Journal of
Electronic Imaging

JElectronicImaging.org

Video quality assessment using visual attention computational models

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

SPIE. IS&T imaging.org

Video quality assessment using visual attention computational models

Welington Y. L. Akamine* **and Mylène C. Q. Farias***
University of Brasília (UnB), Department of Electrical Engineering, Campus Universitário Darcy Ribeiro, 70919-970 Brasília—DF, 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. Among the works that have reported some improvements, most use subjective saliency maps, i.e., saliency maps generated from eye-tracking data obtained experimentally. Other works address the image quality problem, not focusing on how to incorporate visual attention into video signals. We investigate the benefits of incorporating bottom-up video saliency maps (obtained using Itti's computational model) into video quality metrics. In particular, we compare the performance of four full-reference video quality metrics with their modified versions, which had saliency maps incorporated into the algorithm. Results show that the addition of video saliency maps improve the performance of most quality metrics tested, but the highest gains were obtained for the metrics that only took into consideration spatial degradations. *© 2014 SPIE and IS&T* [DOI: 10.1117/1.JEI.23.6.061107]

Keywords: video quality metrics; visual attention; quality assessment; artifacts.

Paper 14184SSP received Apr. 1, 2014; revised manuscript received Jul. 10, 2014; accepted for publication Aug. 13, 2014; published online Sep. 5, 2014.

1 Introduction

In modern digital imaging systems, the quality of the visual content can undergo a decrease due to impairments introduced during capture, transmission, storage, and/or display, as well as by any signal processing algorithm that may be applied to the content along the way (e.g., compression). The most accurate way to determine the quality of a video is by using psychophysical experiments with human subjects. Unfortunately, these experiments are very expensive, time-consuming, and hard to incorporate into a design process or an automatic quality of service control. Therefore, there is a great need for objective quality metrics, i.e., algorithms that can predict visual quality as perceived by human observers.

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. In the past, quality measurements in the area of image processing were largely limited to a few data metrics, such as mean absolute error, mean square error (MSE), and peak signal-to-noise ratio (PSNR), supplemented by limited subjective evaluation. Over the years, data metrics have been widely criticized for not correlating well with perceived quality measurements.[1]

One of the major reasons why data metrics do not generally perform as desired is because they simply perform a pixel to pixel comparison of the data, not considering the visual content characteristics. In the last decades, several image and video quality metrics that incorporate human visual system (HVS) features into their design have been proposed, achieving a better correlation with the human perception of quality.[1,2] Among the HVS features used in quality assessment, we can cite contrast sensitivity function, visual masking (luminance and pattern masking), multichannel modelling, and visual attention.[3,4]

Visual quality metrics that try to incorporate aspects of visual attention into their design use the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying.[5–9] Altough there are many works in this area, this research area is still in its infancy and results obtained by different groups are not yet conclusive, as pointed out by Engelke et al.[10] Some researchers have reported that the incorporation of saliency maps increases the performance of visual quality metrics,[11–21] while others have reported no or very little improvement.[22–25]

In a previous work,[26] we investigated the benefits of incorporating objective saliency maps into three image quality metrics [structural similarity index (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[27], a gaze-attentive fixation finding engine (GAFFE)[28], and Achanta[29]]. Also, we studied the effects that different types of degradations (jpeg or jpeg2k compression, Gaussian noise, 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 tested image quality metrics. The computational model that presented the best performance was GAFFE,[28] with performance gains slightly lower than those obtained with subjective saliency maps. The improvement in

*Address all correspondence to: Welington Y. L. Akamine and Mylène C.Q. Farias, E-mail: welingtonylakamine@gmail.com and mylene@unb.br

0091-3286/2014/$25.00 © 2014 SPIE and IS&T

performance was higher for the simpler metrics (PSNR and MSE) than for the more complex metric (SSIM).

According to previous studies,[10] if motion information is appropriately integrated, visual attention can provide a better measure of video quality metrics than what was obtained with image quality metrics. Therefore, in this work, we investigate the benefits of incorporating video saliency maps generated by Itti's spatiotemporal (ST) attention computational model[27] into full-reference (FR) video quality metrics. With this purpose, we compare the performance of original video quality metrics with the performance of their corresponding metrics with video saliency maps incorporated into their design. We evaluate the performance of the metrics using two video databases: the Laboratory for Image and Video Engineering (LIVE) video quality database,[30,31] which contains videos with common distortions, and the the Institut de Recherche en Communications at Cybernétique de Nantes/Images et Video Communications (IRCCyN/IVC) Eyetracker SD 2009_12 database,[32] which also contains eyetracker data.

This paper is divided as follows. In Sec. 2, we briefly describe the visual attention mechanisms and Itti's visual attention computational model for video signals. In Sec. 3, we describe the video quality metrics tested in this work. In Secs. 4 and 5, the saliency incorporation process is described and the results are presented. Finally, in Sec. 6, the conclusions are presented.

