IMPACT OF TOPOLOGY CHANGES IN VIDEO SEGMENTATION EVALUATION

Elisa Drelie Gelasca, Touradj Ebrahimi

Signal Processing Institute
Swiss Federal Institute of Technology EPFL
CH-1015 Lausanne, Switzerland

Mylène C. Q. Farias, Sanjit K. Mitra

Department of Electrical Engineering
University of California Santa Barbara
Santa Barbara, CA 93106, USA

ABSTRACT

This work addresses the problem of studying and characterizing topology changes between resulting and reference segmentation masks in video sequences. In particular, the goal of this paper is to examine the impact of individual and combined artifacts found in video object segmentation applications (e.g., added regions and holes). Added regions and holes artifacts are synthetically generated and inserted in a segmentation mask. We performed a psychophysical experiment in which human subjects were asked to rate the annoyance of the generated artifacts when presented alone or in combination. The results show how individual objective metrics can be derived and how an overall objective metric can be predicted by linearly combining individual segmentation errors for a specific video content.

1. INTRODUCTION

Applications such as object-based coding, video databases, interactive video and remote surveillance are based on a representation of the video content in terms of video objects. The first step of object-based applications is the identification of the areas of a video sequence that correspond to meaningful regions i.e., objects. This step is generally performed by a segmentation algorithm.

During the past three decades, different video segmentation techniques have been proposed to extract the objects of interest from a video sequence. However, no single segmentation technique is universally useful for all applications and different techniques are not equally suited for a particular task. In recent years, in order to properly evaluate the performance of segmentation techniques, objective metrics have been proposed [1], [2], [3], [4], [5].

To validate an objective metric, subjective experiments need to be performed. Subjective experiments are also used as a research tool to better understand how humans perceive artifacts and judge quality. On the basis of the analysis of the subjective data, topology changes between a reference and the resulting segmentation mask (or segmentation *artifacts*) can be characterized and a more reliable objective metric can be developed. With this purpose, in this paper we present an analysis of two artifacts produced by typical segmentation algorithms (i.e., added regions and holes). A subjective test has been carried out to measure the annoyance of these artifacts when presented alone or in combination. We studied the different levels of annoyance produced by these artifacts at different sizes and by their combinations. The idea is to develop individual metrics for the most relevant artifacts and to combine them into an overall quality metric towards a perceptually driven segmentation evaluation metric.

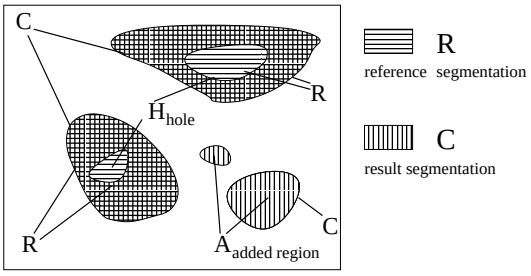


Fig. 1. Reference segmentation R overlapped to the resulting segmentation C. *Spatial artifacts* under investigation are depicted.

In this paper, we also present an experimental method for subjective evaluation of segmented video sequences. The task of defining a formal method for subjective tests in video object segmentation quality assessment is very useful, since to the best of our knowledge, only informal tests have been performed [3], [4].

The paper is organized as follows. The description of synthetic artifacts and the test sequences generated for the subjective experiments are presented in Section 2. The experimental method is described in Section 3. Subjective results are analyzed in Section 4. Finally, Section 5 draws the conclusions.

2. GENERATION OF SYNTHETIC ARTIFACTS

To determine the topology changes between a reference and a resulting segmentation mask, the difference between the two segmentation masks has to be computed. These changes (segmentation artifacts) can affect the quality of a segmented video in two ways: statically (*spatially*) and dynamically (*temporally*). In this work, we concentrated on the annoyance of two kinds of spatial artifacts: *added regions* and *holes* that are among the spatial artifacts typically introduced by the most common segmentation algorithms. Segmentation artifacts are defined by the amount of missegmented pixels (or pixel errors) present in the resulting segmentation mask. An algorithm for object segmentation can in principle be evaluated by estimating only these pixel errors [1], [3], [4] and [5].

