Towards Perceptually Driven Segmentation Evaluation Metrics

Elisa Drelie Gelasca, Touradj Ebrahimi

Signal Processing Institute
Swiss Federal Institute of Technology EPFL
CH 1015 Lausanne, Switzerland
{elisa.drelie,touradj.ebrahimi}@epfl.ch

Mylène C. Q. Farias, Marco Carli
Sanjit K. Mitra

Department of Electrical Engineering
University of California Santa Barbara
Santa Barbara, CA 93106, USA
{mylene,marco,mitra}@ece.ucsb.edu

Abstract

To be reliable, an automatic segmentation evaluation metric has to be validated by subjective tests. In this paper, a formal protocol for subjective tests for segmentation quality assessment is presented. The most common artifacts produced by segmentation algorithms are identified and an extensive analysis of their effects on the perceived quality is performed. A psychophysical experiment was performed to assess the quality of video with segmentation errors. The results show how an objective segmentation evaluation metric can be defined as a function of various error types.

1. Introduction

The unsupervised segmentation of digital images is a difficult and challenging task [1] with several key-applications in many fields: remote sensing, medical diagnosis, vision-driven robotics, interactive entertainment, movie production and so on. The performance of algorithms for subsequent image or video processing, compression, indexing, often depends on a prior efficient image segmentation. Basically, by segmenting an image, several "homogeneous" partitions are created. The number of homogeneity criteria depends on the particular application and on the *a priori* knowledge of the problem. As an example, in a surveillance application every moving object is considered as an object of interest and, therefore, this information is used in the segmentation process.

In literature, many and different video object segmentation algorithms have been proposed. However, no single segmentation technique is universally useful for all applications and different techniques are not equally suited for a particular task. To properly evaluate the performance of segmentation techniques, automatic methods have been proposed [2], [3], [4]. The goal of an automatic segmentation evaluation method is to avoid subjective tests that constitute a time-consuming and expensive process. These objective methods evaluate segmentation algorithms through the quality of their results. A segmentation result can be judged according to general criteria of good segmentation or by comparison with a *reference segmentation* result representing the ideal segmentation [5].

To validate an objective evaluation, subjective experiments need to be performed. For this purpose, in this paper an analysis of artifacts produced by segmentation algorithms has been performed. The most common artifacts have been taken into account and subjective tests have been carried out. A protocol for subjective evaluation of segmented video sequences has been proposed. This protocol is an effort to make subjective evaluations in this field more reliable, comparable and standardized. Little has been done towards defining a procedure to evaluate the performance of objective metrics for segmentation [6]. The task of defining a formal protocol for subjective tests for video object segmentation quality assessment is very useful, since to the best of our knowledge, only informal tests have been performed [3], [4]. In evaluating edge detection algorithms [7] and still image segmentation [8] some experimental methods for subjective tests have been published.

The paper is organized as follows. The analysis of segmentation errors is discussed in Section 2. Section 3 describes the generated test video sequences for the subjective experiments. The experimental method is presented in Section 4. Subjective versus objective data are analyzed in Section 5. Finally, in Section 6 we draw the conclusions and discuss future directions.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

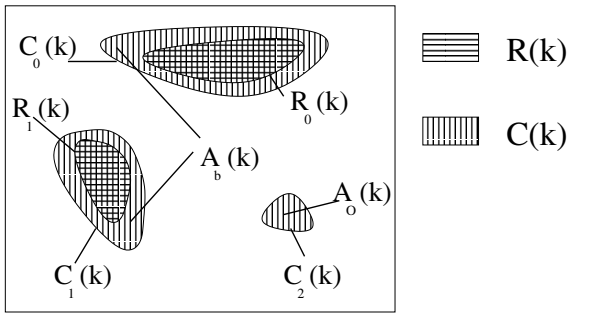

Figure 1. Reference segmentation overlapped to the resulted segmentation, at frame k. In this example, the two kinds of subset of $\mathcal{P}(k)$ are indicated.