2 Visual Attention

When observing a scene, the human eye typically filters the large amount of visual information available on the scene and only attends (focuses on) to selected areas.[33,34] Oculomotor 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 HVS that has the goal of reducing the complexity of scene analysis. It can be divided into two mechanisms that, when 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 selections. The bottom-up mechanism is an automated selection that is mostly controlled 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 and 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, the 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 subjective experiments using eye-tracker equipment. From this type of experiment, gaze patterns are collected and later postprocessed to generate subjective saliency maps. The subjective saliency maps obtained from these experiments are considered ground truths of human visual attention. In a recent work, Engelke et al.[35] compared subjective saliency maps gathered from three independently conducted eye-tracking experiments. The comparison showed that the maps are very similar, and the small differences that were found have a minor impact on the applications.

Some works use subjective saliency maps to study which aspects of visual attention are considered relevant to the quality assessment of images and videos.[6,16,22,36–38] The work of Castelhano et al.[39] studied differences in eye movements for two different tasks: visual search and memorization. Their results show that, for images, the task influences eye movement measures, including the number of fixations and gaze duration. Liu and Heynderickx[16] compared subjective saliency maps for images collected during free-viewing and quality scoring tasks. Their results show that saliency maps are affected by the task. For video signals, Le Meur et al. performed a study to compare eye movements in free-viewing and quality scoring tasks.[22] Their results did not show significant differences between these maps. Due to motion suppression and attentional capture effects, differences in attention due to tasks may be more relevant in images than videos.[10] Studies also show that global distortions do not seem to affect subjective saliency maps very much,[40,41] while localized distortions (e.g., packet loss) change the maps considerably.[42]

There are several works in the literature that incorporate aspects of visual attention into the design of video and image quality metrics.[6,11–16,16–25,36,43,44] Some metrics use subjective saliency maps.[6,16,36] Liu and Heynderickx investigated the incorporation of subjective saliency maps into a set of image quality metrics.[16] Their results show improvements in performance for all metrics. But, the gains in performance varied according to the metric and image content. Le Meur et al. integrated subjective saliency maps into an original video quality metric.[22] Their results show no improvement in comparison to the original metric.

Although subjective saliency maps are considered as the ground truth in visual attention, they cannot be used in real-time applications. To incorporate visual attention aspects into the design of video quality metrics, we have to use some type of visual attention computational model. Several authors have proposed visual quality metrics that use different types of visual attention computational models.[11–25,43,44] Barland and Saadane integrated a saliency model into a blind image quality metric.[12] Cavallaro and Winkler obtained good results using a segmentation algorithm to separate faces and background in videos and a quality metric based on motion and color features.[18] You et al. also used semantic information (faces and text) and a bottom-up model to generate saliency maps for videos, which were later integrated in SSIM and PSNR.[25] Their results show improvements only for PSNR. A later work by You et al. takes into account motion and global quality, obtaining improvements in performance.[43]

Although results are not always consistent, as detailed by Engelke et al.,[35] the integration of visual attention aspects can improve the results of video quality metrics if ST visual attention aspects are correctly modeled and integrated. Since for videos the task effect seem to be smaller than for images,[35] one possible approach is to use a bottom-up computational visual attention model that takes into account the ST characteristics of the video signal. In this work, we incorporate attention into a set of video quality metrics using a bottom-up ST model developed by Itti.[27]

Itti's model analyzes five features from the video, as depicted in Fig. 1.[27] Three of the features are considered spatial features (intensity, contrast, and orientation), since only spatial information is used to compute these features. The intensity feature of a frame is represented by its luminance value. The contrast feature is given by the difference between the color components of the frame, in this case, blue/yellow and green/ red. The orientation feature is given by the direction of the edges of the frame. Itti's model uses four direction angles: 0, 45, 90, and 135 deg.

The other two features of Itti's model (flicker and motion) are considered temporal features, since temporal information from the video is used to compute them. The flicker feature calculates the difference between one frame and the next frame. The motion feature gives the direction of the objects in the scene. As depicted in Fig. 1, Itti's model combines the spatial and temporal features to obtain the saliency maps for each frame of the video. Figure 2 depicts sample frames of two videos and their corresponding saliency maps estimated using Itti's visual attention computational model. In the saliency maps, lighter areas correspond to more salient areas, while darker areas correspond to less salient areas.

3 Video Quality Metrics

For our tests, we selected four FR quality metrics of known good performance: the video quality metric (VQM),[45] the motion-based video integrity evaluation (MOVIE),[46] the SSIM,[1] and the multiscale structural similarity index (MS-SSIM).[47] In this section, we briefly describe each of these metrics.

3.1 *Structural Similarity Index*

SSIM is a very popular metric proposed by Wang and Bovik of LIVE at the University of Texas at Austin.[1] The algorithm used by SSIM estimates the quality of an image using three features: luminance (l), contrast (c), and structure (s). The quality estimate of a test image y, in relation to its original x, is given by

$$\mathrm{SSIM}(x,y) = [l(x,y)].[c(x,y)].[s(x,y)], \\ = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)(\sigma_{xy} + C_3)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)(\sigma_x\sigma_y + C_3)}, \tag{1}$$

where C_1, C_2, and C_3 are fixed constants, μ_x and μ_y are the average values of the pixels in the original and test images, σ_x and σ_y are the standard deviation values of the pixels in the original and test images, and σ_{xy} is the covariance of the pixel values in the original and test images.For video signals, we take the average value of the estimate given by $\mathrm{MS\text{-}SSIM}(x,y)$ for the video frames.

