THE ADDED VALUE OF VISUAL ATTENTION IN OBJECTIVE VIDEO QUALITY METRICS

Welington Y. L. Akamine and Mylène C. Q. Farias

Department of Electrical Engineering, University of Brasília, Brasília - DF, 70919-970, Brazil

ABSTRACT

A recent development in the area of image and video quality consists of trying to incorporate aspects of visual attention in the design of visual quality metrics, mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying. This research area is still in its infancy and results obtained by different groups are not yet conclusive. Besides, most works address the image quality problem, not focusing on how to incorporate visual attention into video signals. In this work, we investigate the benefits of incorporating subjective saliency maps in the design of full-reference video quality metrics. In particular, we compare the performance of four full-reference video quality metrics (SSIM, MS-SSIM, VQM, and MOVIE) with their modified versions, which had saliency maps incorporated to their algorithms. The addition of saliency maps improved the performance of all metrics tested. But, highest gains in performance were obtained for the spatial metrics (SSIM, MS-SSIM, and MOVIE-S metrics), i.e. for the metrics that only took into consideration spatial degradations.

1. INTRODUCTION

Objective visual quality metrics can be classified as *data metrics*, which measure the fidelity of the signal without considering its content, or *picture metrics*, which estimate quality considering the visual information contained in the data. Customarily, quality measurements in the area of image processing have been largely limited to a few data metrics, such as mean absolute error (MAE), mean square error (MSE), and peak signal-to-noise ratio (PSNR), supplemented by limited subjective evaluation. Although over the years data metrics have been widely criticized for not correlating well with perceived quality measurements, it has been shown that such metrics can predict subjective ratings with reasonable accuracy as long as the comparisons are made with the same content, the same technique, or the same type of distortions [1].

One of the major reasons why these simple metrics do not generally perform as desired is because they do not incorporate any human visual system (HVS) features in their computation. It has been discovered that, in the primary visual cortex of mammals, an image is not represented in the pixel domain, but in a rather different manner [2]. Unfortunately, the measurements produced by metrics like MSE or PSNR are simply based on a pixel to pixel comparison of the data, without considering what is the content and the relationships among pixels in an image (or frames).

In the past few years, a big effort in the scientific community has been devoted to the development of better image and video quality metrics that incorporate HVS features (i.e. picture metrics) and, therefore, correlate better with the human perception of quality [1][3]. Recent developments in the area of visual quality include trying to incorporate aspects of visual attention into the design of visual quality metrics [4], mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying [5, 6].

In a previous work [7], we investigated the benefits of incorporating objective saliency maps into three image quality metrics (SSIM, PSNR, and MSE). We compared the performance of the original quality metrics with the performance of quality metrics that incorporate *subjective* saliency maps and saliency maps generated by three different visual attention models (Itti [8], GAFFE [9], and Acchanta [10]). Also, we studied the effects that different types of degradations (jpeg or jpeg2k compression, Gaussian our white noise, and fast fading) have on saliency maps and, consequently, on the performance of the final metric. Our results show that visual attention was able to improve the performance of the image quality metrics tested. The computational model that presented the best performance was GAFFE with gains slightly lower than the subjective saliency maps. The improvement in performance was higher for the simpler metrics (PSNR and MSE) than for the more complex metric (SSIM).

In this work, we investigate the benefit of incorporating subjective saliency maps into full-reference (FR) *video*

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and by University of Brasília ProIC/DPP.

quality metrics. With this purpose, we compare the performance of the original video quality metrics with the performance of the corresponding quality metrics with incorporation of *subjective* saliency maps into their design.

2. VISUAL ATTENTION

When observing a scene, the human eye typically filters the large amount of visual information available on the scene and attend to selected areas [8]. Oculo-motor mechanisms allow the gaze of attention to either hold on a particular location (fixation) or to shift to another location when sufficient information has already been collected (saccades). The selection of *fixations* is based on the visual properties of the scene. Priority is given to areas with a high concentration of information, minimizing the amount of data to be processed by the brain while maximizing the quality of the collected information.

Visual attention is, therefore, a feature of the human visual system that has the goal of reducing the complexity of scene analysis. It can be divided in two mechanisms that, combined, define which areas of the scene are to be considered relevant and, therefore, should be attended. These two mechanisms are known as *bottom-up* and *top-down* attention selection [11]. The bottom-up mechanism is an automated selection that is controlled mostly by the signal, independent of the task being performed. It is fast and short lasting, being performed as a response to low-level features that are perceived as visually salient, standing out from the background of the scene. The top-down mechanism is controlled by higher cognitive factors and external influences, such as semantic information, viewing task, personal preferences, and context. It is slower than bottom-up attention, requiring a voluntary effort.

