmentation evaluation: *mis-classification penalty*, *shape penalty*, and *motion penalty*. In [8], Correia and Pereira first measured the individual segmentation quality through four spatial accuracy criteria: *shape fidelity*, *geometrical fidelity*, *edge and statistical content similarity* and two temporal criteria: *temporal perceptual information* and *criticality*. Then, they computed the similarity factor between the reference and the resulting segmentation. Furthermore, the multiple-object case was addressed by using the criteria of application-dependent "*object relevance*" [16] to provide the weights for the quality metric of each object.

Another way to approach the problem is to consider it as a particular case of shape similarity as proposed in [11] for video object segmentation. In this method, the evaluation of the spatial accuracy and the temporal coherence is based on the mean and standard deviation of the 2-D shape estimation errors.

We proposed to evaluate the quality of a segmented object through spatial and temporal accuracy joined to yield

a combined metric [7]. This work was based on two other discrepancy methods [12, 17] described below.

During the standardization work of ISO/MPEG-4, within the core experiments on automatic segmentation of moving objects, it became necessary to compare the results of different object segmentation algorithms, not only by subjective evaluation, but also by objective evaluation. The proposal for objective evaluation [12] agreed by the working group uses a ground truth. This metric is adopted by the research community due also to its simplicity. A refinement of this metric has been proposed by Villegas *et al.* [17, 13]. These metrics have been chosen as term of comparison for a new metric proposed in this paper.

1.1. MPEG Evaluation Criteria

A moving object can be represented by a binary mask, called *object mask*, where a pixel has object-label if it is inside the object and background-label if it is outside the object. The objective evaluation approach used in the MPEG-4 core-experiments has two objective criteria: the *spatial accuracy* and the *temporal coherence*. Spatial accuracy, Sqm, is estimated through the amount of error pixels (both false positive and false negative pixels) in the resulting mask deviating from an ideal mask.

Temporal coherence is estimated by the difference of the spatial accuracy of the mask, M, at the current and previous frame k,

$$Tqm_M(k) = Sqm(k) - Sqm(k-1). \quad (1)$$

The two evaluation criteria can be combined in a single **MPEG quality measure**, *MPEGqm(k)*, through the sum:

$$MPEGqm(k) = Sqm(k) + Tqm_M(k). \quad (2)$$

In this metric, the perceptual difference of different classes of errors, false positive and false negative, is not considered and they are all treated equally. In fact, different kinds of errors should be combined in the metric in correct proportions to match evaluation results produced by human observers.

1.2. Weighted Evaluation Criteria

Within the project COST 211 [18] the above approach has been further developed by Villegas and Marichal [17, 13]. For the evaluation of the spatial accuracy, as opposed to the previous method, two classes of pixels are distinguished: those which have object-label in the resulting object mask, but not in the reference mask (false positive) and vice versa (false negative), and they are weighted differently. Furthermore, their metric takes into account the impact of these two classes on the spatial accuracy, that is, the evaluation worsens with pixel distance d to the reference object contour. The spatial accuracy, qms, is normalized by the sum of the areas of reference objects as follows:

$$qms(k) = \frac{qms^+(k) + qms^-(k)}{\sum_{i=1}^{N_R} R_i(k)} = \frac{\sum_{d=1}^{D_M^+} w_+(d) \cdot |\mathcal{P}_d(k)| + \sum_{d=1}^{D_M^-} w_-(d) \cdot |\mathcal{N}_d(k)|}{\sum_{i=1}^{N_R} R_i(k)}, \quad (3)$$

where D_M^+ and D_M^- are the biggest distance d for, respectively, false positives and false negatives; N_R is the total number of objects in the reference R; $\sum_{i=1}^{N^R} R_i(k)$ is the sum of the area of all the objects i in the reference; $\mathcal{P}_d(k)$ and $\mathcal{N}_d(k)$ are positive and negative pixels respectively; $w_+(d)$ and $w_-(d)$ are the weights for positives and negatives respectively, expressed as:

$$w_+(d) = b_1 + \frac{b_2}{d + b_3}, \qquad w_-(d) = f_S \cdot d, \quad (4)$$

where the parameters b_i and f_S are chosen empirically [13]: $b_1 = 20$, $b_2 = -178.125$, $b_3 = 9.375$ and $f_S = 2$. These functions represent the fact that the weights for false negative pixels increase linearly and they are larger than those for false positives at the same distance from the border of the object. In fact, as we move away from the border, missing parts of objects become more important than added background.