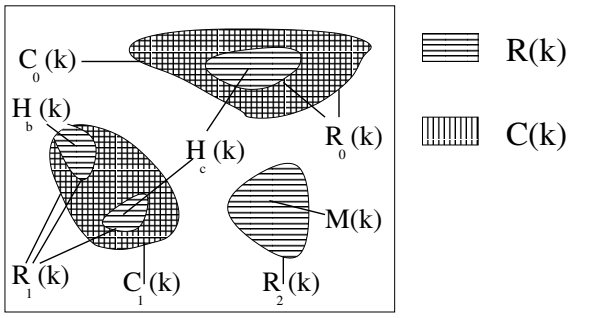

Figure 2. Reference segmentation overlapped to the resulted segmentation, at frame k. In this example, the different kinds of subset of $\mathcal{N}(k)$ are indicated.

2. Segmentation Errors

It is well known that segmentation errors can affect the quality of a segmented video in two ways: statically (*spatially*) and dynamically (*temporally*) [9]. The *spatial errors* of the segmented video are defined by the amount of mis-segmented pixels estimated by a direct comparison between reference and resulted segmentation mask, for a given frame k. An algorithm for object segmentation can in principle be evaluated by estimating only these pixel errors. Nevertheless, since a video is a sequence of images in which spatial errors take place, the temporal effect of segmentation errors must be considered. A given error may be perceived differently, depending on its temporal context. Observers are sensitive to *temporal errors*, i.e., changes in error characteristics along the time.

Pixel errors can be divided into two sets [10]: undetected pixels (*false negative*) and incorrectly detected pixels (*false positive*). Let us define a *region* i, $\mathcal{O}_i(k)$, at frame k as a set of pixels with the following properties: 1) $\mathcal{O}_i(k)$ is connected; 2) $\mathcal{O}_i(k) \cup \mathcal{O}_j(k)$ is disconnected $\forall\ i \neq j$.

We also indicate $R(k)$ as the set of all the j objects (meaningful regions) belonging to the reference segmentation, that can be expressed as:

$$R(k) = \bigcup_{j\in[0,J)} R_j(k) \quad \text{and} \quad \bigcap_{j\in[0,J)} R_j(k) = \emptyset \quad (1)$$

where J is the number of reference segmentation objects. J can also take the value zero when no object is present in the reference segmentation. Similarly, the set of pixels segmented at frame k, $C(k)$ is the union of the i regions/objects $C_i(k)$:

$$C(k) = \bigcup_{i\in[0,I)} C_i(k) \quad \text{and} \quad \bigcap_{i\in[0,I)} C_i(k) = \emptyset \quad (2)$$

where I is the number of resulted segmentation regions/objects. In the case I is zero, no object has been segmented in the resulted segmentation.

The set of *false positive* pixels, $\mathcal{P}(k)$, whose elements are the segmented pixels not belonging to the reference segmentation can be expressed as:

$$\mathcal{P}(k) = C(k) \cap R'(k) \quad (3)$$

where $R'(k)$ denotes the complement of $R(k)$. Similarly, *false negatives* $\mathcal{N}(k)$ appearing in the reference segmentation $R(k)$ and not in the resulted segmentation $C(k)$, can be expressed as:

$$\mathcal{N}(k) = C'(k) \cap R(k) \quad (4)$$

A further investigation of the segmentation errors has been carried out. In the following equations, let us define the condition empty intersection $\gamma_{i,j}(k)$ between the $j-th$ object in the reference segmentation and the $i-th$ in the resulted segmentation as:

$$\gamma_{i,j}(k) = \big(C_i(k) \cap R_j(k) = \emptyset\big) \quad (5)$$

The different errors will be mathematically expressed in Eqs. (6)-(12) and depicted in Figures 1-2. $\mathcal{P}(k)$ can be divided into two different kinds of subsets: *added background* and *added regions*. The added region set, $\mathcal{A}_{\mathcal{O}}(k)$ is a set of regions in $C(k)$ not present in $R(k)$:

$$\mathcal{A}_{\mathcal{O}}(k) = \bigcup_{i\in Q} C_i(k) \quad (6)$$

$$\text{where} \quad Q = \{i \mid \gamma_{i,j}(k),\ \forall\ j \in [0, J)\ \}$$

For the sake of simplicity, let us indicate with $\boldsymbol{A}_o(k)$ the number of regions contained in $\mathcal{A}_{\mathcal{O}}(k)$. $\boldsymbol{A}_o(k)$ therefore represents the number of added regions at frame k.

Added background $\mathcal{A}_b(k)$ does not constitute a region itself in $C(k)$ but is a set of false positive pixels erroneously segmented along the boundary of an object which is an object both in $C(k)$ and $R(k)$. $\mathcal{A}_b(k)$ therefore is composed of those pixels that do not satisfy condition in Eq.(5) and are subsets of $\mathcal{P}(k)$:

$$\mathcal{A}_b(k) = \mathcal{P}(k) \setminus \mathcal{A}_{\mathcal{O}}(k) \tag{7}$$

where $\setminus$ denotes a set difference. Let us express with $|\boldsymbol{A}_b(k)|$ the cardinality of $\mathcal{A}_b(k)$ that is the total amount of added pixels. For the entire video we could calculate the average $|\bar{\boldsymbol{A}}_b|$ over all the frames.

Different kinds of sets depending on the properties of their elements, can also be distinguished inside $\mathcal{N}(k)$. *Missing objects* $\mathcal{M}(k)$ are objects in $R(k)$ not present in $C(k)$:

$$\mathcal{M}(k) = \bigcup_{j \in S} R_j(k) \tag{8}$$
$$\text{where} \quad S = \{ j \mid \gamma_{i,j}(k),\ \forall\ \ i \in [0, I) \ \}$$

Holes $\mathcal{H}(k)$ are sets of $\mathcal{N}(k)$ that intersect the reference segmentation and do not satisfy condition in Eq.(5):

$$\mathcal{H}(k) = \mathcal{N}(k) \setminus \mathcal{M}(k) \tag{9}$$

In $\mathcal{H}(k)$ we can differentiate between *closed holes* $\mathcal{H}_c(k)$ and *boundary holes* $\mathcal{H}_b(k)$. Closed holes are sets of those false negative pixels completely inside the objects and satisfy the following condition:

$$\mathcal{H}_c(k) \subset cl\big(R(k)\big) \tag{10}$$

where $cl(\cdot)$ is the set infinitesimal closure operator. In the following sections, with $\boldsymbol{H}_c(k)$ we mean the number of hole sets contained in $\mathcal{H}_c(k)$.

Boundary holes are sets of false negative pixels that intersect the boundary of the reference object and modify the shape:

$$\mathcal{H}_b(k) \ \cap\ \partial R_j(k) \neq \emptyset \tag{11}$$

where ∂ is the boundary set operator.

In the following let us indicate by $d_H(k)$ the *Hausdorff distance* [11] between the set $\mathcal{H}_b(k)$ and the reference object j which intersects it:

$$d_H(k) = \max_{p \in \mathcal{H}_b(k)} \min_{q \in \partial R_j(k)} \| p - q \| \quad j \in T \tag{12}$$
$$\text{where} \quad T = \{ j \mid \text{not } \gamma_{i,j}(k),\ \forall\ \ i \in [0, I) \ \}$$

where $\| \ \|$ is the Euclidean norm, p is an element of $\mathcal{H}_b(k)$ and q of the boundary of $R_j(k)$.

Table 1. Tested segmentation error and values.

