Incorporating Visual Attention Models into Video Quality Metrics

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

Department of Electrical Engineering,
University of Brasília (UnB),

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 improvement, most use subjective saliency maps, i.e. saliency maps generated from eye-tracking data obtained experimentally. 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 saliency maps obtained with visual attention. In particular, we compare the performance of four full-reference video quality metrics with their modified versions, which had saliency maps incorporated to the algorithm. For comparison proposes, we have used a database of subjective salience maps.

Keywords: Video quality metrics, Visual attention

1. INTRODUCTION

In modern digital imaging systems, the quality of the visual content can undergo a drastic 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.). Impairments are defined as visible defects (flaws) and can be decomposed into a set of perceptual features called artifacts. Being able to detect artifacts and improve the quality of the visual content prior to its delivery to the user is therefore crucial to ensure a good quality of experience. The most accurate way to determine the quality of a video is by using psychophysical experiments with human subjects. Unfortunately, these 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. 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]

Image Quality and System Performance XI, edited by Sophie Triantaphillidou, Mohamed-Chaker Larabi,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9016, 90160O · © 2014 SPIE-IS&T
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2039780

Recent developments in the area of visual quality include trying to incorporate aspects of visual attention into the design of visual quality metrics,[2] mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying.[4,5] 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.[2] Some researchers have reported that the incorporation of saliency maps increases the performance of quality metrics, while others have reported no or very little improvement. Among the works that have reported some improvement, most use subjective saliency maps, i.e. saliency maps generated from eye-tracking data obtained experimentally.[6]

In a previous work,[7] we investigated the benefit 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, GAFFE, and Achanta). 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 saliency maps generated by visual attention models 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 video quality metrics with subjective saliency maps incorporated into their design. We then tested the metrics using the LIVE Video Quality Database[8,9] which contains videos with common distortions.

This paper is divided as follows. In Section 2, we briefly describe the visual attention models. In Section 3, we decribe the video quality metrics tested in this work. In Section 4, the saliency incorporation process is described and the results are presented and compared to the original results. Finally, in Section 5 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 attend to selected areas.[10] 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. 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 subjective experiments using eye-tracker equipments. 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.*[11] 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.

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 visual attention computational models to generate objective saliency maps. In the case of video signals, the number of available computational model is limited. In this work, we use a bottom-up model developed by Itti.[12]

Itt's model analyzes five features from the video, as depicted in Figure 1. Three of the features are considered spatial features (intensity, contrast and orientation), because 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 colors 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. This model uses four direction angles: 0^o, 45^o, $90^o$and 135^o.

The other two features of Itti's model are considered temporal features (flicker and motion), because temporal information from the video frames is used to compute these frames. In other words, more than one frame is necessary to compute these feature. The *flicker* feature gives the difference between one frame and the next frame. The *motion* feature is the direction of the objects in the scene. In Figure 2, sample frames of two videos and their corresponding saliency map estimated using Itti's saliency model are depicted. In the saliency map, the lighter areas correspond to the more salient areas, while the darker areas correspond to less salient areas.

