ANNOYANCE OF SPATIO-TEMPORAL ARTIFACTS IN SEGMENTATION QUALITY ASSESSMENT

Elisa Drelie Gelasca, Touradj Ebrahimi

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

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

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

ABSTRACT

This paper describes the results of a series of subjective experiments that investigated the annoyance caused by the most common artifacts present in segmented video sequences. Various types of artifacts were inserted into a reference segmented video considered as ideal and shown to our test subjects. The artifacts varied in their location, size, appearance and duration. Annoyance of segmentation artifacts are found to be tied up with their intrinsic characteristics (e.g., size, position) but only weakly related to the video content. The results identify the characteristics that should be taken into account in the design of a perceptually driven objective metric.

1. INTRODUCTION

he process of identifying and extracting a collection of *meaningful areas* in an image/video corresponding to *objects* in the real world is referred to as semantic video object extraction or *segmentation*. The main requirements of segmentation are: *spatial accuracy* that is the precise definition of the object boundary, and *temporal coherence* that can be seen as the property of maintaining the spatial accuracy in time. A great variety of segmentation algorithms have been developed in the past and new techniques are proposed each year. However, none of the proposed solutions are applicable to all types of video sequences and applications. These two reasons stress the increasing importance of *objective evaluation* of segmentation algorithms as demonstrated by the efforts of the European *Cost 211* group [1].

While quite a few metrics have been proposed [2, 3, 4, 5], the evaluation of their performance has received much less attention. Ad hoc 'informal tests' are usually carried out [3, 4], and *subjective evaluations* still pose practical (time-consuming, expensive and complex set-up) and theoretical problems (lack of established procedure for comparison and ranking of segmentation quality).

In this paper, we synthetically generated the most common artifacts present in video object segmentation and introduced them in test video sequences. We propose an experimental method for assessing the subjective video segmentation quality. Then, a subjective test was performed with the goal of estimating the degree of annoyance caused by these artifacts. The results show how an objective metric can be derived from the analysis of subjective data. The paper is divided as follows. In Section 2 we describe how the synthetic artifacts were created. In Section 3 we present the proposed experimental method. In Section 4, the subjective evaluation results are analyzed. Section 5 draws some conclusions.

2. GENERATION OF SYNTHETIC SEGMENTATION ERRORS AND TEST SEQUENCES

In literature [5], *reference objective metrics* consider segmentation artifacts as a set of mismatched pixels. These metrics compare the segmentation results with a correct/ideal *reference segmentation*. Generally, the comparison is carried out according to the number and position of misclassified pixels. In our work, we model the segmentation artifacts in a more structured form: each error pixel and its neighborhood are considered as a connected set. For each localized connected set of misclassified pixels (segmentation artifact), we study its characteristics (e.g. shape, location, size) in terms of annoyance.

In this paper, we focus on video segmentation of objects (semantically meaningful regions), and define R as the set of all objects belonging to the reference segmentation. Similarly, C is defined as the set of all objects in the resulting segmentation. Pixels in the resulting object segmentation, C, which do not belong to the reference object segmentation, R, are defined as *false positive* pixels. *False negative* pixels, on the other hand, are defined as pixels which belong to the reference segmentation R, but not to the resulting segmentation C. An initial coarse estimation of the segmentation quality can be done by estimating these false pixels [5].

In our work, false positive pixels are further divided into two categories: *added background* pixels $\mathcal{B}$ and *added region* pixels $\mathcal{A}$. Furthermore, false negative pixels can be grouped in *holes*, $\mathcal{H}$, in which *closed holes* and *boundary holes* can be differentiated. Figure 1 illustrates an example of *spatial artifacts* described above. In this work, we concentrated on estimating the annoyance of two kinds of *spatial artifacts*, namely, *added regions* and *holes*. We did not measure the annoyance of *added background*, $\mathcal{B}$ nor *missing objects*. In addition, a special kind of temporal artifact was investigated, namely, *temporal variation of added regions*.