In this paper, we focused on segmentation of *moving objects* in video sequences. An *object* is a semantically meaningful region. Let us define R as the set of all the objects belonging to the reference segmentation mask. Similarly, C is defined as the set of all the objects and regions in the resulting segmentation mask.

Pixels in the resulting segmentation mask C which do not belong to the reference segmentation mask R are defined as *false*

(a) Segmentation mask $\mathcal{S}$. (b) Segmentation under test.

Fig. 2. Sample frame of *Group* test sequence with combined artifacts: *holes* and *added regions*.

positive pixels. *False negative* pixels, on the other hand, are defined as the pixels which belong to the reference segmentation R, but not to the resulting segmentation mask C.

Added region pixels, $\mathcal{A}$, are those sets of false positive pixels not connected to any objects of R. Let us define $|\mathcal{A}|$ as the cardinality of $\mathcal{A}$, i.e., the number of pixels contained in $\mathcal{A}$. *Holes* are those sets, $\mathcal{H}$, of false negative pixels completely inside the objects of R. In the following, $|\mathcal{H}|$ is the cardinality of $\mathcal{H}$. These spatial artifacts are indicated in Figure 1.

To generate the test sequences, we chose a set of four video sequences of 60 frames each: Coastguard, Hall monitor, Group (a European IST project *Art.live* sequence), and Highway (an MPEG-7 test sequence). To insert the artifacts, we modified the reference segmentation masks. For two of the sequences (Group and Highway) the reference masks were obtained by hand. For the other two sequences, they were made available by the MPEG committee [1].

Many and different experiments can be generated by changing the position of artifacts along the test sequence in order to investigate their annoyance in the temporal dimension. In this experiment, the temporal variation of spatial artifacts is not under investigation. We keep the same position for the spatial artifacts, since we want to study the interactions between different artifacts and to find whether a simple model is able to predict how the two artifacts combine to determine the overall annoyance. For each reference segmentation mask, two new test segmentation masks were created: one with only *added regions* and one with only *holes*. The test segmentation masks, $\mathcal{S}$, were obtained by combining *added regions* and *holes* (see Fig. 2 (a)) of different sizes as given by the following equation:

$$\mathcal{S} = (R \backslash \mathcal{H}) \cup \mathcal{A} \tag{1}$$

where $\backslash$ denotes a set difference. By varying $|\mathcal{A}|$ and $|\mathcal{H}|$ we obtained the twenty combinations shown in columns 2-3 of Table 1. We did not use all possible combinations because that would make the experiment too long. The artifacts presented the same shape (square) and were constant in number (three). Column 1 shows the indices of the test combinations. Finally, we obtained the test sequences by showing the texture of the original video in correspondence to the segmented *moving* objects/regions over a uniform green background, as shown in Figure 2 (b). A total of 80 test sequences were used in this experiment (20 combinations $\times$ 4 reference segmentation masks).

[1] MPEG Home Page, http://mpeg.telecomitalialab.com/

Test comb.	$\|\mathcal{A}\|$	$\|\mathcal{H}\|$	Coast. *MAV*	High. *MAV*	Group *MAV*	Hall *MAV*
1	0	0	4.6	3.9	6.7	7.0
2	9	0	8.7	23.0	16.2	12.2
3	25	0	13.5	22.3	20.9	17.6
4	49	0	20.3	27.7	23.0	24.1
5	81	0	24.2	30.1	27.9	28.6
6	169	0	29.9	33.7	30.0	30.3
7	0	9	28.3	23.2	50.0	42.2
8	0	25	43.9	36.3	54.6	63.1
9	0	49	45.8	35.3	58.2	47.1
10	0	81	56.8	50.8	62.8	52.7
11	0	169	62.4	68.0	70.4	64.1
12	25	9	40.5	27.6	50.1	41.3
13	9	25	38.9	32.8	53.1	70.7
14	25	25	45.1	36.5	52.1	67.0
15	81	25	61.6	46.2	54.8	65.1
16	169	49	68.8	62.7	65.1	56.7
17	25	81	61.5	51.5	65.6	56.3
18	81	81	59.9	65.3	74.8	54.7
19	49	169	72.8	70.8	76.5	63.6
20	169	169	77.8	64.0	77.7	66.7