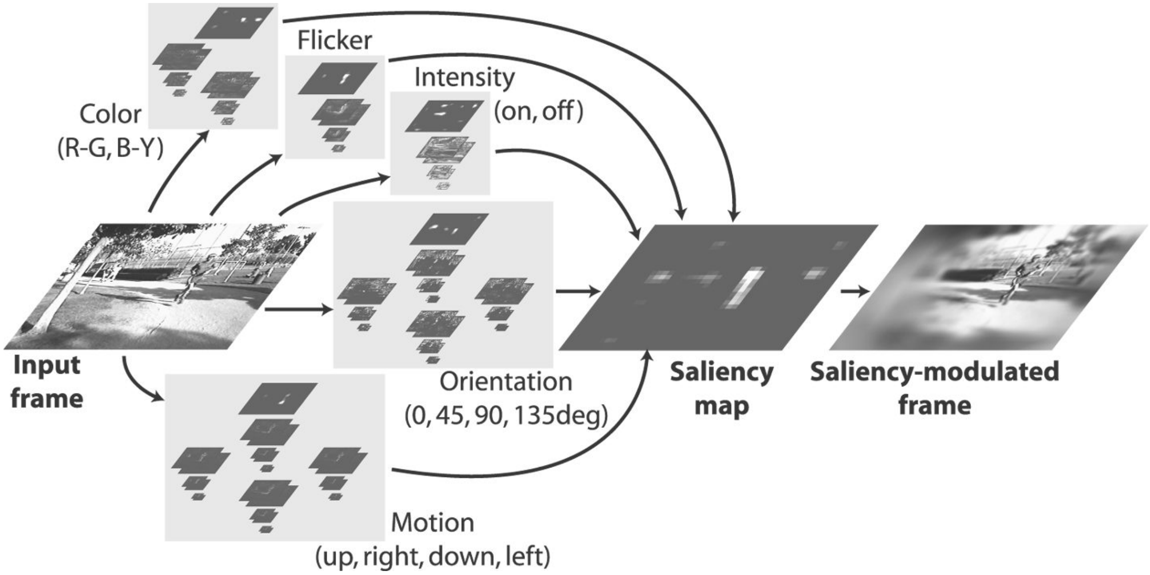
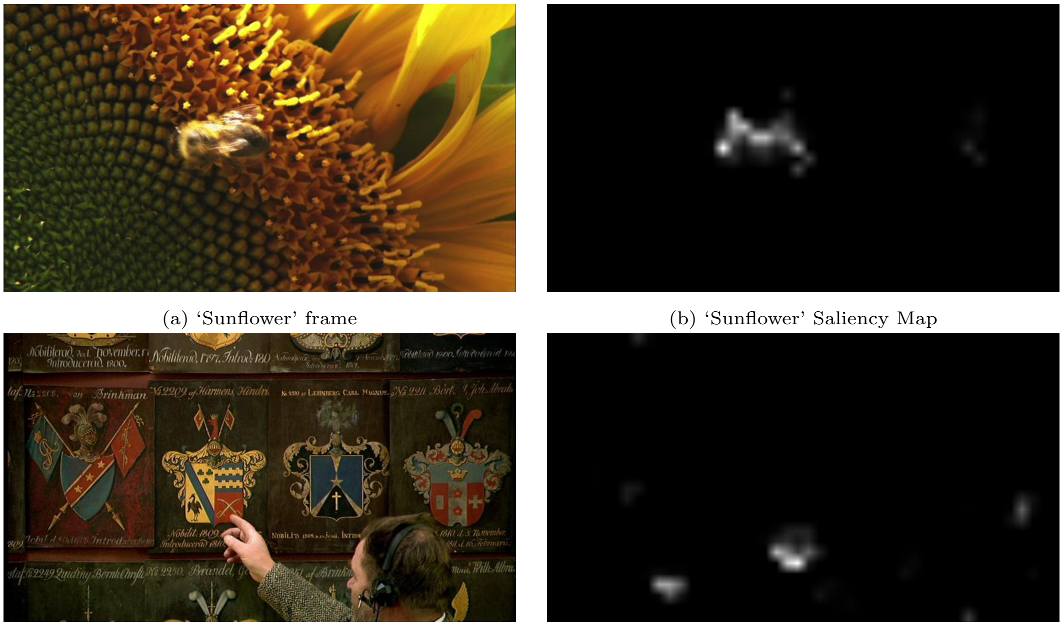


Figure 1. Block diagram of Itti's visual attention model for video signals (original from Itti[12]).