Two criteria are used for estimating temporal coherence, the temporal stability $qmt(k)$ and the temporal drift $qmd(k)$ of the mask. First, the variation of spatial accuracy criterion between successive frames is investigated as follows. The temporal stability is equal to the normalized sum of the differences of the spatial accuracy for two consecutive frames for false positive and false negative pixels:

$$qmt(k) = \frac{qms^+(k, k-1) + qms^-(k, k-1)}{\sum_{i=1}^{N_R} R_i(k)}. \quad (5)$$

where $qms^*(k, k-1) = |qms^*(k) - qms^*(k-1)|$.

Second, the displacement of the gravity center, $\overrightarrow{G}$, of the resulting object and the reference object mask is computed for successive frames to estimate possible *drifts* of the object mask, $\overrightarrow{qmd}(k)$:

$$\overrightarrow{qmd}(k) = [\overrightarrow{G}_E(k) - \overrightarrow{G}_R(k)] - [\overrightarrow{G}_E(k-1) - \overrightarrow{G}_R(k-1)] \quad (6)$$

that is displacement from time $(k-1)$ to time (k) of the centers of gravity $\overrightarrow{G}$, of the estimated E and reference R masks. The value of drift is the norm of the displacement vector divided by the sum of the reference object bounding boxes,

$$qmd(k) = \frac{||\overrightarrow{qmd}(k)||}{\frac{1}{N_R} \sum_{i=1}^{N_R} BB_i^{x,y}(k)}, \quad (7)$$

where $BB_i^{x,y}(k)$ is the bounding box of the object i in the reference mask R at time k. The authors proposed to define a single quality value by linearly combining all the three presented measures as the **weighted quality metric**, $wqm(k)$:

$$wqm(k) = w_1 \cdot qms(k) + w_2 \cdot qmt(k) + w_3 \cdot qmd(k). \quad (8)$$

The values of the weights w_i are very much application dependent. If no application is specified, all three weights can be assumed equal to $\frac{1}{3}$.

In this method, the perceptual difference between two kinds of errors is taken into account. The drawback is that the weighting functions defined in Eq. (4), that should be 'perceptual' weights of the evaluation criteria, are defined by means of empirical tests. These empirical tests are not generally sufficient. As well in all other proposed evaluation criteria in the literature, the relevance and the corresponding weight of different kinds of errors should be supported by formal subjective experiments performed under clear and well defined specifications.

1.3. Object Matching Evaluation Criteria

Nascimento and Marques [15] used several simple discrepancy metrics to classify the errors into region splitting, merging or split-merge, detection failures and false alarms. In this scenario, the most important thing is that all the objects have to be detected and tracked along time. Object matching is performed by computing a binary correspondence matrix between the segmented and the ground truth images. The advantage of the method is that ambiguous segmentations are considered (e.g., it is not always possible to know if two close objects correspond to a single group or a pair of disjoint regions: both interpretations are adopted in such cases). In fact, by analyzing this correspondence matrix, the following measures are computed: Correct Detection (C_D): the detected region matches one and only one region; False Alarm (F_A): the detected region has no correspondence; Detection Failure (D_F): the test region has no correspondence; Merge Region (M): the detected region is associated to several test regions; Split Region (S): the test region is associated to several detected regions; Split-Merge Region (S_M): when the conditions M and S simultaneously occur.