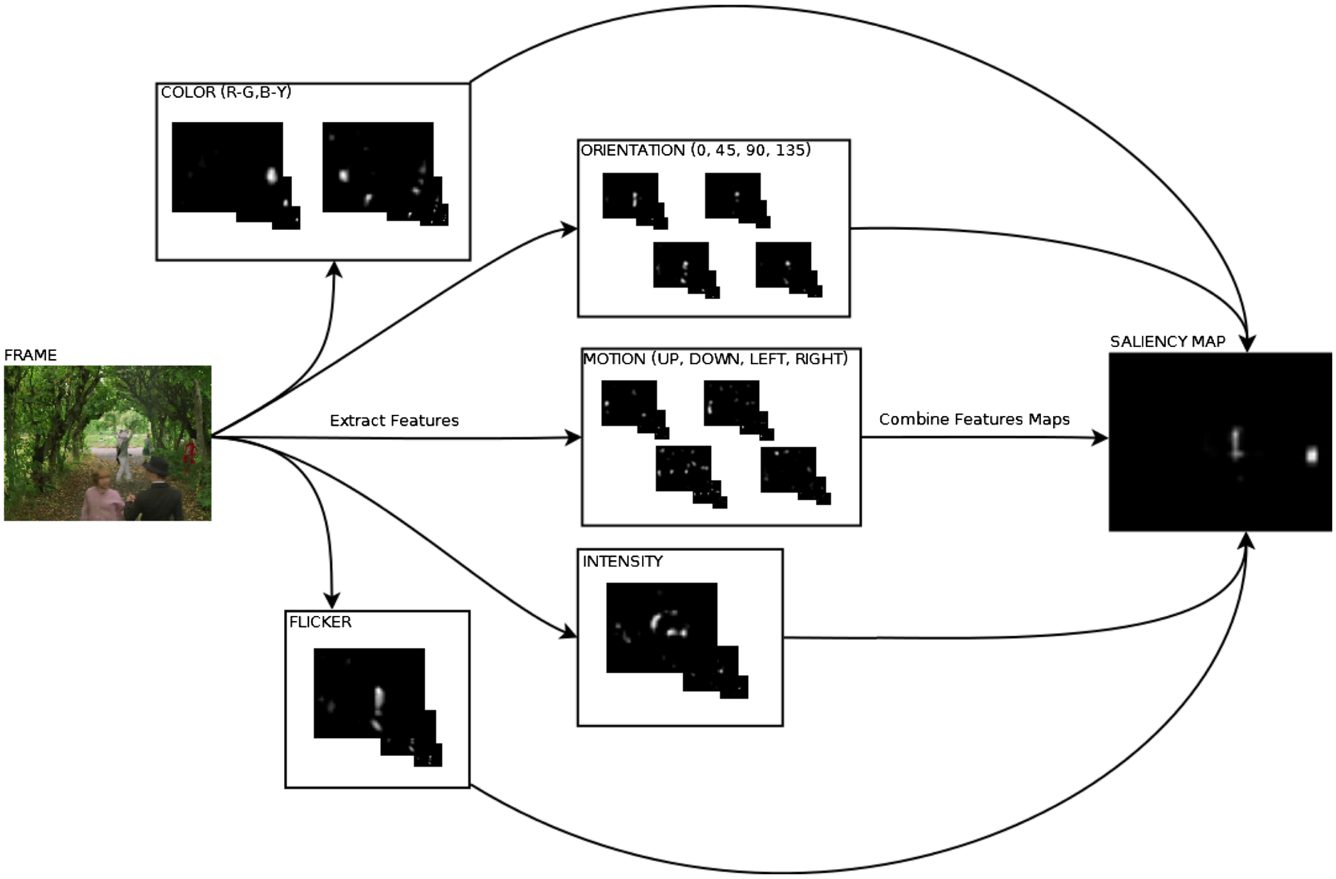
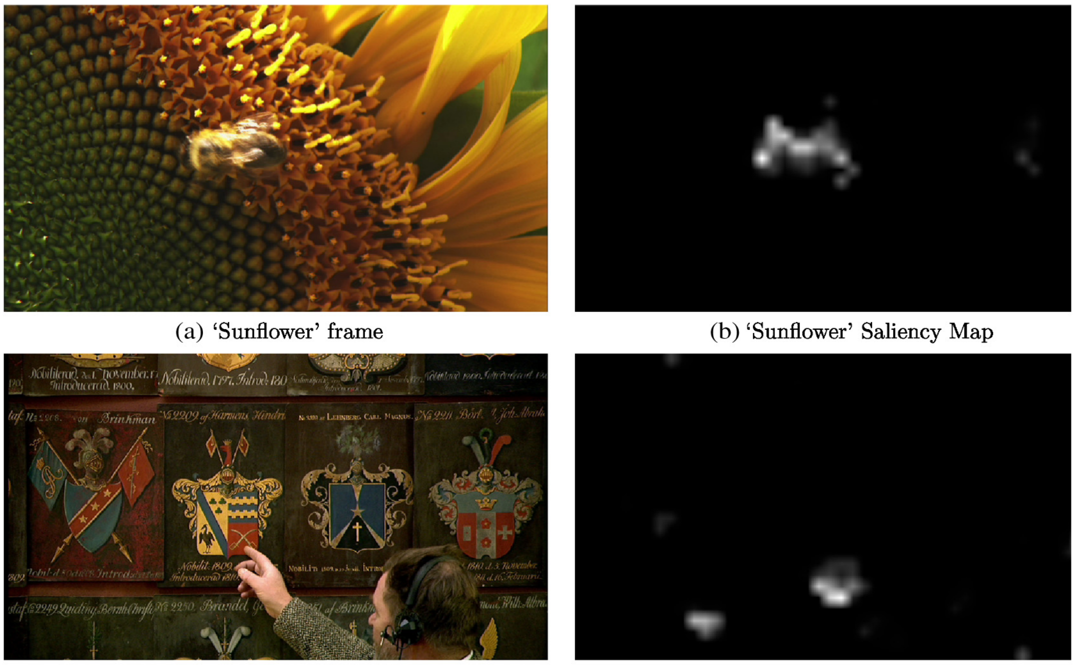


