SEGMENTATION-DRIVEN PERCEPTUAL QUALITY METRICS

Andrea Cavallaro

Multimedia and Vision Laboratory
Queen Mary, University of London
London E1 4NS, United Kingdom

Stefan Winkler

Genista Corporation
Rue du Theâtre 5
1820 Montreux, Switzerland

ABSTRACT

We present a full-reference and a no-reference perceptual video quality metric that incorporate both low-level and high-level aspects of vision. Low-level aspects include color perception, contrast sensitivity, masking as well as artifact analysis. High-level aspects take into account the cognitive behavior of an observer when watching a video by means of semantic segmentation. Using the special case of semantic face segmentation, we evaluate the proposed segmentation-driven perceptual quality metrics using a range of test sequences and demonstrate an improvement of their prediction performance.

1. INTRODUCTION

Reducing the bandwidth and storage requirements of images and video while increasing their visual quality is a priority in the development of new compression or transmission systems, and guaranteeing a certain level of quality has become an important concern for content providers. As it is typically the viewer who judges quality, subjective experiments have been the only accepted way of obtaining reliable quality ratings. Predicting these subjective ratings using an automatic visual quality metric with higher accuracy than peak signal-to-noise ratio (PSNR) has been the topic of much research in recent years.

Two approaches for perceptual quality metric design can be distinguished [1]: One class of metrics implements a general model of low-level visual processing in the retina and the early visual cortex. Metrics in this class typically require access to the reference video for difference analysis. The other class of metrics looks for specific features in the image, for example compression artifacts arising from a certain type of codec, and estimates their annoyance. However, none of today's metrics quite achieve the reliability of subjective experiments.

One of the common shortcomings of quality metrics is the fact that they analyze the entire scene uniformly, assuming that people look at every pixel of the image or video. In reality, we do not scan a scene in raster fashion. Our visual attention tends to jump from one point to another. These jumps are called *saccades*. Yarbus [2] demonstrated that the saccadic patterns depend on the visual scene as well as the cognitive task to be performed. The studies of Bajcsy [3] also led to the conclusion that *we do not see, we look*. We focus our visual attention according to task at hand and the scene content.

In this work, we attempt to emulate the human visual system to prioritize the visual data in order to improve the prediction performance of perceptual quality metrics. Section 2 discusses the factors influencing the cognitive behavior of people watching a video. In Section 3 we describe how this behavior can be incorporated into a quality metric by means of semantic segmentation. The prediction performance of the proposed metrics is discussed for the special case of face segmentation in Section 4. Finally, we draw some conclusions and describe the directions of our current work in Section 5.

2. COGNITIVE BEHAVIOR

While most vision models and quality metrics are limited to lower-level aspects of vision, the cognitive behavior of people when watching video cannot be ignored. However, cognitive behavior may differ greatly between individuals and situations, which makes it very difficult to generalize. Nevertheless, two important aspects can be pointed out, namely the *focus of attention* and the *tracking of moving objects*.

2.1. Focus of Attention

When watching video, we focus on particular areas of the scene. Studies have shown that the direction of gaze is not completely idiosyncratic to individual viewers. Instead, a significant number of viewers will focus on the same regions of a scene [4]. Naturally, this focus of attention is highly scene-dependent. Maeder *et al.* [5] proposed constructing an importance map for the sequence as a prediction for the focus of attention, taking into account perceptual factors such as edge strength, texture energy, contrast, color variation, homogeneity, etc.

0-7803-8554-3/04/$20.00

One of the objects attracting most of our attention are people and especially human faces. If there are faces of people in a scene, we will look at them immediately. Furthermore, because of our familiarity with people's faces, we are very sensitive to distortions or artifacts occurring in them. The importance of faces is also underlined by a study of image appeal in consumer photography [6]. People in the picture and their facial expressions are among the most important criteria for image selection.

2.2. Object Tracking

In a similar manner, viewers may also track specific moving objects in a scene. In fact, motion tends to attract the viewers' attention. Now, the spatial acuity of the human visual system depends on the velocity of the image on the retina: as the retinal image velocity increases, spatial acuity decreases. The visual system addresses this problem by tracking moving objects with smooth-pursuit eye movements, which minimizes retinal image velocity and keeps the object of interest on the fovea. Smooth pursuit works well even for high velocities, but it is impeded by large accelerations and unpredictable motion [7]. On the other hand, tracking a particular movement will reduce the spatial acuity for the background and objects moving in different directions or at different velocities. An appropriate adjustment of the spatio-temporal contrast sensitivity function (CSF) as outlined in [8] to account for some of these sensitivity changes can be considered as a first step in modeling such phenomena.

3. SEGMENTATION-ENABLED QUALITY METRICS

Based on the observations of the previous section, the proposed perceptual quality metrics take into account both low-level and high-level aspects of vision. To achieve this, a segmentation stage is added to the metrics to find regions of interest. Its output then guides the pooling process by giving higher weight to the regions with semantically higher importance.