The analysis of how humans perceive scenes can be performed by tracking eye movements in a subjective experiment using an eye-tracker equipment. From this type of experiment, gaze patterns are collected and later post-processed to generate *saliency maps*. The saliency maps obtained from these experiments are considered ground truths of human visual attention. In a very recent work, Engelke *et al.* [12] compared saliency maps gathered from three independently conducted eye tracking experiments. The comparison showed that the maps are very similar and the small differences found have minor impact on the applications.

In this work, we use a public database created by the Institut de Recherche en Communications at Cybernétique de Nantes/ Images et Video Communications (IRCCyN/IVC). The chosen database is the IRCCyN/IVC Eyetracker SD 2009_12 Database [13], which contains eyetracker data and the associated videos with various contents. The eye tracking information was gathered from a subjective experiment in which observers performed a quality scoring task. So, as well as the eyetracking data, the database also contains subjective ratings (quality scores) from 30 observers. Twenty standard definition original videos (720×576, interlaced, 50 Hz) were used in the experiment. Sample frames of 8 of the 20 originals of the database are depicted in Figure 1. The videos were coded with H.264 (JM coder version 16.1) and, then transmission errors were inserted. The bit-rates are selected to have a good quality if no transmission errors are present. The transmission errors were varied in spatial position and duration. There are 5 test coditions in the database, the reference plus 4 simulations of transmission errors, what resulted in $20 \times 5 = 100$ test sequences.

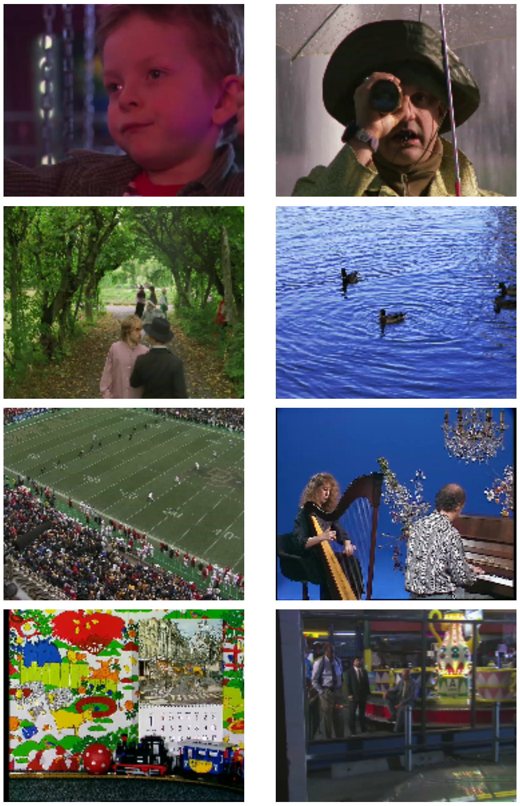

Fig. 1. Sample frames of 8 of the original videos in the IRCCyN/IVC Eyetracker SD 2009_12 Database [13].

Our goal in this paper is to incorporate the subjective saliency maps from this database into the process of estimation the quality of videos, which may contain common transmission and compression artifacts. The use of subjective saliency maps helps to understand if attention information can improve the performance of objective video quality metrics. In the next two sections, we describe the four full-reference quality metrics and how they are modified in order to take into account the information provided by the subjective saliency maps.

3. FR VIDEO QUALITY METRICS

Among the available quality metrics, we have selected four quality metrics for our tests: the Video Quality Metric (VQM) [14] , the MOtion-based Video Integrity Evaluation (MOVIE) [15], the Structural Similarity Index (SSIM) [16], and the Multi-Scale Structural Similarity Index (MS-SSIM) [17]. In this section, we briefly describe these metrics.

3.1. VQM

The video quality metric (VQM) is a metric proposed by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA) [14]. This metric has recently been adopted by ANSI as a standard for objective video quality. In VQEG Phase II (VQEG, 2003), VQM presented a very good correlation with subjective scores, showing one of the best performances among the competitors.

The algorithm used by VQM includes measurements for the perceptual effects of several video impairments, such as blurring, jerky/unnatural motion, global noise, block distortion, and color distortion. These measurements are combined into a single metric that gives a prediction of the overall quality.

The VQM algorithm can be divided into the following stages:

- Calibration – Estimates and corrects the spatial and temporal shifts, as well as the contrast and brightness offsets of the processed video sequence with respect to the original video sequence.

- Extraction of quality features – The set of quality features that characterizes perceptual changes in the spatial, temporal, and chrominance domains are extracted from spatial-temporal sub-regions of the video sequence. For this, a perceptual filter is applied to the video to enhance a particular type of property, such as edge information. Features are extracted from spatio-temporal (ST) subregions using a mathematical function and, then, a visibility threshold is applied to these features.