Tested Error	Values
Added Back. (10^3), $\|\boldsymbol{A}_b\|$	0.6,1.6,2.2,2.7,4.4
Added Region number, $\boldsymbol{A}_o$	3,4,7,12
Closed Hole number, $\boldsymbol{H}_c$	2,3,6,9
Boundary Hole dist., $d_H(k)$	5,10,15,20
Flickering Period, $\boldsymbol{f}_T$	1, 3, 5, 12, 30

3. Generation of Synthetic Segmentation Errors and Test Sequences

To generate the test video sequences, we modified the ideally segmented reference mask of a 300 frame MPEG-4 test sequence: *Hall monitor*. Different kinds and amounts of artifacts were added to the reference, as described in the next sections and the results were analyzed in Sec. 5. Since evaluation of the same sequence with different artifacts could cause *fatigue* in subjects, two other segmented video sequences were used in the test. The European IST project *Art.live*[1] sequence *Group*, as well as the MPEG-7 test sequence *Highway* were segmented by an automatic method of segmentation with different parameter sets [12] (as no reference segmentations for these two video sequences were available).

3.1. Synthetic Spatial Errors

A combination of the errors described in Sec. 2 is typically introduced by an automatic method of segmentation. When the segmentation quality is objectively evaluated in comparison with a reference segmentation, some features related to the artifact are derived (such as the number of added regions, distance of boundary holes from the ideal contour and so on). Many objective segmentation quality metrics have been proposed, but no work has been done on studying and characterizing these errors from a perceptual significance point of view.

The segmented images or video sequences can be thought to be made of a combination of the reference segmentation and some errors. In this work, the idea consists in producing segmentation errors that are relatively pure and studying their perceptual contribution.

Four different kinds of *spatial errors* have been synthesized and combined to the reference segmentation: added background, added regions, closed holes and boundary holes. The added background test sequence was synthetically generated by adding

[1] http://www.tele.ucl.ac.be/PROJECTS/art.live/

Table 2. Viewing conditions during subjective test.

Variable	Values
Peak luminance	<= 0.04
Maximum observation angle	10 degrees
Monitor resolution	1024×768
Viewing Distance	$35 - 40$ cm
Monitor Size	19"

increasingly more background to the R_j objects. By dilating the reference mask, five levels of dilation were generated (1,3,4,5,8). Therefore, five values of added background $|\boldsymbol{A}_b|$ were investigated in the experiment. Other test sequences were generated by adding four different amounts of added regions. The inserted added regions were constant in shape and size but varied in number $\boldsymbol{A}_o$ (3,4,7,12) in order to study the perception of an increasing number of added regions. Similarly, other four test sequences were generated by subtracting closed holes from the reference. The closed holes presented the same size and shape but varied in the number $\boldsymbol{H}_c$ (2,3,6,9). The last kind of spatial error investigated was the distance of $\mathcal{H}_b$ from the contour of the reference. Four test sequences with different distances $d_H(k)$ were generated for the boundary hole artifact. The distance was kept constant for each frame k ($\boldsymbol{d}_H$) along the same video sequence. The four values of $\boldsymbol{d}_H$ (5,10,15,20) for the four generated video sequences investigated the perception of object shape modification.

3.2. Synthetic Temporal Errors

In video segmentation, an error may vary its characteristics through time. A non smooth change of any spatial error deteriorates the perception of the error itself. The temporal artifact caused by a variation of the spatial error is called *flickering*. By carrying out subjective tests on real segmentation, flickering has been observed to be one of the most annoying artifacts introduced by segmentation algorithms. In fact, if an imprecise segmentation mask is stable along the time, it is perceived less annoying than a more precise segmentation presenting abrupt changes along the time.

Different variations of any spatial error could be implemented to test the flickering perception. We chose to change the position of added regions along the test sequence. The test video sequences with the temporal errors presented the same number $\boldsymbol{A}_o$ by the same shape and size. But their position changed each 1, 3, 5, 12 and 30 frames (flickering period, $\boldsymbol{f}_T$) by starting from a very fast and annoying flickering, and by ending with a temporally smooth change of added region position.