3. VIDEO QUALITY METRICS

There is an ongoing effort to develop video quality metrics that are able to detect impairments and estimate their annoyance, as perceived by human viewers.[3] Quality metrics can be classified according to the amount of reference (original) information used: Full Reference (FR), Reduced Reference (RR), and No-Reference (NR) metrics. On the FR approach the entire reference is available at the measurement point. On the RR approach, only part of the reference is available through an auxiliary channel. In this case, the information available at the measurement point generally consists of a set of features extracted from the reference. Finally, on the NR approach the quality estimation is obtained only from the test video.

To date, most of the achievements have been in the development of FR video quality metrics. FR Quality metrics that analyze visible differences between a test and a reference signal, taking into account aspects of the human visual system, usually have the best performance, but are often computationally expensive and hardly applicable in real-time contexts.[13,14] We have selected four FR quality metrics for our tests: the Video Quality Metric (VQM)[15] , the MOtion-based Video Integrity Evaluation (MOVIE),[16] the Structural Similarity Index (SSIM),[1] and the Multi-Scale Structural Similarity Index (MS-SSIM).[17] In this section, we briefly describe each of these metrics.



(a) ‘Sunflower’ frame

(b) ‘Sunflower’ Saliency Map

(a) ‘Shield’ frame

(b) ‘Shield’ Saliency Map

Figure 2. Sample frames of two videos (left) and their corresponding saliency map (right) estimated using Itti’s saliency model.

3.1 Structural Similarity Index (SSIM)

The Structural Similarity Index (SSIM) is a very popular metric proposed by Wang and Bovig of the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin.[1] The algorithm used by SSIM estimates the quality of an image using three features: luminance, contrast, and structure. For an original image x and a test image y, these features are calculated using the following equations:

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

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

and

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

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

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. The overall SSIM equation is given by

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{5}$$

For video signals, we down-sample the video signal by a factor of 8 and calculate the average value of SSIM(x,y) for the resulting set of frames.

3.2 Multi-Scale Structural Similarity Index(MS-SSIM)

Multi-Scale Structural Similarity Index (MS-SSIM) is a variation of the SSIM metric proposed by Wang *et al.*[17] The method provides more flexibility than single-scale SSIM in incorporating the variations of image resolution and viewing conditions. 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 (last) iteration to Scale M.

At all scales, the *contrast* feature (c in Eq. 2) and the *structure* feature (s in Eq. 3) of SSIM are calculated. The *luminance* feature ($l$in Eq. 1) 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:

$$\text{MS-SSIM}(x,y) = [l(x,y)]_M^{\alpha} \prod_{j=1}^{M} .[c(x,y)]_j^{\beta}.[s(x,y)]_j^{\gamma}. \tag{6}$$

For video signals, we down-sample the video signal by a factor of 8 and calculate the average value of the estimate given by MS-SSIM(x,y) for the resulting set of frames.

3.3 VQM

The video quality metric (VQM) is a metric proposed by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA).[15] 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.

3.4 MOVIE

The MOVIE metric was proposed by the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin.[16] 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).

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 estimate MOVIE, the algorithm combines MOVIE-S and MOVIE-T, as shown in the block diagram depicted in Figure 3.

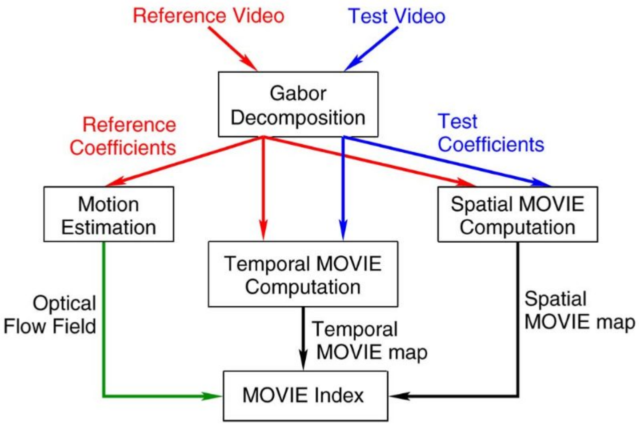


Figure 3. Block Diagram of the MOVIE algorithm taken (original taken from Seshadrinathan and Bovik[16]).