To generate the test sequences, a set of four video sequences of 60 frames of size 352×288 pixels, representing a sequence of duration 4.8 seconds each were chosen: Coastguard, Hall monitor, Group (a European IST project *Art.live* sequence), and Highway (an MPEG-7 test sequence). To introduce the artifacts, the reference segmentation masks were modified. For two of the sequences (Group and Highway) the reference masks were obtained by hand. For the other two sequences, we used publicly available masks produced by the MPEG committee [1].

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

2.1. Added Regions

In this work, the annoyance produced by *added region* artifacts, $\mathcal{A}$, was studied by varying its size, position and shape. We artificially mis-segmented three portions of the background completely disconnected from the correctly segmented foreground objects. In a previous experiment, we noticed that an increase in the number of added regions follows a non-symmetrical function approximating the standard logistic function [6]. In this experiment, the number of added regions was not under investigation. Therefore, we kept the number of regions equal to three and varied the size, position and shape of the artifact for each test sequence.

The shape of the *added region* was varied by adopting a super-ellipse function. By modifying the super-ellipse parameters, a continuum of several shapes can be formed, ranging from a circle to a square. The topology of the reference segmentation was varied in the following way. First, we positioned the group of three added regions in three different random positions (p_1, p_2 and p_3) far from the reference objects. Then, for each of these positions, two different shapes were generated with four different sizes (2×2, 5×5, 10×10, 20×20). The total number of test sequences for this part of the experiment was 75 which included 72 test sequences (3 reference segmentations × 3 positions × 4 sizes × 2 shapes) plus the 3 reference segmentations without any artifact of *Hall monitor*, *Highway* and *Group*.

2.2. Holes

In the objective metrics proposed in the literature, holes are only considered in terms of uncorrelated set of pixels and their distances from the reference boundary of the object [3, 4]. According to [5] the more distant a hole is from the boundary of the object, the more annoying becomes the artifact. The authors conclude that as we move away from the border, holes become more annoying. Boundary holes only make the object thinner. Therefore, they are less annoying for the human observer.

In our experiment, we studied if this condition is still valid for large holes. In this case the annoyance caused by a boundary hole could be worse than for a closed hole (completely inside the object). This could be justified by the fact that if the shape of the object is completely modified by a large hole on the boundary, the object can become harder to recognize. On the other hand, in the presence of a large closed hole completely inside an object, the object can be still recognizable and, consequently, this artifact becomes less annoying. For this purpose, we synthetically inserted a group of three holes at three positions: on the contour of the object, d_0 (boundary hole), and in two inner positions, d_1 and d_2 (closed holes). For each position, we generated 4 sizes (3×3, 5×5, 9×9,13×13) of holes. The total number of test sequences for this part of the experiment was 52 which included: 48 test sequences (4 reference segmentations × 3 positions × 4 sizes) plus the 4 reference segmentations of *Hall monitor*, *Highway*, *Coastguard* and *Group*.

2.3. Temporal Error

In video segmentation, an artifact often varies its characteristics through time. In this work, we considered the appearance and disappearance of added regions through time as a typical temporal artifact. Different variations of spatial artifacts can be implemented to test the effect of temporal artifacts. In a previous experiment, we chose to change the position of added regions along the test sequence. Their position changed after every N frames. N was different for each test sequence (starting from a temporally smooth change of added region position, for a large N, until a very fast and annoying flickering, for a small N).

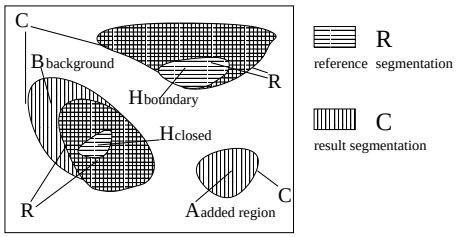

Fig. 1. Reference segmentation R overlapped to the resulting segmentation C. Different kinds of *spatial* errors are depicted.