- Estimation of quality parameters – A set of quality parameters that describes the perceptual changes is calculated by comparing features extracted from the processed video with those extracted from the reference video.

- Quality estimation – The final step consists of calculating an overall quality metric using a linear combination of parameters calculated in previous stages.

3.2. MOVIE

The MOVIE metric was proposed by the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin [15]. It also has a good performance, but is a high complexity algorithm. The MOVIE metric generates three quality estimates: a global quality estimate (MOVIE), a spatial quality estimate (MOVIE-S), and a temporal quality estimate (MOVIE-T). To generate MOVIE-S, the algorithm uses Gabor filters and measures the degradations in each video frame separately. To generate MOVIE-T, on the other hand, the algorithm takes into consideration temporal degradations/features affecting the video quality. MOVIE-S and MOVIE-T are combined in order to obtain the overall estimate MOVIE, as shown in the block diagram depicted in Figure 2.

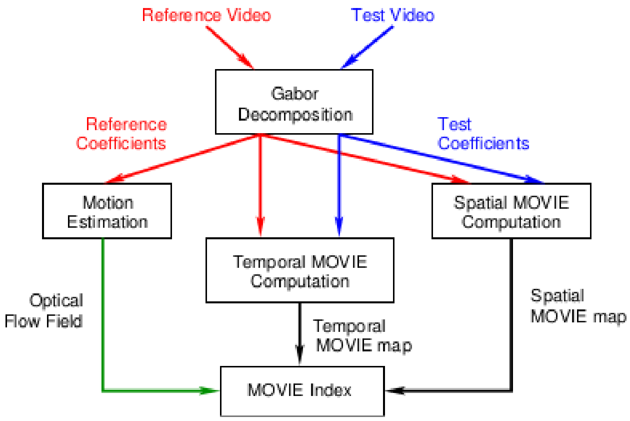

Fig. 2. Block Diagram of the MOVIE algorithm taken from [15].

3.3. SSIM

The SSIM is a very popular metric. It calculates the quality of an image using three features: luminance, contrast, and structure. These features are calculated using the following equations:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{1}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{2}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{3}$$

where C_1, C_2, C_3 are fixed constants, μ_x and μ_y are the average of the original and test image, σ_x and σ_y are the standard deviation of the original and test image, and σ_{xy} is the covariance between x and y.

The quality estimate of a test image y, in relation to its original x, is given by:

$$\text{SSIM}(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma}, \tag{4}$$

where α, β e γ are paramenters that define the importance of each feature. Generally, to simplify, $\alpha = \beta = \gamma = 1$ is used. For video signals, the quality estimate is given by the average value of SSIM(x, y) over all frames.

3.4. M-SSIM

MS-SSIM is a variation of the SSIM metric. The algorithm iteratively applies a low-pass filter to the image and downsamples the filtered image by a factor of two. The original image corresponds to Scale 1 and the (M-1)-th (last) iteration to Scale M. At all scales, the contrast feature (c, eq. 2) and the structure feature (s, eq. 3) of SSIM are calculated. The luminance feature (l, eq. 1) is only calculated for Scale M.

The MS-SSIM estimate for the quality of an image y, in relation to its original x, is given by the following equation:

$$\text{MS-SSIM}(x, y) = [l(x, y)]_M^\alpha \prod_{j=1}^{M} .[c(x, y)]_j^\beta . [s(x, y)]_j^\gamma . \tag{5}$$

For video signals, the quality estimate is given by the average value of SSIM(x, y) over all frames.

4. INCORPORATION OF SALIENCY INTO QUALITY METRICS

The visual attention integration process consists of using the gray-scale pixel values of the subjective saliency maps as *weights* for the error maps generated by the quality metrics. The modified saliency-based quality metrics for the corresponding FR metrics are given by the folowing expression:

$$\text{MET-AV} = \frac{\sum_{x=1}^{L} \sum_{y=1}^{C} \text{MET}(x, y) \cdot SAL(x, y)}{\sum_{x=1}^{L} \sum_{y=1}^{C} SAL(x, y)}, \tag{6}$$

where $SAL(x, y)$ is the saliency map pixel and MET(x, y) is the error map pixel calculated using the FR quality metric. This particular integration process is used because it is the simplest solution that allows the same model to be used for all metrics [18].

This integration approach also makes it easier to compare the performance of different metrics. For the quality metrics SSIM and MS-SSIM, the integration consists simply of using the error difference map generated by these metrics in the place of MET(x, y) in eq.6. On the other hand, for MOVIE and VQM some adaptation is necessary.