3.1. Low-level Contribution

We used two different metrics in this work, a full-reference perceptual distortion metric (PDM) based on a vision model, and a no-reference video quality metric based on the analysis of common artifacts.

The full-reference PDM is based on a contrast gain control model of the human visual system that incorporates spatial and temporal aspects of vision as well as color perception [9]. The metric requires both the reference sequence and the distorted sequence as inputs. After their conversion to a perceptual opponent-color space, each of the resulting three components is subjected to a spatio-temporal filter bank decomposition, yielding a number of perceptual channels. They are weighted according to contrast sensitivity data and subsequently undergo contrast gain control for the modeling of pattern masking.

The no-reference quality metric estimates visual quality based on the analysis of blockiness, blur and jerkiness artifacts found in the video [10]. It does not need any information about the reference sequence. The metric is part of Genista's *Media Optimacy* and *Stream PQoS* tools. The use of a no-reference metric is particularly interesting here because semantic segmentation does not require a reference video either.

Both of these metrics can make local quality measurements in small subregions over a few frames in every video. The process of combining these low-level contributions into an overall quality rating is guided by the result of the semantic segmentation stage described in the following section.

3.2. High-level Contribution

The high-level contribution to the quality metrics takes into account the cognitive behavior of people when watching a video. To represent the semantic model of a specific cognitive task, we decompose each frame of the reference sequence into sets of mutually exclusive and jointly exhaustive segments. This semantic model corresponds to a specific human abstraction, which need not necessarily be characterized by perceptual uniformity. The semantics (i.e. the meaning) are defined through human abstraction. Consequently, the definition of the semantic partition depends on the task to be performed. The partition is then derived by semantic segmentation. In general, the topology of the semantic partition cannot be expressed using homogeneity criteria, because the elements of such a partition do not necessarily possess invariant properties. Some knowledge of the objects we want to segment (*a priori* information) is therefore required.

For example, for segmenting moving objects, motion information can be used as semantics. The motion of an object is usually different from the motion of background and other surrounding objects. For this reason, many extraction methods make use of motion information in video sequences to segment objects [11]. An example of semantic segmentation result is shown in Figure 2.

If we want to segment faces of people, color-based segmentation can be used. A number of relatively robust algorithms for face segmentation are based on the fact that human skin colors are confined to a narrow region in the chrominance (C_B, C_R) plane [13], and their distribution is quite stable [14]. When the goal is to detect the presence of faces in a video and their location, a cascade of simple classifiers can be used [12]. Each classifier is trained to detect a specific face feature, such as the intensity difference

Fig. 1. Example of semantic segmentation result.

between the eye region and the upper cheek or between the eye region and the bridge of the nose. An example of face detection result is shown in Figure 3.

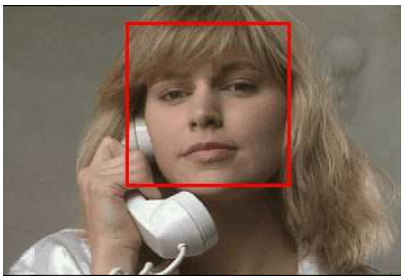

Fig. 2. Example of face detection result.

4. EVALUATION

4.1. Test Material

We used test sequences from three different subjective testing databases available to us:

1. VQEG Phase I database [15]. This database comprises mainly TV material with 16 test conditions. 3 relevant scenes were selected from this database to evaluate the full-reference PDM.

2. PC video database [16]. This database was created with CIF-size video and various DirectShow codecs at bitrates of 1-2 Mb/s, for a total of 8 test conditions. We picked 2 scenes from this database to evaluate the full-reference PDM.

3. Internet streaming database [10]. This database contains clips encoded with MPEG-4, Real Media and Windows Media at 256 and 512 kb/s as well as some packet loss (7 conditions in total). 4 scenes from this database were used. Due to the test conditions here, these sequences cannot be properly aligned with the reference. Therefore, we use this set for the evaluation of our no-reference metric.

The scenes we selected from these databases contain faces at various scales and with various amounts of head and camera movements. Some examples are shown in Figure 4.

Fig. 3. Sample frames from selected test sequences.

4.2. Prediction Performance

To evaluate the improvement of the prediction performance due to face segmentation, we compare the predictions of the regular full-frame metrics with those of the segmentation-supported metrics for the different data sets.