The spatial and temporal errors and their values are summarized in Table 3.

4. Experimental Method

A set of standards and grading techniques to evaluate quality of video and multimedia content have been defined by ITU-R [13] and ITU-T [14]. However, there are no prescribed standards for the evaluation of segmented video sequences. In this paper, we also propose a protocol for subjective evaluation of segmented video sequences based on ITU recommendations [13] and [14]. This protocol is an effort to make subjective evaluations in this field more reliable, comparable and standardized.

The usual approach to subjective quality assessment is to ask for a quality rating [14]. We used a Single Stimulus Continuous Quality Scale Method (SSCQS) [14]. In this method, only the video sequence under test is shown and subjects are asked to vote on a continuous scale after the video is shown. The display configuration showed the portion of the original image corresponding to the area of the segmented objects under analysis over a uniform green background. The scale used was a continuous quality scale between 0 (bad) and 10 (excellent). The continuous scale gave the user the ability to differentiate more easily between the qualities of segmentation. The reference segmentation or the original video were not used in the main experiment trials, but only in the training part for two reasons. First, in real applications the reference is not always available. Second, in the pilot tests, we noticed that subjects do not pay attention to the reference after the training stage of the test.

The viewing conditions of the experiments are given in Table 2. These conditions comply as much as possible with [13] and [14]. Each test session was composed of four stages: instructions, practice trials, experimental trials, and interview.

In the first stage, the subject was made familiar with the task of segmentation and then shown the reference segmentation. The second stage, practice trials, was used to familiarize the subject with the experiment and to allow the subjects' responses to stabilize before the main experiment began.

The experimental trials were performed with the complete set of test sequences presented in a random order. The 8 subjects were asked to rate the quality

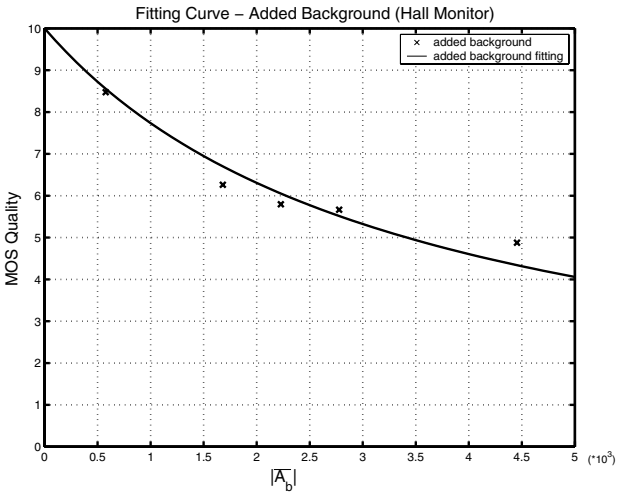

Figure 3. Mean quality curve corresponding to the added background error $|\bar{A}_b|$.

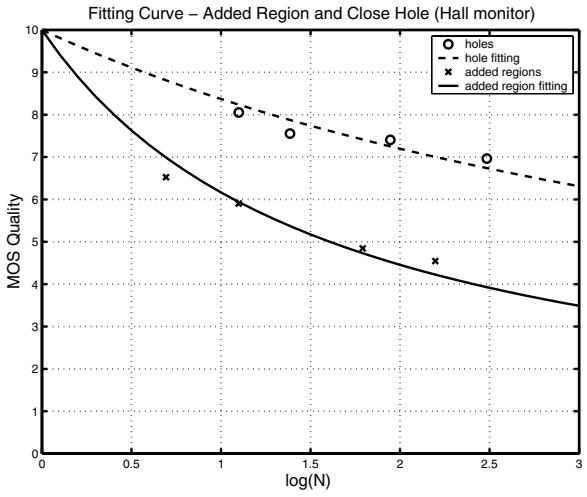

Figure 4. Mean quality curve corresponding to logarithm of the added region number A_o and closed hole number H_c.

of the segmented video under test. 56 test sequences were rated (2 repetitions × (22 artifacts×1 video sequences + 3 parameter set×2 video sequences)). Finally, in the interview stage, we asked the test subjects for qualitative descriptions of the artifacts that were perceived.