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 following expression:

$$\text{MET-AV} = \frac{\sum_{x=1}^{L} \sum_{y=1}^{C} \text{MET}(x,y) \cdot SAL(x,y)}{\sum_{x=1}^{L} \sum_{y=1}^{C} SAL(x,y)}, \tag{7}$$

where $SAL(x, y)$ is the saliency map pixel and $\text{MET}(x, y)$ 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 map pixel generated by these metrics in the place of $\text{MET}(x, y)$ in eq.7. On the other hand, for MOVIE and VQM some adaptation is required.

For the metric MOVIE, besides of incorporating the saliency maps using the final error map pixel (MOVIE-AV), we also independently incorporate it to the spatial error map pixel (MOVIE-S-AV) and to the temporal error map pixel (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 error map pixel of these two metrics.

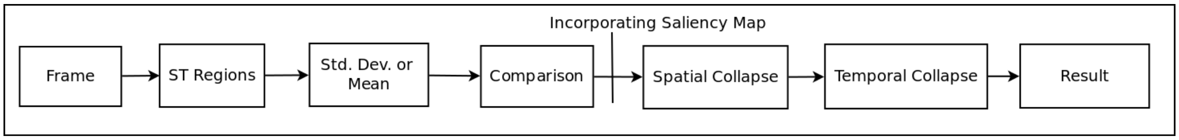


Figure 4. Block diagram of the 1st attention incorporation method for VQM: VQM-C-AV.

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, as shown in Figure 4. For that, the saliency map has to be divided by the exactly same number of regions than the comparison map generated by the VQM algorithm. The values of each region in the saliency map is set as the average value of the saliency of the region. Then, we use Eq. 7, where we make MET the comparison map and SAL the saliency map divided in regions.

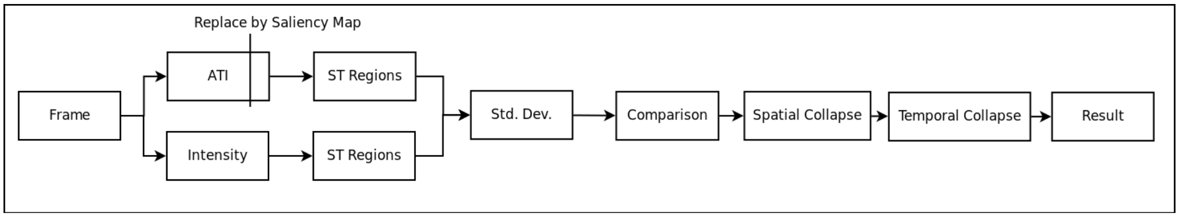


Figure 5. Block diagram of the 2nd attention incorporation method for VQM: VQM-A-AV.

The second approach (VQM-A-AV) used for VQM consists of replacing the "absolute value of temporal information" (ATI), given by the VQM algorithm, by the saliency map, as shown in Figure 5. The parameter ATI is used because in the VQM metric 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.

We use a public database created by the Laboratory for Image and Video Engineering (LIVE) and Center for Perceptual Systems (CPS) at the University of Texas at Austin (UT Austin) to test the tested metrics with incorporated saliency maps. The chosen database is the LIVE Video Quality Database[8],[9] which contains a set of 150 distorted videos and corresponding Difference Mean Opinion Scores (DMOS) values. There are four different types of distortions in this database: MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks.

In Table 1, we present the Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) for all the metrics tested in this work, when compared to the subjective data of the LIVE database. As can be observed, with the addition of the saliency maps the performance of most metrics improved. It is interesting to notice that the highest improvements in performance corresponded to the SSIM and MOVIE-S metrics. These particular metrics are the ones which only take into consideration the spatial information of the video. The best improvement was for SSIM. In comparison with the original metric, the gain was 17,36% (Pearson Correlation) and 25,75% (Spearman Correlation).

