On the Performance of Image Quality Metrics Enhanced with Visual Attention Computational Models

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

In this work, we investigate the benefits of incorporating saliency maps obtained with visual attention *computational* models into three image quality metrics. In particular, we compare the performance of simple quality metrics with quality metrics that incorporate saliency maps obtained using three popular visual attention computational models. Results show that performance of simple quality metrics can be improved by adding visual attention information. Nevertheless, gains in performance depend on the precision of the visual attention model, the type of distortion, and the characteristics of the quality metric.

Introduction: A big effort in the scientific community has been devoted to the development of better image and video quality metrics that incorporate human visual system (HVS) features and, therefore, correlate better with the human perception of quality [2]. A recent development in the area consists of trying to incorporate aspects of visual attention in the design of quality metrics, using the assumption that distortions appearing in less salient areas might be less visible and, therefore, less annoying.

Initial studies have reported that the incorporation of *subjective* saliency maps increases the performance of quality metrics [5]. Subjective saliency maps are obtained through psycho-physical experiments using an eye-tracker equipment which records where subjects are fixating as they look at pictures. Although subjective saliency maps are considered as the ground-truth in visual attention, they cannot be used in real-time applications. Thus, in order to incorporate visual attention aspects into the design of image quality metrics, we have to use visual attention *computational* models to generate *objective* saliency maps.

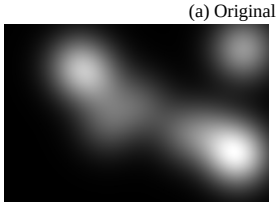
Very few works tested the incorporation of specific computational attention models into image quality metrics [9]. Up to date, there has been no work that compared the incorporation of visual attention computation models versus subjective saliency maps. In this work, we investigate the benefit of incorporating objective saliency maps into three image quality metrics. We compare the performance of the original quality metrics with the performance of quality metrics that incorporate *objective* saliency maps. Also, we study the effects that different types of degradations have on the computational model and, consequently, on the performance of the final metric.

Incorporation of Visual Attention Models: Visual attention is a feature of the HVS that is responsible for defining which areas of the scene are relevant and should be attended. There are two visual selection mechanisms: *bottom-up* and *top-down*. The bottom-up mechanism is an automated selection that is controlled mostly by the signal. It is fast and short lasting, being performed as a response to low-level features that are perceived as *visually salient*. The top-down mechanism is controlled by higher cognitive factors and external influences, such as semantic information, viewing task, and personal preferences, context. It is slower and requires a voluntary effort.

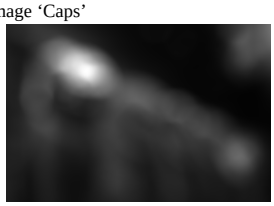
In this work, we consider three popular bottom-up visual attention computational models: Itti's model [3], Achanta *et al.*'s model [1], and GAFFE model (Gaze-Attentive Fixation Finding Engine) [6]. For a given image, these models generate a gray-scale saliency map indicating image regions that are most likely to attract attention. In the saliency maps, higher luminance values correspond to higher saliency pixels, while lower values correspond to lower saliency ones. Fig. 1(a) depicts the image 'Caps', while the corresponding saliency maps generated using Itti's, Achanta's, and GAFFE models are depicted in Figs. 1.(b), (c), and (d), respectively. We used the subjective saliency maps from the TUD LIVE Eye Tracking database as our visual attention *ground-truth* [4]. These saliency maps were collected in a subjective experiment that used twenty-nine source images from the LIVE database [7]. Fig. 1(e) depicts the subjective saliency map corresponding to the image 'Caps'.