The normalized measures are obtained by normalizing the amount of F_A by the number of objects in the segmentation, N_C, all the others by the number of objects in the reference, N_R, and by multiplying the obtained numbers by 100. The **object matching quality metric** at frame k, $mqm(k)$, is finally given by:

$$\begin{aligned} mqm(k) &= w_1 \cdot \frac{C_D(k)}{N_R} + w_2 \cdot \frac{F_A(k)}{N_C} + w_3 \cdot \frac{D_F(k)}{N_R} \\ &+ w_4 \cdot \frac{M(k)}{N_R} + w_5 \cdot \frac{S(k)}{N_R} + w_6 \cdot \frac{S_M(k)}{N_R} \quad (9) \end{aligned}$$

where w_i are the weights for the different discrepancy metrics. It is evident that this metric is able to describe quantitatively the correct number of detected objects and their correspondence with the ground truth only, while the metrics described in the previous sections are able to monitor intrinsic properties of the segmented objects such as shape irregularities and temporal instability of the mask along time.

2. PROPOSED PERCEPTUAL METRIC

The perceptual objective metric proposed here is defined based on two types of errors, namely, objective errors and perceptual errors. Objective metrics quantify the deviation (objective error) of the segmentation under test from the ground truth and are described in this section. Perceptual metrics weight these deviations (perceptual errors) according to human perception by means of subjective experiments described in [19].

The proposed objective metric, as discussed in the next section, will be compared to the MOS (Mean Opinion Score) to provide the final perceptual objective assessment. The novelty of our approach consists in classifying the different clusters of error pixels according to the following characteristics: If they do or do not modify the shape of the object and afterward their size. Border holes, $\mathcal{H}_b$, and added backgrounds, $\mathcal{A}_b$, modify the shape while inside holes, $\mathcal{H}_i$, and added regions, $\mathcal{A}_r$ preserve the segmented object shapes.

The relative spatial error $\mathbf{S}_{A_r}(k)$, for all the j added regions at frame k, $\mathcal{A}_r^j(k)$, is obtained by simply applying:

$$\mathbf{S}_{A_r}(k) = \frac{\sum_{j=1}^{N_{Ar}} |\boldsymbol{A}_r^j(k)|}{|n(k)|}, \quad (10)$$

where $|\cdot|$ is the set cardinality operator; $n(k)$ is the sum of the reference and the resulted segmentation areas; N_{Ar} is the total number of added regions.

Similarly, for all the j holes inside the segmented objects, $\mathcal{H}_i^j(k)$, the relative spatial error, $\mathbf{S}_{H_i}(k)$, is given by:

$$\mathbf{S}_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |\boldsymbol{H}_i^j(k)|}{|n(k)|}, \quad (11)$$

where N_{Hi} is the total number of holes inside the objects. The spatial error for added background and holes on the border of the object is formulated in a different way. In fact, both kinds of errors are located around the object contours and have to be distinguished from the numerous deviations around the object boundary and a few but larger

deviation [11] by adding this weighting factor:

$$1 + \frac{\overline{d} + \sigma_d}{d_{max}}, \tag{12}$$

where d are the distance values[1] of error pixels from the correct object contour. The mean $\overline{d}$ and the standard deviation σ_d are calculated and are normalized by the maximal diameter, d_{max}, of the reference object to which the cluster of errors belongs. By combining this last Eq. (12) and Eq. (10), we obtain for the border artifacts the corrected relative spatial error $\mathbf{S}_{A_b}(k)$, for j added backgrounds:

$$\mathbf{S}_{A_b}(k) = \big(\frac{1}{|n(k)|} + \frac{\sum_{j=1}^{N_{Ab}} (\overline{d}^j_{Ab} + \sigma^j_{dAb}) \cdot |\boldsymbol{A}^j_b(k)|}{d_{max} \cdot |n(k)|}\big), \tag{13}$$

similarly for j holes on the border, $\mathcal{H}^j_b(k)$, the relative spatial error $\mathbf{S}_{H_b}(k)$ is:

$$\mathbf{S}_{H_b}(k) = \big(\frac{1}{|n(k)|} + \frac{\sum_{j=1}^{N_{Hb}} (\overline{d}^j_{Hb} + \sigma^j_{dHb}) \cdot |\boldsymbol{H}^j_b(k)|}{d_{max} \cdot |n(k)|}\big). \tag{14}$$