Fig. 1 Block diagram of Itti's visual attention computational model for video signals.[27]



(a) 'Sunflower' frame

(b) 'Sunflower' Saliency Map

(a) 'Shield' frame

(b) 'Shield' Saliency Map

Fig. 2 Sample frames of two videos (left) and their corresponding saliency maps (right) estimated using Itti's visual attention computational model for video signals.[27]

3.2 *Multiscale Structural Similarity Index*

MS-SSIM is a variation of the SSIM metric proposed by Wang et al.[47] The method provides more flexibility than single-scale SSIM in incorporating the variations of image resolution and viewing conditions. The MS-SSIM 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 iteration to scale M.

At all scales, the contrast feature (c) and the structure feature (s) of SSIM are calculated. The luminance feature (l) is only calculated for scale M. The MS-SSIM quality estimate of an image y, in relation to its original x, is given by the following equation:

$$\text{MS-SSIM}(x,y) = [l(x,y)]_M^{\alpha} \prod_{j=1}^{M} .[c(x,y)]_j^{\beta}.[s(x,y)]_j^{\gamma}. \tag{2}$$

For video signals, we take the average value of the estimate given by $\text{MS-SSIM}(x,y)$ for the video frames.

3.3 *VQM*

VQM is a metric proposed by Wolf and Pinson from the National Telecommunications and Information Administration.[45] In video quality experts group (VQEG) Phase II (VQEG, 2003), VQM presented a very good correlation with subjective scores, showing one of the best performances among the competitors. This metric has recently been adopted by ANSI as a standard for objective video quality.

The algorithm used by VQM includes measurements of the perceptual effects caused by several types of 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: It 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 subregions 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 ST subregions using a mathematical function, 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 the parameters calculated in previous stages.

3.4 *MOVIE*

The MOVIE metric was proposed by LIVE at the University of Texas at Austin.[46] It also has a good performance, but has a very high computational complexity. The MOVIE algorithm generates three quality estimates: a global quality estimate (MOVIE), a spatial quality estimate (MOVIE-S), and a temporal quality estimate (MOVIE-T).

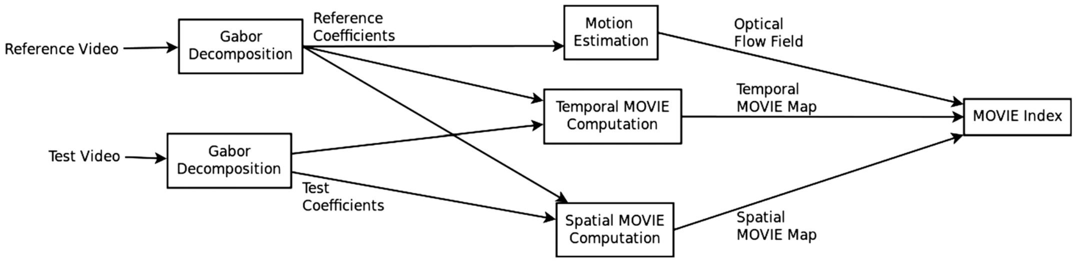


Fig. 3 Block diagram of the motion-based video integrity evaluation algorithm (adapted from Ref. 46).

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 and features affecting the video quality. To generate the overall estimatTo generate the overall MOVIE index, the algorithm combines MOVIE-S and MOVIE-T, as shown in the block diagram depicted in Fig. 3.

4 Incorporation of Saliency into Quality Metrics

The visual attention integration process consists of using the gray-scale pixel values of the saliency maps as weights for the error maps generated by the video quality metrics. The error maps give a perceptual measure of the physical errors (i.e., errors calculated by the quality metric) for the pixels of the video frame. In other words, the error maps depict the

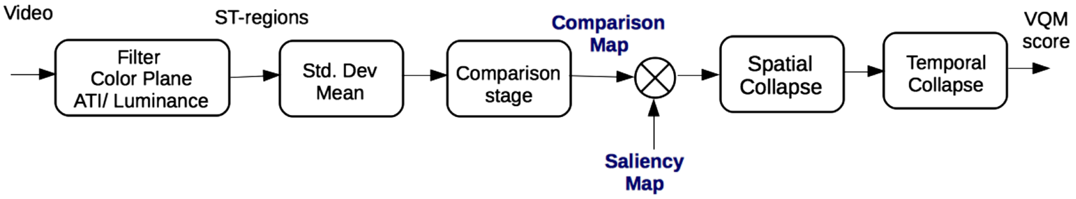


Fig. 4 Simplified block diagram of video quality metric (VQM), showing the incorporation of the saliency map with VQM comparison map to generate the modified metric VQM-C-VA.