The results of the evaluation for our three data sets are shown in Figure 5. Segmentation generally leads to a better agreement between the metric's predictions and the subjective ratings. Some caution must be used when interpreting these results as some of the differences between correlations are not very significant. However, the trend is the same for all three data sets, which indicates that face segmentation is useful for augmenting the predictions of quality metrics. The fact that giving lower weights to the faces from the analysis generally leads to a a reduced prediction performance also supports this conclusion. As expected, the improvement is most noticeable for the scenes where faces cover a substantial part of the frame. Segmentation is least beneficial for sequences in which the faces are quite small and the distortions in the background introduced by some test con-

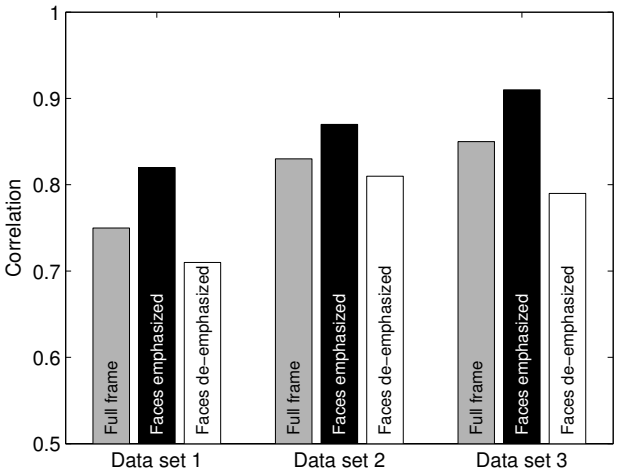

Fig. 4. Prediction performance with and without segmentation. Correlations are shown for the metrics applied uniformly across the full frame (gray bars), with an emphasis on the areas resulting from face segmentation (black bars), and the complementary emphasis (white bars).

ditions are more annoying to viewers than in other regions (as is the case with data set 2).

5. CONCLUSIONS

We presented a full-reference and a no-reference perceptual quality metric that account for visual attention using semantic segmentation. The advantages of segmentation support were demonstrated with test sequences showing human faces, resulting in better agreement of the predictions of our perceptual quality metrics with subjective ratings.

Obviously, face segmentation alone is not sufficient for improving the accuracy of metric predictions in all cases, but the results show that it is an important aspect. Our current research aims to generalize the proposed segmentation-driven quality metrics to detect more features and objects of interest [17] and to include object tracking [18].

6. REFERENCES

[1] S. Winkler, "Video quality metrics – A review." to appear in H. R. Wu, K. R. Rao (eds.), Digital Video Image Quality and Perceptual Coding, Marcel Dekker, 2004.

[2] A. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[3] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1005, 1988.

[4] L. B. Stelmach and W. J. Tam, "Processing image sequences based on eye movements," in *Proc. SPIE*, vol. 2179, pp. 90–98, San Jose, CA, 1994.

[5] A. Maeder, J. Diederich, E. Niebur, "Limiting human perception for image sequences," in *Proc. SPIE*, vol. 2657, pp. 330–337, San Jose, CA, 1996.

[6] A. E. Savakis, S. P. Etz, A. C. Loui, "Evaluation of image appeal in consumer photography," in *Proc. SPIE*, vol. 3959, pp. 111–120, San Jose, CA, 2000.

[7] M. P. Eckert, G. Buchsbaum, "The significance of eye movements and image acceleration for coding television image sequences," in *Digital Images and Human Vision*, A. B. Watson (ed.), pp. 89–98, MIT Press, 1993.

[8] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Proc. SPIE*, vol. 3299, pp. 180–191, San Jose, CA, 1998.

[9] S. Winkler, "A perceptual distortion metric for digital color video," in *Proc. SPIE*, vol. 3644, pp. 175–184, San Jose, CA, 1999.

[10] S. Winkler, R. Campos, "Video quality evaluation for Internet streaming applications." in *Proc. SPIE*, vol. 5007, pages 104–115, Santa Clara, CA, Jan. 21–24, 2003.

[11] A. Cavallaro, T. Ebrahimi, "Accurate video object segmentation through change detection," in *Proc. Int. Conf. on Multimedia and Expo*, Lausanne, Switzerland, 2002.

[12] P. Viola, M. Jones, "Robust Real-time Object Detection," in *Proc. of the Second International Workshop on Statistical Learning and Computational Theories of Vision Modeling, Learning, Computing and Sampling*, Vancouver, Canada, 2002.

[13] L. Gu, D. Bone, "Skin color region detection in MPEG video sequences," in *Proc. Int. Conf. Image Analysis and Processing*, pp. 898–903, Venice, Italy, 1999.

[14] J. Yang, W. Lu, A. Waibel, "Skin-color modeling and adaptation," in *Proc. Asian Conf. Computer Vision*, vol. 2, pp. 687–694, Hong Kong, 1998.

[15] VQEG, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," Available at http://www.vqeg.org, 2000.

[16] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics." in *Proc. SPIE*, vol. 4299, pages 114–125, San Jose, CA, Jan. 21–26, 2001.

[17] W. Osberger, A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proc. SPIE*, vol. 4299, pp. 361–372, San Jose, CA, 2001.

[18] A. Cavallaro, O. Steiger, T. Ebrahimi, "Tracking video objects in cluttered background," to appear in *IEEE Trans. Circuits and Systems for Video Technology*, 2004.