The temporal artifact caused by an abrupt variation of the spatial errors between consecutive frames is called *flickering*. To take this phenomenon into account in the objective metric, a measure of flickering is introduced, $\mathbf{F}(k)$ that can be computed for each artifact Λ=[$\mathcal{A}_r$, $\mathcal{A}_b$, $\mathcal{H}_i$, $\mathcal{H}_b$] as follows:

$$\mathbf{F}_\Lambda(k) = \frac{|\Lambda(k)| - |\Lambda(k-1)|}{|\Lambda(k)| + |\Lambda(k-1)|}, \tag{15}$$

The difference of the amount of an artifact between two consecutive frames is normalized by the sum of the amount of this artifact in the current frame k and the previous frame $k-1$. To model this effect, Eq. (15) is combined to the relative spatial artifact measures to construct an objective *spatio-temporal* error measure $\mathbf{ST}(\mathbf{k})$ for each artifact, and finally the artifact is summed along the time axis to obtain the overall objective spatio temporal metric $\mathbf{ST}$ for each artifact Λ:

$$\begin{aligned} \mathbf{ST}_\Lambda(k) &= \mathbf{S}_\Lambda(k) \cdot \frac{1 + \mathbf{F}_\Lambda(k)}{2}, \\ \mathbf{ST}_\Lambda &= \frac{1}{K} \sum_{k=1}^{K} w_t(k) \mathbf{ST}_\Lambda(k), \end{aligned} \tag{16}$$

where the temporal weights $w_t(k)$ that model the *human memory effect* have been empirically defined [19] as:

$$w_t(k) = (a \cdot e^{\frac{k-30}{b}} + c) \tag{17}$$

[1]For distance computation, 8-connectivity has been used.

with $a = 0.02$, $b = 7.8$, $c = 0.0078$, $K = 60$ (total number of frames).

Synthetic artifacts have been used to study and to characterize the perception of spatial and temporal artifacts previously described. The experimental protocol to carry out the subjective experiments is described in [19] and the viewing conditions are defined by the ITU Recommendations [20, 21].

In the following, a brief description of the parameters obtained for the perceptual metric is given and in the next section, the proposed metric is tested and compared to the state to the art metrics. The $\mathbf{ST}$ values of each artifact metric were plotted versus the values of MOS and the best fitting psychometric curves have been found to describe the human perception of errors [19]. Four psychometric curves have been derived through subjective experiments, one for each artifact, to obtain four *perceptual artifact metrics*: **PST**. The best fitting function for each artifact is the Weibull function, W. Thus the perceptual artifact metrics are described by:

$$\begin{aligned} W(x,S,k) &= 1 - e^{-(Sx)^k} \quad where \quad x = \mathbf{ST}_\Lambda \\ \mathbf{PST}_\Lambda &= W(\mathbf{ST}_\Lambda, S, k) \end{aligned} \tag{18}$$

where the parameters S and k have been obtained in [19] for general application scenarios: $S = 0.014$, $k = 0.304$ for $\mathbf{PST}_{A_r}$; $S = 0.026$, $k = 0.653$ for $\mathbf{PST}_{A_b}$; $S = 0.331$, $k = 0.2339$ for $\mathbf{PST}_{H_i}$; $S = 0.771$, $k = 0.641$ for $\mathbf{PST}_{H_b}$.

The overall perceptual metric is given by the combination of all the four kinds of artifacts. The total annoyance can be so estimated by a simple linear combination of artifacts [19]:

$$\mathbf{PST} = a_1 \cdot \mathbf{PST}_{A_r} + a_2 \cdot \mathbf{PST}_{A_b} + a_3 \cdot \mathbf{PST}_{H_i} + a_4 \cdot \mathbf{PST}_{H_b} \tag{19}$$

The perceptual weights were found by means of subjective experiments on combined errors: $a_1 = 2.86$, $a_2 = 4.50$, $a_3 = 4.77$, $a_4 = 5.82$. This equation can be further extended and the predicted annoyance by the proposed metric **PST** can be formulated by the following expression (Minkowski metric)[22]:

$$\mathbf{PST} = \sum a_i \cdot (\mathbf{PST}^p_\Lambda)^{\frac{1}{p}} \tag{20}$$

To find the Minkowski coefficients and exponent, we performed a nonlinear least-squares data fitting using the data obtained from the subjective experiment of combined artifacts. In this case, the Minkowski coefficients are: $p = 1.6$, $a_1 = 11.36$, $a_2 = 19.54$, $a_3 = 26.58$, and $a_4 = 32.52$. The correlation coefficients and the Minkowski exponent are reported in the 1st and 2nd rows of Tab. 1. The squared correlation coefficient (r) is used, also called *goodness of fit* [19].