In this experiment, we want to find whether there is an *expectation* effect and how this affects the overall perceived quality. By *expectation* we mean the effect that a good segmentation at the beginning could create a good overall impression on assessing the quality of the sequences under test. Three regions of the same size (10×10) were added always at the same position along the entire video sequence. The added regions appeared and disappeared along the time causing a temporal artifact. The mathematical expression for this temporal artifact, B, is given by the following formula:

$$B(t_1, t_2) = S(t - t_1) - S(t - t_2) \quad (1)$$

where S is the step function, with t_1 the start and t_2 the end of the temporal artifact.

Figure 2 shows an illustration of how added regions were inserted in the sequences in order to create the temporal artifacts. Condition 1 corresponds to the reference sequence, while condition 2 corresponds to a sequence with the added regions present in all 60 frames. Conditions 3-5 are cases where the added regions were inserted in 10 out of 60 frames. They were inserted in different parts of the video sequence: at the beginning ($B(1, 10)$), at the end ($B(50, 60)$), and in the middle ($B(25, 35)$). Conditions 5-9 correspond to combinations of these three previous occurrences. A total of 9 test conditions and two test video sequences *Hall monitor* and *Coastguard* were used.

3. EXPERIMENTAL METHOD

Standard subjective evaluation methodologies for video segmentation quality are not yet available. We propose an experimental method for subjective evaluation based on those established for video quality evaluation [6], [7]. This method is an effort to make subjective evaluations in this field more reliable and comparable.

Each test session was composed of five stages: instruction, training, practice trials, experimental trials, and interview. In the first stage, the subject was verbally given instructions. He/she was made familiar with the task of segmentation by considering the specific case in which only moving objects had to be segmented. In the training stage, the original video, the reference segmentation masks, and samples of test segmentations were shown to establish the range of the annoyance scale. The implemented graphical interface displayed the texture of the original image in correspondence to the segmented moving objects/regions over a uniform green background. After the training, in order to familiarize

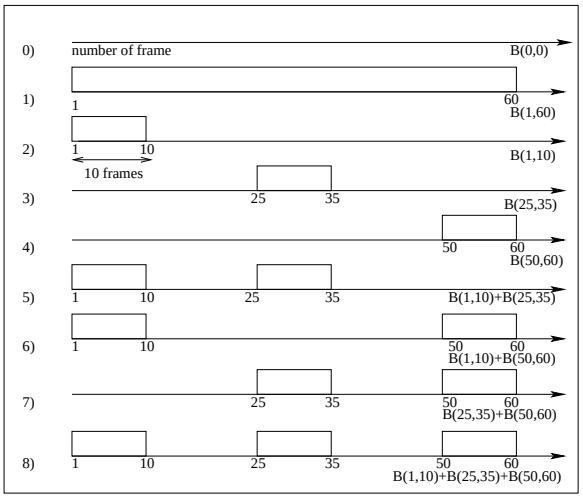

Fig. 2. Temporal insertion of added object for 10 frames in different moments of the video sequence

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 28 students in the introductory psychology class at UCSB. The subjects were asked one question after each segmented video sequence was presented, "How annoying was the defect relative to the worst example in the sample video?". The subject was instructed to enter a numerical value greater than 0. The value 100 was to be assigned to the most annoying artifacts in the sample video sequences. We specifically instructed subjects to go above 100 in case they might find some of the test sequences to be worse than the worst case in the sample set.

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

4. DATA ANALYSIS

We used the standard methods [7] to analyze and to screen the judgments provided by the test subjects. From the data gathered we calculated the Mean Annoyance Value (*MAV*) of each test sequence. First, we obtained the *MAV* values versus the cardinality $|\mathcal{A}|$ of **added regions**, (i.e., the number of pixels contained in $\mathcal{A}$) by averaging the *MAV* values gathered for the two different shapes (squares and circles). The idea was to derive an objective measure based on the annoyance of added regions independent of their shapes. Since there was very little difference among the *MAV* values for the different positions (p_1, p_2, and p_3), we averaged the *MAV* values for the three positions. We obtained a more general result that is independent of the position and shape of the added region. In order to illustrate this result, we plot in Figure 3 the *MAV* values versus $|\mathcal{A}|$ for all test sequences.