Table 1. *MAV*s for all segmented video sequences and all combination $|\mathcal{A}|$ and $|\mathcal{H}|$ used in the experiment.

3. EXPERIMENTAL METHOD

For the subjective evaluations to be meaningful and comparable, a set of standards and grading techniques are defined by ITU-R [6] and ITU-T [7]. However, there are no recommended standards for the evaluation of segmented video sequences. Nevertheless, the subjective video segmentation quality evaluation is not completely *ad hoc* and an indicative set of guidelines has been provided [8]. In this paper, we present a new method for subjective evaluation of segmented video sequences. This experimental method is an effort to make subjective evaluations in this field more reliable, comparable and standardized.

Each test session was composed of five stages: instructions, training, practice trials, experimental trials, and interview. In the first stage, the subject was verbally given instructions and was made familiar with the task of segmentation of moving meaningful objects. In the training stage, the original sequences, the reference segmentation masks and sample segmented masks were shown to establish the range for the annoyance scale. The display configuration showed the texture of the original image in correspondence to the segmented objects/regions over a uniform green background. The reference segmented masks and the original sequences were only shown in the training stage for two reasons. First, in real applications the reference and the original video are not always available. Second, in earlier experiments we noticed that subjects do not pay attention to the reference after the training. After the training, in order to familiarize the subject with the experiment and to stabilize the subjects' responses, practice trials were performed with a small subset of the test sequences.

The experimental trials were performed with the complete set of test sequences presented in a random order. Our test subjects were drawn from a pool of students in the introductory psychology class at UCSB. The 28 subjects were asked one question after each

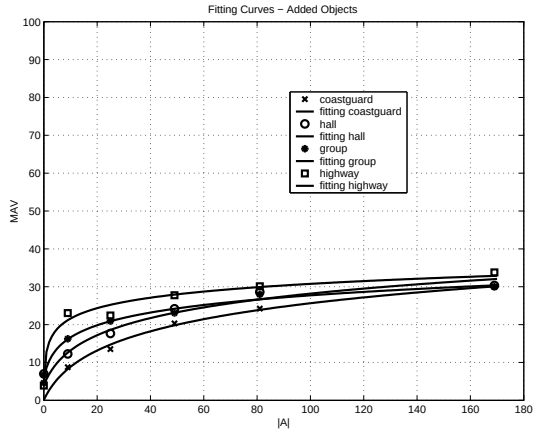


Fig. 3. Mean annoyance curves corresponding to error measure *added regions* $|\mathcal{A}|$. Fitting parameters for Coastguard a=-15.75, b=8.87, c=5.95 and absolute sum of residuals r=7.20. For Hall a=-6.76, b=7.52, c=4.13 and r=9.08. For Group, a=4.30, b=5.06, c=1.31 and r=3.90. For Highway a=12.32, b=4.00, c=0.11 and r=6.10.

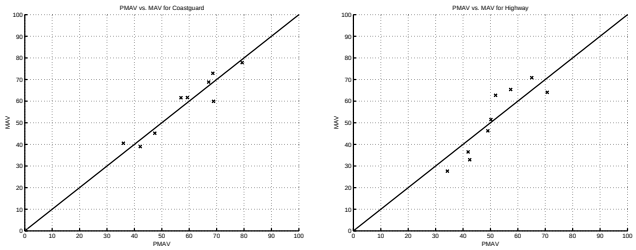
segmented video sequence was presented, "How annoying was the defect relative to the worst example in the sample videos". The subject was instructed to enter a numerical value greater than 0. The value 100 was to be assigned to artifacts as annoying as the most annoying artifacts in the sample video sequences. Although we tried to include the worst test sequences in the sample set, we acknowledge the fact that the subjects might find some of the other test sequences to be worse, and we specifically instructed them to go above 100 in those cases. The subjects were then told that artifacts would appear combined or by themselves and they should rate the overall annoyance in both cases.