The linear model is simpler and the correlation slightly decreases between the linear ($r = 0.86$) and the more generic

Table 1. Performance of the proposed metric **PST** and state of the art metrics, *MPEGqm*, *wqm* and *mqm*: correlation with Mean Opinion Score, MOS.

Metric	*Correlation*
PST (p=1.6)	0.90
PST (p=1)	0.86
MPEGqm	0.71
wqm	0.56
mqm	0.21

Minkowski model ($r = 0.90$). Since there is not a significant difference (by means of F-test) between the two models, the simpler linear model has been chosen for the proposed metric ($p = 1.6$, $r = 0.86$).

3. PERFORMANCE COMPARISON

In the previous section, we presented several artifact metrics that measured the annoyance of four of the most common spatial and two temporal artifacts found in video object segmentation.

Given the results of the subjective experiments carried out on synthetic artifacts [19], we propose a ground truth based perceptual objective metric that uses the metrics for spatial artifacts: $\mathbf{PST}_{A_r}$, $\mathbf{PST}_{A_b}$, $\mathbf{PST}_{H_i}$, and $\mathbf{PST}_{H_b}$ of Eqs.(18)-(20); for temporal artifacts: the **flickering** metric in Eq. (15) and the **expectation** effect of Eq. (17).

In order to compare the results of the proposed method to the state of the art metrics, we ran the three metrics described on the synthetically generated test sequences of the combined artifact experiment. The state of the art metrics described in Secs. 1.1 -1.3 are: the MPEG metric, *MPEGqm*, Villegas' metric, *wqm*, and Nascimento's metric, *mqm* .

In the 3rd, 4th and 5th rows of Tab. 1 are presented the correlation for respectively, *MPEGqm*, *wqm* and *mqm* metrics. The correlations coefficients are respectively: $r = 0.71$, $r = 0.56$ and $r = 0.21$. Our proposed method with a correlation of $r = 0.86$ outperforms the other state of the art metrics.

MPEG metric is the second best metric in fitting the subjective data. This result is surprising since no distinction between different kinds of error is applied in the MPEG metric in contrast with Villegas' and our metric. It has to be mentioned that all the weights for the Villegas' metric in Eq. (8) are set the same value and it is possible that by tuning a better fit could have been obtained. However, if no specific application is specified, as in these subjective experiments, using equal weights seems to be a good compromise.

As predicted in Sec. 1.3, the Nascimento's metric does not provide a good fit with the subjective data. In fact this metric is more suitable to predict object tracking quality than object segmentation quality. In our subjective experiments, subjects were told to judge in general the quality of segmentation without specifically taking into special account the quality of object tracking.

4. CONCLUSIONS

First, video object segmentation evaluation is reviewed in this paper. To this end, three state of the art metrics whose performance are analyzed are described in details. The first state of the art metric, *MPEGqm* is a simple sum of spatial and temporal errors commonly used by the research community. The second metric, *wqm* is a refinement of the first one where false positive and false negative errors are distinguished and weighted differently in the final formula. The third state of the art metric, *mqm* combines several simple metrics to classify the errors into split and merge errors, detection failures and false alarms. None of the state of the art objective methods includes the characterization of artifact perception in their models.

Second, this paper proposes a new objective metric which includes the study and characterization of segmentation artifact perception obtained by means of subjective experiments. Four spatial artifacts are deeply analyzed, namely, added regions, added background, inside holes and border holes. Two temporal effects are also studied, namely, the temporal flickering and the expectation effect. Objective measures are proposed to estimate these artifacts. Through subjective experiments the objective measures are modeled by psychometric curves found to assess the annoyance of the artifact perception. In the combined artifact subjective experiment, the relationship between the individual spatial artifact weights and the overall annoyance is found. It was found that added region weight is smaller than that of the inside hole and added background, and almost half of the border hole artifact weight (the most annoying artifact).