For the temporal metrics (MOVIE and VQM), the MOVIE had the least improvement, 0,0036% (Pearson Correlation) and -0,28% (Spearman Correlation). This is comprehensible, because MOVIE is the most complex metric of all metrics analized. VQM-C-AV had the best improvement, with a gain of 0,096% (Pearson Correlation) and 0,43% (Spearman Correlation). The MOVIE-T-AV with visual attention had the best Spearman correlation of all metrics tested, equal to 0.8011. This corresponded to an improvement of 0,062%.

Quality Metric	**PCC**	**SCC**
SSIM	0.5437	0.5401
SSIM-AV	**0.6381**	**0.6792**
MS-SSIM	0.7084	0.7445
MS-SSIM-AV	0.7031	**0.7578**
VQM	0.7297	0.7153
VQM-C-AV	**0.7304**	**0.7184**
VQM-A-AV	**0.7301**	**0.7156**
MOVIE	0.7898	0.7893
MOVIE-AV	**0.7901**	0.7871
MOVIE-S	0.7185	0.7077
MOVIE-S-AV	**0.7201**	**0.7173**
MOVIE-T	0.8200	0.8006
MOVIE-T-AV	0.8199	**0.8011**

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.

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 and MOVIE-S metrics), i.e. for the metrics that only took into consideration spatial degradations.

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] Wang, Z. and Bovik, A., "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE* **26**, 98 –117 (jan. 2009).

[2] Engelke, U., Kaprykowsky, H., Zepernick, H.-J., and Ndjiki-Nya, P., "Visual attention in quality assessment," *Signal Processing Magazine, IEEE* **28**, 50 –59 (nov. 2011).

[3] Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L., "Objective video quality assessment methods: A classification, review, and performance comparison," *Broadcasting, IEEE Transactions on* **57**, 165 –182 (june 2011).

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

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

[6] Liu, H. and Heynderickx, I., "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in [*Image Processing (ICIP), 2009 16th IEEE International Conference on*], 3097 –3100 (nov. 2009).

[7] Farias, M. and Akamine, W., "On performance of image quality metrics enhanced with visual attention computational models," *Electronics Letters* **48**(11), 631–633 (2012).

[8] Seshadrinathan, K., Soundararajan, R., Bovik, A., and Cormack, L., "A subjective study to evaluate video quality assessment algorithms," (2010).

[9] Seshadrinathan, K., Soundararajan, R., Bovik, A., and Cormack, L., "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on* **19**(6), 1427–1441 (2010).

[10] Itti, L. and Koch, C., "Computational modelling of visual attention," *Nature Reviews Neuroscience* **2**(3), 194–203 (2001).

[11] Engelke, U., Liu, H., Wang, J., Le Callet, P., Heynderickx, I., Zepernick, H., and Maeder, A., "Comparative study of fixation density maps," *Image Processing, IEEE Transactions on* **22**(3), 1121–1133 (2013).

[12] Itti, L., "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing* **13**, 1304–1318 (Oct 2004).

[13] Pinson, M. H. and Wolf, S., "An objective method for combining multiple subjective data sets," in [*Proc. SPIE Conference on Visual Communications and Image Processing*], **5150**, 583–592 (2003).

[14] Wang, Z., Lu, L., and Bovik, A. C., "Video Quality Assessment Based on Structural Distortion Measurement," *Signal Processing: Image Comm.* **vol19**, 121–132 (2004).

[15] Pinson, M. and Wolf, S., "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting* **50**, 312–322 (Sept. 2004).

[16] Seshadrinathan, K. and Bovik, A., "Motion tuned spatio-temporal quality assessment of natural videos," *Image Processing, IEEE Transactions on* **19**(2), 335–350 (2010).

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

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