The *MAV* data suggested a logarithmic curve to fit the data:

$$y = a + b * log(x + c) \quad (2)$$

Fig. 3 contains both the *MAV added regions* values and the fitting curves for each video. This function can be used to derive an objective metric taking into account the following considerations: the perceived annoyance of added regions has logarithmic behavior as a function of the size of the artifact; the shape and the position of the added region do not influence the annoyance; their perceived annoyance changes little with the video content, but a difference can be noticed between *Group*/*Hall* and *Highway*. This could be explained by the fact that the segmented objects in *Highway* (segmented cars in a traffic monitoring scene) are smaller than those in *Hall* and *Group* (people walking in front of the camera). This shows that the size of the added regions in relation to the size of the correctly segmented objects should be taken into account while designing an objective metric. This result also shows that the type of correctly segmented video objects do not influence the visibility of this artifact. However, the size of correctly segmented object have some influence on the overall perceived annoyance.

Figures 4-5 shows the plots of the *MAVs* as a function of the cardinality $|\mathcal{H}|$ of **hole** artifacts for all video sequences. Each graph shows two curves for each video sequence, one corresponding to the boundary holes ($\mathcal{H}_b$) *MAVs* and the other corresponding to the closed holes ($\mathcal{H}_c$) *MAVs*. The boundary holes curve increases faster than the closed holes curve. For small values of the size, $\mathcal{H}_b$ is more annoying than $\mathcal{H}_c$, as already reported in the literature [5]. By increasing the size of the holes $|\mathcal{H}|$, a *point of inversion* can be noticed concerning the annoyance of the two kinds of artifacts (see Fig.4). After that point of inversion, $\mathcal{H}_b$ is more annoying than $\mathcal{H}_c$, since the shape of the object becomes less recognizable. For all the sequences tested (*Highway*, *Coastguard*, *Hall Monitor* and *Group*) independently from the content, the point of such inversion starts between sizes $\mathbf{5}\times\mathbf{5}$ and $\mathbf{9}\times\mathbf{9}$. This subjective experiment indicates that both distance and size of the hole should be jointly taken into account when an objective metric is proposed.

Figure 6 shows the plots of the *MAVs* for the **temporal artifact** $B(t_1, t_2)$ for each test sequence. Even though the contents of the two test sequences are very different, the two curves obtained for the *MAVs* look quite similar. This is especially true for complex temporal artifacts and when the temporal defects become similar to a *flickering*. For both video sequences, the most annoying artifacts are those with more temporal variation of added regions. By looking at Figure 6, a surprising result is that the initial temporal variation ($B(1, 10)$+$B(25, 35)$) is worse than the final temporal variation ($B(25, 35)$+$B(50, 60)$) for both video sequences. This could be explained in terms of *expectation*. A good segmentation at the beginning creates a good impression. A bad segmentation at the beginning puts the overall impression of the segmentation quality in jeopardy. It is possible that for longer video sequences, a *memory effect* would prevail the *expectation*. More tests with longer sequences are needed to confirm this result.

5. CONCLUSIONS

In this paper, we identified the degree of annoyance of some common spatial-temporal artifacts in segmentation quality assessment. To do so, a series of typical segmentation artifacts were generated, namely *added regions*, *holes* and *temporal artifacts*. These artifacts varied in their location, size, shape, as well as duration. An evaluation methodology derived from video quality evaluation was then used to carry out subjective tests. As a result, a logarithmic function was derived to model the degree of annoyance of added

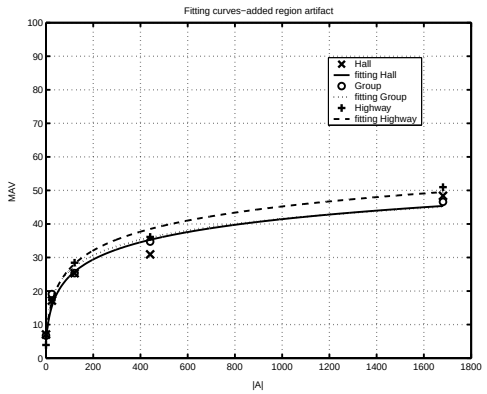

Fig. 3. Mean annoyance curves corresponding to the size $|\mathcal{A}|$ of the artifact *added regions*. Fitting parameters for Highway a=-12.24, b=8.3087 c=7.57 and absolute sum of residuals r=5.31. For Hall a=-11.25, b=7.6182, c=8.02, r=12.01. For Group, a=-6.04, b=6.88, c=5.27 and r=7.60.