Finally, an overall perceptual objective metric was proposed on the basis of the results described above. The performance of the new metric was analyzed in terms of correlation with subjective scores and compared to those of the three considered state of the art metrics. It could be shown that the proposed perceptual objective metric provides superior performance to those of the state of the art *MPEGqm*, *wqm* and *mqm*.

5. ACKNOWLEDGMENT

The authors would like to thank John Foley and Mylene Farias for their valuable inputs, suggestions and feedback, particularly in the set up of subjective experiments and Marco Carli and Gaia Arrigoni for their contribution in running the tests.

6. REFERENCES

[1] Nikhil R. Pal and Sankar P. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, pp. 1277–1293, 1993.

[2] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.

[3] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern Recognition Letters*, vol. 19, pp. 741–747, 1998.

[4] C. Rosenberger and K. Chehdi, "Genetic fusion: application to multi-components image segmentation," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. IEEE International Conference on*, 5-9 June 2000, vol. 6, pp. 2223–2226.

[5] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Still image objective segmentation evaluation using ground truth," in *5th COST 276 Workshop 2003*, J. Prikryl B. Kovar and M. Vlcek, Eds., 2003, pp. 9–14.

[6] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the 8th International Conference On Computer Vision (ICCV-01), July 7-14, 2001, Vancouver, Canada*, IEEE, Ed., 2001, vol. 2, pp. 416–425.

[7] A. Cavallaro, E. Drelie, and T. Ebrahimi, "Objective evaluation of segmentation quality using spatio-temporal context," in *Proc. IEEE International Conference on Image Processing, Rochester(NY),22-25 September 2002*, 2002, pp. 301–304.

[8] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Transaction on Image Processing*, vol. 12, pp. 186–200, 2003.

[9] C. Erdem, B. Sankur, and A. M. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Transactions on Image Processing*, vol. 13, no. 7, pp. 937–951, July 2004.

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

[11] R. Mech and F. Marques, "Objective evaluation criteria for 2d-shape estimation results of moving objects," in *Workshop on Image Analysis for Multimedia Interactive Services*, Tampere, Finland, 16-17 May 2001.

[12] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," in *ISO/IECJTCI/SC29/WG11 M3448*, 43rd MPEG Meeting, Tokyo, Japan 1998, 1998.

[13] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1092–1103, August 2004.

[14] R. Piroddi and T. Vlachos, "A framework for single-stimulus quality assessment of segmented video," *EURASIP Journal on Applied Signal Processing*, to appear 1st quarter 2006.

[15] Jacinto Nascimento and Jorge S. Marques, "New performance evaluation metrics for object detection algorithms," in *6th International Workshop on Performance Evaluation for tracking and Surveillance (PETS, ECCV), Prague, May 2004*, 2004.

[16] P. Correia and F. Pereira, "Estimation of video object's relevance," in *Proc European Signal Processing Conf. - EUSIPCO*, IEEE, Ed., 2000, pp. 925–928.

[17] X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. Of X European Signal Processing Conference, Tampere, Finland*, 2000, pp. 2139–2196.

[18] Call for AM Comparisons, "Compare your segmentation algorithm to the cost 211 quat analysis model," 1996, http://www.iva.cs.tut.fi/COST211/Call/Call.htm.

[19] Elisa Drelie Gelasca, *Full Reference Obejctive Quality Metrics for Video Watermarking, Video Segmentation and 3D Model Watermarking*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, EPFL, Lausanne, Switzerland, Dec. 2005.

[20] ITU, *Methodology for Subjective Assessment of the Quality of Television Pictures Recommendation BT.500-11*, International Telecommunication Union, Geneva, Switzerland, 2002.

[21] ITU, *Subjective Video Quality Assessment Methods for Multimedia Applications Recommendation P.910*, International Telecommunication Union, Geneva, Switzerland, 1996.

[22] H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *Human Vision, Visual Processing, and Digital Display III; Bernice E. Rogowitz; Ed.*, 1992, vol. 1666, pp. 16–26.