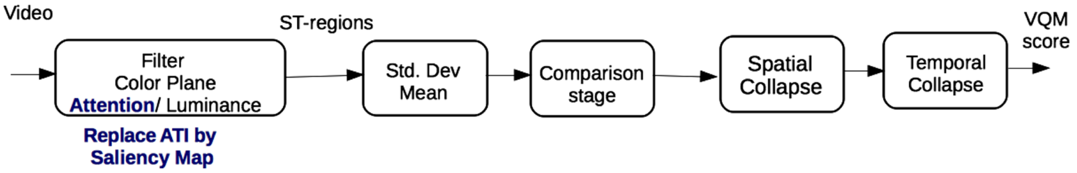


Fig. 5 Simplified block diagram of VQM, showing the incorporation of the saliency map by substitution of absolute value of temporal information by the saliency map: VQM-A-VA.

Fig. 6 Sample frames of 4 of the 10 original videos in the Laboratory for Image and Video Engineering video quality database.[30,31]

spatial distribution of the perceptual errors in each video frame. For SSIM and MS-SSIM, the error maps are obtained using Eqs. (1) and (2), respectively, for all pixels in the video frames.

For a given original metric (MET), the modified saliency-based quality metric (MET-VA) is given by the following expression:

$$\text{MET-VA} = \frac{\sum_{x=1}^{L}\sum_{y=1}^{C}\text{MET}(x,y)\cdot[1+\text{SAL}(x,y)]}{\sum_{x=1}^{L}\sum_{y=1}^{C}[1+\text{SAL}(x,y)]}, \tag{3}$$

where $\text{SAL}(x,y)$ is the saliency map pixel and $\text{MET}(x,y)$ is the error map pixel calculated using the original FR quality metric (unmodified). This integration process is used because it is the simplest solution that allows the same model to be used for all metrics,[6] making it easier to compare different metrics. Also, according to Redi et al.,[6] the weighting function in Eq. (3) provides the best performance gain since it weighs salient regions higher, but does not nullifiy the contribution of nonsalient regions.

For the quality metrics SSIM and MS-SSIM, the integration consists simply of using the error maps generated by these metrics [Eqs. (1) and (2), respectively] in place of $\text{MET}(x,y)$ in Eq. (3). For the metric MOVIE, besides incorporating the saliency map to the final error map (MOVIE-VA), we also independently incorporate it into the spatial error map (MOVIE-S-VA) and to the temporal error map (MOVIE-T-VA). 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 the error maps of these two metrics.

Since VQM does not generate error maps as do the previous metrics, some adaptation of the incorporation process is required. In this work, we propose two different approaches to incorporate visual attention into VQM. The first approach (VQM-C-VA) consists of multiplying the saliency map by the comparison map generated by the VQM algorithm, as shown in a simplified block diagram in Fig. 4. Given that the comparison stage divides the frame into several subregions, the saliency map is also divided into exactly the same number of subregions in order to make integration possible. After dividing the saliency maps into subregions, the values of each region in the new region-based saliency map are set as the average saliency value of the region. Then, we substitute MET for the comparison map and SAL for the region-based saliency map in Eq. (3).

The second approach (VQM-A-VA) used for VQM consists of replacing the absolute value of temporal information (ATI) of the VQM algorithm by the saliency map. ATI is generated in the first stage of the simplified block diagram shown in Fig. 4. In Fig. 5, the simplified block diagram of the second approach is shown, in which ATI is substituted for the saliency map. The feature ATI is chosen in this approach because, in the VQM algorithm, it 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.

5 Results

As mentioned earlier, we tested the saliency-based metrics using two databases: the LIVE video quality database[30,31] and the IRCCyN/IVC Eyetracker SD 2009_12 database.[32] The LIVE public database was created by LIVE and the Center for Perceptual Systems at the University of Texas at Austin. It contains a set of 150 distorted videos and corresponding subjective quality scores. There are four different types of distortions in this database: MPEG-2 compression, H.264 compression, and simulated transmission of H.264 compressed bitstreams through error-prone IP and wireless networks. Sample frames of the 10 originals of the database are depicted in Figs. 2 and 6.

In Table 1, we present the Pearson correlation coefficient (PCC) and the Spearman correlation coefficient (SCC) for all quality metrics (with and without integration of saliency maps) when compared to the subjective data of the LIVE

Table 1 Laboratory for Image and Video Engineering database: Pearson correlations coefficients (PCC) and Spearman correlation coefficients (SCC) for the tested video quality metrics [structural similarity index (SSIM), multiscale SSIM (MS-SSIM), video quality metric (VQM), and motion-based video integrity evaluation (MOVIE)] and their corresponding versions with incorporation of saliency maps. The abbreviation VA corresponds to the incorporation of Itti's saliency maps and RN corresponds to the the incorporation of random saliency maps.