Finally, in the interview stage, we asked the test subjects for qualitative descriptions of the defects that were perceived. The qualitative descriptions are useful for categorizing the defect features seen in each experiment and help in the design of future experiments.

4. DATA ANALYSIS

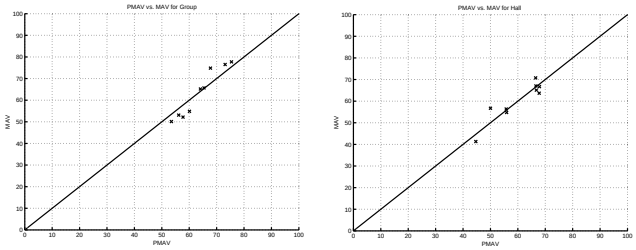
We used the standard methods [7] to analyze and screen the judgments provided by the test subjects. From the data gathered we calculated the Mean Annoyance Value (*MAV*) of each test sequence.

The values for the *MAV* for all the video sequences are shown in columns 4-7 of Table 1. The test combination number 1 corresponds to the reference segmentation without any artifact. It is interesting to notice that the values for the *MAV*s corresponding to the reference are not zero, indicating that subjects reported that these sequences, normally used as terms of comparison in many objective metrics, contain annoyance levels different from zero. The test combinations 2-6 and 7-11 correspond to sequences with only one type of synthetic artifact: *added regions* or *holes*, respectively. It can be observed from these combinations that the pure artifact *holes* obtained higher *MAV*s compared to the pure *added region* artifact, both being equal in size, number and shape. This confirms the hypothesis used by most of the objective metrics that *holes* contribute the most in annoyance with respect to *added regions* with the same size.



(a) Coastguard. (b) Highway.

Fig. 4. Predicted Mean Annoyance Value versus Mean Annoyance Value for the segmented video sequence Coastguard and Highway.



(a) Group (b) Hall monitor

Fig. 5. Predicted Mean Annoyance Value versus Mean Annoyance Value for the segmented video sequence Group and Hall.

On the other hand, the *MAV* values for *added regions* artifact (test combinations 2-6) point out another important aspect not yet remarked. Independently from the video content, the *added regions* have perceived annoyance values which quickly reach a saturation level for larger sizes. In order to illustrate this result, we plotted all the *MAV* values versus the the *added region* error measures $|\mathcal{A}|$ for all the reference video sequences. The *MAV* data suggested a logarithmic curve to fit the data:

$$y = a + b * log(x + c) \tag{2}$$

Figure 3 contains both the *MAV added regions* values and the fitting curves for each video. The perceived added region annoyance data follow a logarithmic behavior as the size of the artifact increases. Their perceived annoyance changes very little with the video content for larger sizes of artifacts. This result has to be taken into consideration when an objective metric is formulated. The type and the salience of the correctly extracted video objects do not influence the perception of this kind of artifact.

As can be observed from the *MAV* values of test combinations 7-11 in Tab.1, the previous conclusion is not valid anymore for the artifact *holes*. For the pure *holes* artifact, the annoyance changes according to the content, the size and the relevance of the extracted objects. That is a consequence of the fact that more relevant objects attract more the eye gaze. Therefore, a bad segmentation of those objects implies in higher annoyance values, as remarked by Correia in [4]. A further investigation of this artifact needs to be carried out.