For the metric MOVIE, besides of incorporating the saliency maps using the final error difference map (MOVIE-AV), we also independently incorporate it to the spatial difference map (MOVIE-S-AV) and to the temporal difference map (MOVIE-T-AV). In other words, we consider the intermediates estimates MOVIE-S and MOVIE-T as two other metrics and perform the incorporation of saliency maps for these two metrics.

Two different approaches are used to incorporate visual attention into the VQM metric. The first approach (VQM-C-AV) consists of multiplying the saliency map by the *comparison* map generated by the VQM algorithm. For that, the saliency map has to be divided by the exactly same number of regions than the comparison map. The values of each region in the saliency map is the average value of the saliency inside the region. Then, we use eq. 6, where we make MET the comparison map and SAL the saliency map divided in regions.

The second approach (VQM-A-AV) used for VQM consists of multiplying the saliency map by the "absolute value of temporal information" (ATI) given by the VQM algorithm. We chose this parameter because in the VQM metric ATI is used to give more importance to certain areas of the frame. In other words, ATI is used in the same way as the saliency map is used.

In Table 1, we present the Pearson and Spearman correlation coefficients for all the metrics tested in this work. As can be observed, with the addition of the saliency maps the performance of all metrics improved. It is interesting to notice that the highest improvements in performance corresponded to the SSIM, MS-SSIM, and MOVIE-S metrics. These particular metrics are the ones which only take into consideration the spatial information of the video.

Table 1. Pearson and Spearman correlations coefficients for the video quality metrics tested (SSIM, MS-SSIM, VQM, and MOVIE). The abreviation AV corresponds to the models with saliency maps integrated.

Quality Metric	**PCC**	**SCC**
SSIM	0.5345	0.6761
SSIM-AV	**0.7219**	**0.8199**
MS-SSIM	0.7182	0.7913
MS-SSIM-AV	**0.7403**	**0.9169**
VQM	0.5598	0.6838
VQM-C-AV	**0.5745**	**0.7009**
VQM-A-AV	**0.5604**	**0.6852**
MOVIE	0.5254	0.7307
MOVIE-AV	**0.5488**	**0.7450**
MOVIE-S	0.6528	0.7128
MOVIE-S-AV	**0.6715**	**0.7335**
MOVIE-T	0.6671	0.7172
MOVIE-T-AV	**0.6788**	**0.7263**

5. CONCLUSIONS

In this work, we investigated the benefits of incorporating subjective saliency maps in the design of full-reference video quality metrics. In particular, we compared the performance of four full-reference video quality metrics (SSIM, MS-SSIM, VQM, and MOVIE) with their modified versions, which had saliency maps incorporated to their algorithm. Results showed that the addition of saliency maps improved the performance of all quality metrics tested. But, highest gains in performance were obtained for the spatial metrics (SSIM, MS-SSIM, and MOVIE-S metrics), i.e. for the metrics that only took into consideration spatial degradations.

6. REFERENCES

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?" *IEEE Signal Processing Magazine*, vol. 1, no. January, pp. 98–117, 2009.

[2] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 50 –59, nov. 2011.

[3] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165 –182, june 2011.

[4] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 50 –59, nov. 2011.

[5] C. Oprea, I. Pirnog, C. Paleologu, and M. Udrea, "Perceptual Video Quality Assessment Based on Salient Region Detection," in *Telecommunications, 2009. AICT '09. Fifth Advanced International Conference on*, May 2009, pp. 232–236.

[6] J. Redi, H. Liu, P. Gastaldo, R. Zunino, and I. Heynderickx, "How to apply spatial saliency into objective metrics for jpeg compressed images?" in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, nov. 2009, pp. 961 –964.

[7] M. Farias and W. Akamine, "On performance of image quality metrics enhanced with visual attention computational models," *Electronics Letters*, vol. 48, no. 11, pp. 631–633, 2012.

[8] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[9] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science, A. Gasteratos, M. Vincze, and J. Tsotsos, Eds. Springer Berlin / Heidelberg, 2008, vol. 5008, pp. 66–75.

[10] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *Image Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 564 –573, april 2008.

[11] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology*, vol. 14, no. 19, pp. R850 – R852, 2004.

[12] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H. Zepernick, and A. Maeder, "Comparative study of fixation density maps," *Image Processing, IEEE Transactions on*, vol. 22, no. 3, pp. 1121–1133, 2013.

[13] P. L. C. Ulrich Engelke, Marcus Barkowsky and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, june 2010.

[14] M. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[15] K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *Image Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 335–350, 2010.

[16] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98 –117, jan. 2009.

[17] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[18] J. A. Redi, H. Liu, P. Gastaldo, R. Zunino, and I. Heynderickx, "How to apply spatial saliency into objective metrics for JPEG compressed images?" in *IEEE ICIP2009 International Conference on Image Processing*, nov 2009.