Quality metric	PCC	SCC
SSIM	0.5437	0.5401
SSIM-VA	**0.6381**	**0.6792**
SSIM-RN	0.5317	0.5257
MS-SSIM	0.7084	0.7445
MS-SSIM-VA	0.7031	**0.7578**
MS-SSIM-RN	0.6944	0.7360
VQM	0.7297	0.7153
VQM-C-VA	0.7289	0.7138
VQM-A-VA	0.7297	**0.7156**
VQM-C-RN	0.7233	0.7100
MOVIE	0.7898	0.7893
MOVIE-VA	**0.7901**	0.7871
MOVIE-RN	0.5357	0.5808
MOVIE-S	0.7185	0.7077
MOVIE-S-VA	**0.7201**	**0.7173**
MOVIE-S-RN	0.5091	0.5282
MOVIE-T	0.8200	0.8006
MOVIE-T-VA	0.8199	**0.8011**
MOVIE-T-RN	0.5811	0.5943

Note: Bold values correspond to an improvement in performance, when compared with the original metric.

database. The abbreviation VA (visual attention) corresponds to the models with integrated saliency maps. To make sure that the differences in performance are not attained by chance, we also test the performance of models integrated with random saliency maps. These random maps were computed by picking five random (fixation) points in the frame and processing them with Gaussian filters. The abbreviation RN corresponds to the models with integrated random saliency maps.

As can be observed, the performance of most metrics improves with the addition of the saliency maps. All models with random maps perform worse than the original models. It is interesting to note that the highest improvements in performance correspond to SSIM-VA and MOVIE-S. These particular metrics are the ones that only take the spatial information of the video into consideration. The best improvement is obtained for SSIM-VA: a gain of 17.36% in PCC and 25.75% in SCC (when compared to the original SSIM).

For the metrics with temporal information (MOVIE and VQM), MOVIE-VA has the smallest improvement: a gain of 0.0036% in PCC and a loss of −0.28% in SCC (in comparison to the original MOVIE). VQM-A-VA has the biggest improvement: a gain of 0.042% in SCC (in comparison to the original VQM). MOVIE-T-VA is the metric with the best SCC: 0.8011 (an increase of 0.062% in comparison to MOVIE-T). These results are expected because MOVIE is the most complex metric of the tested metrics and, therefore, has less room for improvement.

The IRCCyN/IVC Eyetracker SD 2009_12 database[32] was created by IRCCyN/IVC. This database contains 20 standard definition original videos (720 × 576, interlaced, 50 Hz). Sample frames of 9 of the 20 originals of the database are depicted in Fig. 7. To generate the test sequences, the original videos are encoded with H.264 (JM coder version 16.1) and transmission errors are inserted. The chosen bit rates generate test sequences with good quality if no transmission errors are present. The transmission errors were varied in spatial position and duration. There are five test conditions in the database (reference + four values of transmission errors), which resulted in $20 \times 5 = 100$ test sequences. The database contains eyetracker data collected from a free-viewing experiment and subjective quality scores from 30 observers.

In Table 2, we present the Pearson correlation coefficients (SCC) and Spearman correlation coefficients (PCC) for all the metrics tested in the IRCCyN/IVC Eyetracker SD 2009_12 database. Since in this database subjective saliency maps of the originals are available, we also tested the incorporation of subjective saliency maps. To identify the metrics with incorporation of subjective saliency maps, we added the abbreviation VA-SUB (visual attention-subjective) to the initials of the metrics in Table 2.

As can be observed, with the incorporation of saliency maps (subjective and objective), the performance of all tested metrics improves. The only exception is MS-SSIM-VA, which has a PCC value of 0.7027, which is smaller than

Fig. 7 Sample frames of nine of the original videos in the IRCCyN/IVC Eyetracker SD 2009_12 database.[32]

Table 2 IRCCyN/IVC Eyetracker SD 2009_12 database: PCC and SCC for the tested video quality metrics (SSIM, MS-SSIM, VQM, and MOVIE) and their corresponding versions with incorporation of saliency maps. The abbreviation VA corresponds to the incorporation of Itti's saliency maps, VA-SUB corresponds to the incorporation of subjective saliency maps, and RN corresponds to the incorporation of random saliency maps.

Quality metric	PCC	SCC
SSIM	0.5345	0.6761
SSIM-VA-SUB	**0.7216**	**0.8302**
SSIM-VA	**0.6063**	**0.7537**
SSIM-RN	0.5275	0.6781
MS-SSIM	0.7182	0.7913
MS-SSIM-VA-SUB	**0.7295**	**0.9191**
MS-SSIM-VA	0.7027	**0.8396**
MS-SSIM-RN	0.6682	0.7731
VQM	0.5598	0.6838
VQM-C-VA-SUB	**0.6063**	**0.7395**
VQM-C-VA	**0.5744**	**0.6949**
VQM-A-VA-SUB	**0.5599**	**0.6839**
VQM-A-VA	**0.5744**	**0.6839**
VQM-C-RN	**0.5648**	**0.6904**
MOVIE	0.5254	0.7307
MOVIE-VA-SUB	**0.5814**	**0.7690**
MOVIE-VA	**0.5285**	**0.7377**
MOVIE-RN	**0.5255**	0.7307
MOVIE-S	0.6528	0.7128
MOVIE-S-VA-SUB	**0.7045**	**0.7654**
MOVIE-S-VA	**0.6560**	**0.7235**
MOVIE-S-RN	0.6528	**0.7182**
MOVIE-T	0.6671	0.7172
MOVIE-T-VA-SUB	**0.6973**	**0.7532**
MOVIE-T-VA	**0.6694**	**0.7210**
MOVIE-T-RN	**0.6673**	0.7172