We combine the information from the saliency maps into three different full-reference (FR) image quality metrics: Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM) index [8]. The MSE and PSNR error maps are calculated using the following equations:

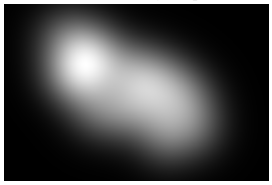
(a) Original image 'Caps'



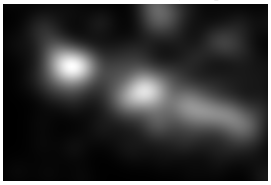
(b) Itti's saliency map [3]



(c) Achanta's saliency map [1]



(d) GAFFE saliency map [6]



(e) subjective saliency map [4]

Fig. 1. Saliency maps corresponding to the image 'Caps'.

$$MSE(x,y) = (I_o(x,y) - I_t(x,y))^2 \,, \tag{1}$$

and

$$PSNR(x,y) = 20\log_{10}\left(MAX_i/\sqrt{\text{MSE}(x,y)}\right) \tag{2}$$

where $I_o(x,y)$ is the original image pixel, $I_t(x,y)$ is the test image pixel, MAX_i is the highest intensity level of the pixels, and x and y are the horizontal and vertical coordinates. For SSIM, we used the local $SSIM(x,y)$ map, as described in [7]. The combination or integration process consists of using the gray-scale pixel values of the saliency maps as *weights* for the error maps generated by the three quality metrics. The modified saliency-based quality metrics for the corresponding FR metrics are given as

$$\text{SM-MEM} = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\text{MEM}(x,y)\cdot SAL(x,y)}{\sum_{x=1}^{M}\sum_{y=1}^{N}SAL(x,y)}, \tag{3}$$

where $SAL(x,y)$ is the saliency map pixel and $\text{MET}(x,y)$ is the error map pixel calculated using one of the FR quality metric (SSIM, PSNR or MSE).

Simulation Results: The performance of an image quality metric is measured by how well its output scores (quality estimates) correlate to the Mean Observer Scores (MOS) given by observers in a subjective experiment. To compare the performance of the three original quality metrics with the saliency-based quality metrics (Eq. 3), we used the LIVE database [7] that contains images with the following degradations: JPEG, JPEG2k, Gaussian Blur (GB), Fast Fading (FF), and White Noise (WN).

Although the saliency map is generally estimated using the original images, for all computational models we obtain saliency maps using both the original and test images since we want to analyze how the performance of the saliency-based metrics is affected by the use of degraded maps. To make sure that the differences in performance are not by chance, we also test the performance with 'switched' (sw) saliency maps that consisted of picking a random saliency map corresponding to another image in the database. We also test the incorporation of subjective (su) saliency maps. To identify the different models, we substitute the initials SM in Eq. 3 by the first letter of the saliency map used ('I' for Itti, 'A' for Achanta, and 'G' for GAFFE) followed by 'o' (original) or 't' (test), indicating whether the computational model obtained the saliency map using the original or test image, SAL. In Tables 1-3, we present the Spearman correlation coefficients for MSE, PSNR and SSIM and their saliency-based versions. Correlation values of saliency-based metrics that represent a gain in comparison to the original metrics are depicted in **bold**.

For saliency-based MSE and PSNR metrics, when we consider each individual distortion (columns 2-6 of Tables 1 and 2), the correlation coefficients improve for almost all distortions, with performance gains varying from 1.2% to 2.1%. The only exception is the degradation White

Table 1: Spearman correlation coefficients for MSE metric.

Model	JPEG	JPEG2k	GB	FF	WN	All
MSE	0.90117	0.88872	0.78249	0.88549	0.98564	0.87270
Su-MSE	**0.91891**	**0.91338**	**0.81171**	**0.90894**	0.98526	**0.89080**
Io-MSE	**0.91702**	**0.90620**	**0.80246**	**0.90200**	0.98530	**0.88650**
It-MSE	**0.91895**	**0.90500**	**0.80360**	**0.90250**	0.98550	**0.88620**
Ao-MSE	**0.91187**	**0.90757**	**0.80522**	**0.90432**	0.98548	**0.88650**
At-MSE	**0.91170**	**0.90470**	**0.79640**	**0.90180**	0.98560	**0.88480**
Go-MSE	**0.91670**	**0.91020**	**0.79140**	**0.90510**	0.98530	**0.88790**
Gt-MSE	**0.91780**	**0.91150**	**0.79030**	**0.90430**	0.98520	**0.88830**
Sw-MSE	0.90091	**0.89250**	0.74067	0.87540	0.98497	0.86860

Table 2: Spearman correlation coefficients for PSNR metric.