5. Data Analysis

In many quality assessment problems, the Mean Opinion Score (MOS) provides a numerical measure of the subjective quality. To determine MOS, a number of subjects rate the quality of system under test. The MOS is the arithmetic mean over all individual scores that can range from bad to excellent.

Standard methods [13] have been used to analyze the data provided by the test subjects and to screen the observers. In our case, the MOS values for *Hall monitor* test sequence have been fitted with a non-symmetrical function approximating the standard logistic function [13]:

$$y = y_{min} + \frac{(y_{max} - y_{min})}{1 + (\frac{x}{x_{mean}})\beta} \qquad (13)$$

where y is the predicted quality and x is the error measure. The parameters y_{max} and y_{min} establish the limits of the quality value range. The parameter x_{mean} translates the curve in the x-direction and the parameter β controls the steepness of the curve.

Figures 3-6 depict the perceived quality (MOS) for the artifacts caused by non ideal segmentations. Each figure contains both the experimental data and the fitted curve for *Hall monitor*.

Figure 3 shows the MOS versus the sum of added background over frames divided by the total number of frames, $|\bar{\boldsymbol{A}}_b|$. The artifact considered here is the presence of added background. It can be noticed that added background impact of perceived quality tends to reduce with the increase of the amount of such artifacts.

Figure 4 reports the MOS versus the logarithm of the amount N of artifacts introduced. We used the logarithm of N because it provided a better fit. This means that the differences in the amount of artifacts are not linearly perceived. The MOS curves corresponding to holes (dashed line) and added regions (continuous line) are plotted on the same graph for comparison purposes. By comparing the two curves, it is evident that the presence of holes in the segmented object affects more importantly the perceived quality when compared to the presence of added regions.

Figure 5 depicts the MOS versus the distance in pixels from the boundary of the reference. In this figure, the artifact is analyzed in terms of the distance $\boldsymbol{d}_H$. As can be noticed, the deeper the hole, the more annoying the artifact, as the artifact affects the shape of the object. The curve for such artifact does not exhibit an as good fit, probably because, in this case also size and shape of the holes should be taken into account. The fitting parame-

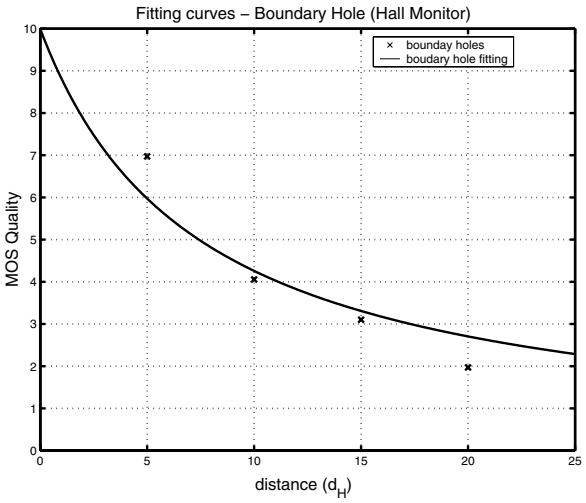

Figure 5. Mean quality curve corresponding to the d_H of the boundary hole errors.