Note: Bold values correspond to an improvement in performance, when compared with the original metric.

the PCC of 0.7182 for MS-SSIM. Similar to what happened for the LIVE database, almost all models with random maps perform similar to or worse than the original models. The exception is VQM for which the random saliency maps present the best performance. The highest improvements in performance correspond to the spatial metrics: SSIM-VA, SSIM-VA-SUB, MS-SSIM-VA, MS-SSIM-VA-SUB, MOVIE-S-VA, and MOVIE-S-VA-SUB. Among all metrics, the best improvements are obtained for SSIM-VA-SUB and SSIM-VA. The increase in PCC is 35.06% for subjective maps and 13.43% for Itti's objective model, while the increase in SCC is 21.26% for subjective maps and 11.48% for Itti's objective model.

It is worth pointing out that, as expected, the increase in performance due to incorporation of subjective maps is higher than that for the objective maps obtained using Itti's computational model. This result is different from what was obtained earlier by Le Meur et al., who did not find significant performance improvements by incorporating subjective saliency maps into a video quality metric.[22] On the other hand, the metric used in their work considers both temporal and spatial distortions.[48] As shown in this work, performance is higher for simpler metrics. Given these results, we believe there is still room for performance improvement by choosing a computational attention model that generates saliency maps that are more similar to subjective maps.

We can notice by comparing the results in Tables 1 and 2 that the performance of metrics with incorporation of saliency maps is better for the IRCCyN/IVC Eyetracker SD 2009_12 database than for the LIVE database. One possible reason for this is the fact that most videos in the IRCCyN/IVC database have a clear attention focus, as opposed to the videos in the LIVE database. As pointed out by Le Meur et al.,[22] the absence of a clear attention focus causes variability between observers' saliency maps. And, as shown by Liu et al., the smaller the variation in saliency maps, the larger is the performance gain achieved by integrating saliency maps into objective metrics.[37]

6 Conclusions

In this work, we investigated the benefits of incorporating subjective saliency maps in the design of FR video quality metrics. In particular, we compared the performance of four FR 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 most quality metrics tested. But the highest gains in performance were obtained for the spatial metrics (SSIM and MOVIE-S metrics), i.e., for the metrics that only took spatial degradations into consideration. As expected, the increase in performance due to the incorporation of subjective maps is higher than that for objective saliency maps obtained using Itti's computational model.

Acknowledgments

This work was supported in part by Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq) Brazil and in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Brazil.