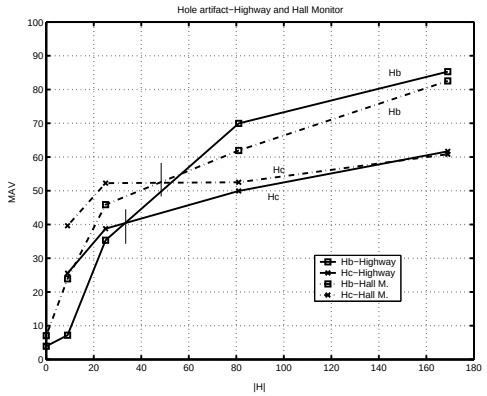

Fig. 4. Mean annoyance curves corresponding to *hole* artifacts versus their sizes $|\mathcal{H}|$ ($\mathcal{H}_c$ inside the object and $\mathcal{H}_b$ on the boundary of the object) for Highway and Hall video sequences.

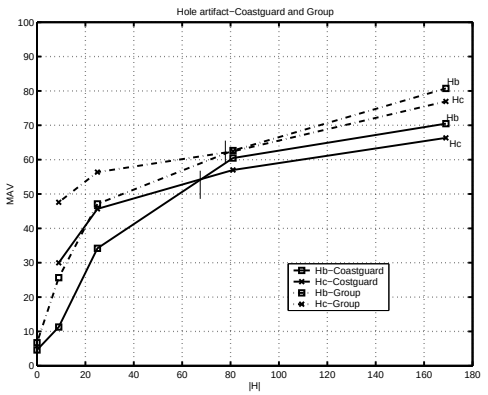

Fig. 5. Mean annoyance curves corresponding to *hole* artifacts versus their sizes $|\mathcal{H}|$ (inside the object $\mathcal{H}_c$ and boundary holes $\mathcal{H}_b$ on the boundary of the object) for Coastguard and Group sequences.

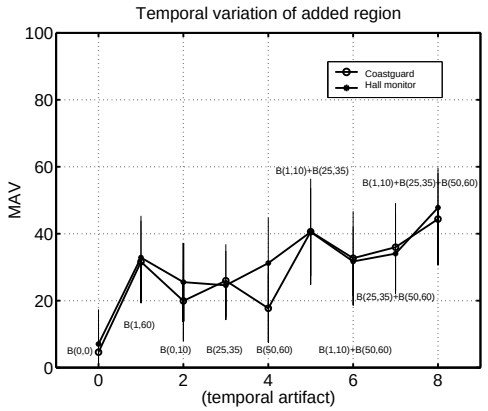

Fig. 6. Mean annoyance curves corresponding to *temporal variation* $B(t_1, t_2)$ and the corresponding confidence intervals for the video sequences *Hall Monitor* and *Coastguard*. The two curves look very similar for large combination of $B(t_1, t_2)$.

regions. It was further identified that annoyance of the latter artifacts do not depend on their location and shape but are weakly related to the video content. We further showed that a third dimension related to the size has to be taken into account for the evaluation of degree of annoyance due to hole artifacts. The number and the position along time of temporal artifacts influence differently the perceived annoyance. An early temporal artifact seems to affect most the overall perceived quality. This indicates that an expectation effect could play a role in segmentation quality assessment. More tests are needed to confirm this hypothesis and to find out possible interactions between an expectation effect and a memory effect.

6. REFERENCES

[1] 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," .

[2] 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.

[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 Assessment of the Quality of Television Pictures Recommendation BT.500-11*, International Telecommunication Union, Geneva, Switzerland, 2002.

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