To complete the analysis of the data we tried to predict the *MAV* of the combined artifacts from the two *MAV* values, MAV_{add} and MAV_{hole}, respectively for the pure artifacts *added regions* and

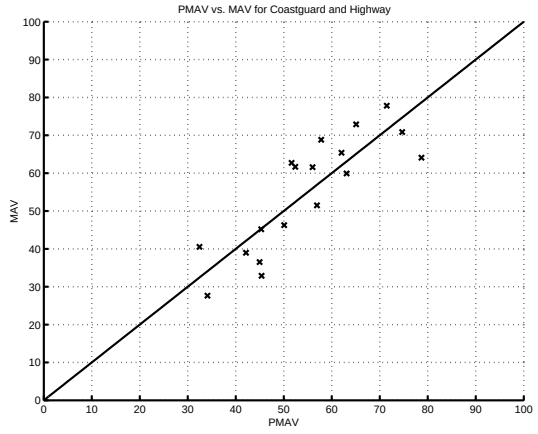


Fig. 6. Predicted Mean Annoyance Value versus Mean Annoyance Value for the segmented video sequences Coastguard and Highway.

Video	α	β	R^2
Coastguard	1.11	0.73	89.62
Highway	0.93	0.57	78.31
Group	0.33	0.92	85.82
Hall	0.01	1.05	85.52
Coastguard and Highway	0.66	0.82	85.41

Table 2. Regression analysis parameters for Equation 3.

holes. We fitted the following equation and estimated the coefficients:

$$PMAV_j = \alpha * MAV_{add} + \beta * MAV_{hole} \qquad (3)$$

PMAV is the predicted value of *MAV*, α and β are the regression coefficients. Table 2 summarizes the results obtained for the fit. We found that a linear regression model provided a good fit to the data. Column 4 shows the squared correlation (R^2) for the fit.

In Figures 4-5, we plotted the *PMAV* versus the *MAV* for the video Coastguard, Highway, Group and Hall. The regression coefficients for Coastguard and Highway, of similar content, are close. In Figure 6, we plotted the *PMAV* versus the *MAV* for the data set containing both the video sequences Coastguard and Highway. The fit to the data set in Fig. 6 is reasonably good and the correlation coefficient is 85.41.

The regression coefficients for the videos for which the segmented objects contained people (Hall and Group) are quite different from those of the other two sequences. Not only the size of the artifact influences the annoyance but also the interaction of the artifact with the relevant features of the correctly segmented video objects (such as the size, the position and the kind, e.g., people faces).

In summary, this simple linear model with no interactions fits well the data. The fitting coefficients are close for two test sequences (Coastguard and Highway). In the case of segmented video content where people take part (Hall and Group), the results of the linear model could not be grouped together, since other high level factors need to be taken into consideration. Further investigation on these cases needs to be carried out.

5. CONCLUSIONS

In this paper, we created two synthetic segmentation artifacts - added regions and holes. We presented them alone or in various combinations and had subjects rate the annoyance of the perceived artifacts. When the artifacts were presented alone, the annoyance judgments show that the annoyance of *added regions* is almost independent of the video content especially for larger artifacts. The *MAVs* suggest a logarithmic curve (see Equation 2) to describe the perceived annoyance caused by added regions. This function can be used in an objective metric with *reference*. The size of the added region (x in Eq. 2) can be provided by computing the difference between the reference and the resulting segmentation masks. The annoyance (y in Eq. 2) is then obtained by using the coefficients a, b, c indicated in Fig. 3.

When the artifacts were presented in combination, a simple linear model with no interactions predicted how the artifacts combine to determine the overall annoyance. The linear model of Equation 3 provides a good fit. The estimated coefficients in Table 2 were close for segmented video content not containing specific regions of interest such as people. This simple model with the estimated coefficients α and β shows how an overall objective metric can be predicted by linearly combining the individual metrics for spatial artifacts *added regions* and *holes*.

6. REFERENCES

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

[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] A. Cavallaro, E. Drelie Gelasca, 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.

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

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

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

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

[8] 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," .