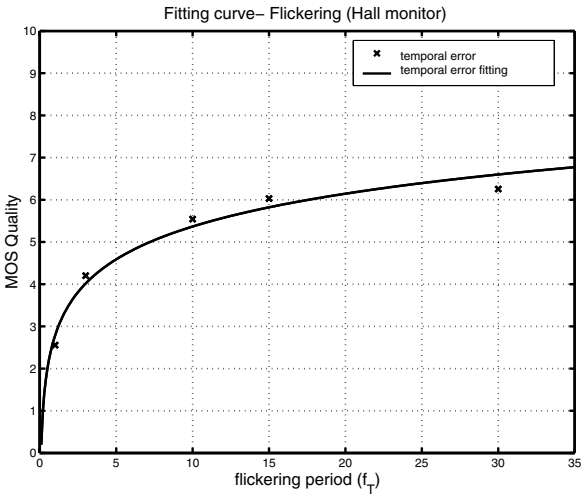

Figure 6. Mean quality curve corresponding to the temporal flickering period f_T.

Table 3. Fitting parameters for logistic function approximation curves of objective errors versus subjective MOS. The sum of the absolute values of residuals is r.

Test Error	x_{mean}	β	r
Added background	5.0557	1.4806	0.5779
Added regions	2.4220	0.4717	0.2146
Closed Holes	1.6225	1.0092	0.3290
Boundary holes	13.7185	1.8497	1.6190

ters are shown in columns 2, 3 and 4 of Table 3.

In Figure 6 we show the MOS versus the temporal error expressed as the period of flickering $\boldsymbol{f}_T$. The perceived flickering follows a logarithmic behavior, as the period of the artifact $\boldsymbol{f}_T$ increases. This can be explained by the fact that beyond a certain threshold such temporal artifacts are perceived similarly.

In this case, the MOS data behavior suggests a logarithmic curve to fit the data:

$$y = a + b * log(x) \tag{14}$$

The fit returned the coefficients $a = 2.7814$ and $b = 1.122$. The sum of absolute values of residuals for this fit was 0.1193.

6. Conclusions and Future Work

A perceptually driven segmentation evaluation is presented in this paper together with a method to carry out subjective tests on video object segmentation quality assessment. A psychophysical experiment was performed to assess the different perceptual importance of errors. The analysis of the subjective data versus the objective measures of errors introduced in the segmentation has been done. The fitted curves indicate how segmentation errors can be objectively described as a function of the error measures. Such a description can be used further in an objective segmentation evaluation metric.

At the moment, we are performing other subjective tests on other video sequences to further confirm the derived conclusions. To this end, we have refined the subjective protocol in collaboration with psychophysics test experts as follows. The subjects are not asked anymore about the *quality* but about the *annoyance* caused by the artifact.

Acknowledgments

The authors would like to thank John M. Foley for his valuable input, particularly in the data collection and experiment design.

References

[1] Rafael C.Gonzalez and Richard E. Woods, *Digital Image Processing*, Addison Wesley Longman, 2000.

[2] B. Sankur Ç. E. Erdem and A. M. Tekalp, "Metrics for performance evaluation of video object segmentation and tracking without ground-truth," in *Proc. SPIE, Int. Conf. on Visual Communications and Image Processing, Lugano, Switzerland*, 2003, vol. 5150, pp. 29–40.

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

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

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

[6] Call for AM Comparisons, "Compare your segmentation algorithm to the cost 211 quat analysis model http://www.iva.cs.tut.fi/cost211/call/call.htm," .

[7] Sarkar S. Sanocki T. Bowyer K.W. Heath, M.D., "A robust visual method for assessing the relative performance of edge detection algorithms," *Transactions on Pattern Analyis and Machine Intelligence*, vol. 19, pp. 1338–1359, 1997.

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

[9] 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.

[10] 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.

[11] D. Huttenlocher, D. Klanderman, and A. Rucklige, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, September 1993.

[12] E. Drelie, E. Salvador, and T. Ebrahimi, "Intuitive strategy for parameter setting in video segmentation," in *Proc. of Visual Communication and Image Processing,Lugano,Switzerland*.

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

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