References

1. Z. Wang and A. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.* **26**(1), 98–117 (2009).
2. S. Chikkerur et al., "Objective video quality assessment methods: a classification, review, and performance comparison," *IEEE Trans. Broadcast.* **57**(2), 165–182 (2011).
3. D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Process.* **2013**, 905685 (2013).
4. B. A. Wandell, *Foundations of Vision*, Sinauer Associates Inc., Sunderland, MA (1995).
5. C. Oprea et al., "Perceptual video quality assessment based on salient region detection," in *Fifth Advanced Int. Conf. on Telecommunications*, pp. 232–236, IEEE (2009).
6. J. Redi et al., "How to apply spatial saliency into objective metrics for jpeg compressed images?," in *16th IEEE Int. Conf. on Image Processing*, pp. 961–964, IEEE (2009).
7. A. Ninassi et al., "Which semi-local visual masking model for wavelet based image quality metric?," in *15th IEEE Int. Conf. on Image Processing*, pp. 1180–1183, IEEE 2008).
8. J. You, A. Perkis, and M. Gabbouj, "Improving image quality assessment with modeling visual attention," in *2nd European Workshop on Visual Information Processing*, pp. 177–182, IEEE (2010).
9. B. Patrick, L. Callet, and E. Niebur, "Visual attention and applications in multimedia technologies," *Proc. IEEE* **101**(9), 2058–2067 (2013).
10. U. Engelke et al., "Visual attention in quality assessment: theory, advances, and challenges," *IEEE Signal Process. Mag.* **28**(6), 50–59 (2011).
11. W. Osberger, A. Maeder, and N. Bergmann, "A technique for image quality assessment based on a human visual system model," in *Proc. European Signal Processing Conf.*, Vol. 2, pp. 1049–1052, Eurasip (1998).
12. R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for jpeg-20000 compressed images," in *IEEE Int. Conf. on Image Processing*, pp. 2941–2944, IEEE (2006).
13. N. G. Sadaka et al., "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in *15th IEEE Int. Conf. on Image Processing*, pp. 369–372, IEEE (2008).
14. A. Moorthy and A. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.* **3**(2), 193–201 (2009).
15. I. Gkioulekas, G. Evangelopoulos, and P. Maragos, "Spatial Bayesian surprise for image saliency and quality assessment," in *17th IEEE Int. Conf. on Image Processing*, pp. 1081–1084, IEEE (2010).
16. H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.* **21**(7), 971–982 (2011).
17. U. Engelke and H.-J. Zepernick, "Framework for optimal region of interest-based quality assessment in wireless imaging," *J. Electron. Imaging* **19**(1), 011005 (2010).
18. A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Int. Conf. on Image Processing*, Vol. 5, pp. 3543–3546, IEEE (2004).
19. Z. Lu et al., "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.* **14**(11), 1928–1942 (2005).
20. Q. Ma, L. Zhang, and B. Wang, "New strategy for image and video quality assessment," *J. Electron. Imaging* **19**(1), 011019 (2010).
21. X. Feng et al., "Saliency inspired full-reference quality metrics for packet-loss-impaired video," *IEEE Trans. Broadcast.* **57**, 81–88 (2011).
22. O. Le Meur et al., "Overt visual attention for free-viewing and quality assessment tasks: impact of the regions of interest on a video quality metric," *Signal Process.: Image Commun.* **25**(7), 547–558 (2010).
23. A. Ninassi et al., "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *IEEE Int. Conf. on Image Processing*, Vol. 2, pp. II-169–II-172, IEEE (2007).
24. E. C. Larson, C. Vu, and D. M. Chandler, "Can visual fixation patterns improve image fidelity assessment?," in *15th IEEE Int. Conf. on Image Processing*, pp. 2572–2575, IEEE (2008).
25. J. You et al., "Perceptual quality assessment based on visual attention analysis," in *Proc. of the 17th ACM Int. Conf. on Multimedia*, pp. 561–564, ACM (2009).
26. M. Farias and W. Akamine, "On performance of image quality metrics enhanced with visual attention computational models," *Electron. Lett.* **48**(11), 631–633 (2012).
27. L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.* **13**(10), 1304–1318 (2004).
28. U. Rajashekar et al., "Gaffe: a gaze-attentive fixation finding engine," *IEEE Trans. Image Process.* **17**(4), 564–573 (2008).
29. R. Achanta et al., *Salient Region Detection and Segmentation, International Conference on Computer Vision Systems (ICVS '08)* Springer Lecture Notes in Computer Science, Vol. 5008, pp. 66–75, Springer (2008).
30. K. Seshadrinathan et al., "A subjective study to evaluate video quality assessment algorithms," *Proc. SPIE* **7527**, 75270H (2010).
31. K. Seshadrinathan et al., "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.* **19**(6), 1427–1441 (2010).
32. U. Engelke et al., "Modelling saliency awareness for objective video quality assessment," in *Int. Workshop on Quality of Multimedia Experience*, pp. 212–217, IEEE (2010).
33. L. Itti and C. Koch, "Computational modelling of visual attention," *Nat. Rev. Neurosci.* **2**(3), 194–203 (2001).
34. T. Judd et al., "Learning to predict where humans look," in *IEEE Int. Conf. on Computer Vision*, pp. 2106–2113, IEEE (2009).
35. U. Engelke et al., "Comparative study of fixation density maps," *IEEE Trans. Image Process.* **22**(3), 1121–1133 (2013).
36. H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *16th IEEE Int. Conf. on Image Processing*, pp. 3097–3100, IEEE (2009).
37. H. Liu et al., "How does image content affect the added value of visual attention in objective image quality assessment?," *IEEE Signal Process. Lett.* **20**(4), 355–358 (2013).
38. A. Ninassi et al., "Task impact on the visual attention in subjective image quality assessment," in *Proc. of the 14th European Signal Processing Conf. (Eurasip Eusipco)*, Eurasip (2006).
39. M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *J. Vis.* **9**(3), 6 (2009).
40. O. Le Meur et al., "Do video coding impairments disturb the visual attention deployment?," *Signal Process.: Image Commun.* **25**(8), 597–609 (2010).
41. J. You et al., "Balancing attended and global stimuli in perceived video quality assessment," *IEEE Trans. Multimed.* **13**(6), 1269–1285 (2011).
42. U. Engelke, "Modelling perceptual quality and visual saliency for image and video communications," PhD Thesis, Blekinge Institute of Technology, Sweden (2010).
43. J. You, J. Korhonen, and A. Perkis, "Attention modeling for video quality assessment: balancing global quality and local quality," in *IEEE Int. Conf. on Multimedia and Expo*, pp. 914–919, IEEE (2010).
44. J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.* **23**(1), 200–213 (2014).
45. M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.* **50**(3), 312–322 (2004).
46. K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.* **19**(2), 335–350 (2010).
47. Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Conf. Record of the Thirty-Seventh Asilomar Conf. on Signals, Systems and Computers*, Vol. 2, pp. 1398–1402, IEEE (2003).
48. A. Ninassi et al., "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.* **3**(2), 253–265 (2009).

Welington Y. L. Akamine is a MSc student in electrical engineering at the University of Brasília. He received his BSc degree in computer engineering from the University of Brasília in 2014. His current research interests include visual attention, quality metrics, and image processing.

Mylène C. Q. Farias received her BSc in electrical engineering from the Universidade Federal de Pernambuco, Brazil, in 1995, her MSc in electrical engineering from the Universidade Estadual de Campinas, Brazil, in 1998, and her PhD in electrical engineering from the University of California Santa Barbara, in 2004. He worked at CPqD, Brazil, Philips Research Laboratories, The Netherlands, and Intel Corporation, Phoenix. Currently, she is a professor of electrical engineering at the University of Brasília.