Model	JPEG	JPEG2k	GB	FF	WN	All
PSNR	0.90120	0.88872	0.78249	0.88549	0.98564	0.87270
Su-PSNR	**0.91891**	**0.91338**	**0.81159**	**0.90894**	0.98523	**0.89080**
Io-PSNR	**0.91696**	**0.90620**	**0.80246**	**0.90200**	0.98530	**0.88240**
It-PSNR	**0.91696**	**0.90500**	**0.80370**	**0.90250**	0.98550	**0.88620**
Ao-PSNR	**0.91187**	**0.90757**	**0.80522**	**0.90432**	0.98548	**0.88650**
At-PSNR	**0.91170**	**0.90470**	**0.79640**	**0.90180**	0.98560	**0.88480**
Go-PSNR	**0.91670**	**0.91020**	**0.79140**	**0.90510**	0.98530	**0.88790**
Gt-PSNR	**0.91780**	**0.91150**	**0.79030**	**0.90430**	0.98520	**0.88830**
Sw-PSNR	0.90091	**0.89250**	0.74070	0.87550	0.98500	0.86860

Noise for which the performance decreases with the incorporation of any type of saliency map. Because of the similarity between PSNR and MSE, the correlation values of their corresponding saliency-based metrics are very similar. The best gains are obtained for the subjective maps (1.7% to 2.5%) and GAFFE objective maps (1.6% to 2.2%). Achanta's model presents the best performance for the degradation Gaussian Blur.

For saliency-based SSIM, considering again only the individual distortions (columns 2-6 of Table 3), the performance improves when subjective and GAFFE saliency maps are used. Although GAFFE is the computational model with the best performance, the gains in performance vary with the distortions. The gain for JPEG is only 0.03%, while for other distortions they range from 0.2% to 1%. When Achanta and Itti models are used there is no improvement for JPEG and JPEG2k. For the other degradations, using Itti and Achanta models provides improvement gains from 0.06% to 1.2%. The degradation corresponding to the worst performance is White Noise, with gains from 0.06% to 0.4%. For Gaussian Blur, Achanta's model incurs in a higher performance gain (1.2%) than Itti's model (0.16%) or GAFFE (0.28%).

For the set containing all types of distortions ('All' – column 7 of Tables 1-3), the correlation coefficients of saliency-based metrics show gains raging from 1.1% to 1.9%. The subjective saliency maps show the highest gain in performance, followed by the GAFFE saliency maps generated from original images. The saliency maps obtained from test images presented an inferior performance for both GAFFE and Itti models, but a better performance for Achanta. For the switched saliency maps, the performance was worse than with any other saliency map. These results seem to point out that the precision of the saliency map has an impact on the performance of the metrics. The correlation values are comparable to the values found by other researchers [5].

Overall, the performance gains for MSE and PSNR were higher than for SSIM. This is expected since SSIM already includes some of the same parameters (e.g. contrast and texture) that are taken into account by attention models. The computational model that presents the best performance is GAFFE. Although the results of GAFFE for SSIM are not as significant as for PSNR and MSE, the gains in performance are close (sometimes higher) than what is obtained with subjective saliency maps. For Gaussian Blur, the best performance model is Achanta's – the simplest of the three models tested. Blur removes image details making it easier for simpler models to find salient areas. Most models shows no or very small gain in performance for White Noise. Noise adds more details to the saliency map, making more difficult to find salient areas.

Conclusions: Our results show that the computational models were 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. Nevertheless, the improvement in performance was higher for the simpler metrics (PSNR and MSE) than for the more complex metric (SSIM). The results also showed that the performance depended on distortion type, with White Noise presenting the lowest gains.

Acknowledgment: 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.

Mylène C.Q. Farias and Welington Y.L. Akamine (*Department of Computer Science, University of Brasília, Brasília - DF, 70910-900, Brazil*)

E-mail: mylene@ieee.org

Table 3: Spearman correlation coefficients for SSIM metric.

Model	JPEG	JPEG2k	GB	FF	WN	All
SSIM	0.96958	0.95060	0.92506	0.93681	0.96410	0.92210
Su-SSIM	**0.97029**	**0.95188**	**0.92709**	**0.94480**	**0.96889**	**0.93250**
Io-SSIM	0.96874	0.94850	**0.92657**	**0.94287**	**0.96760**	**0.93090**
It-SSIM	0.95795	0.95060	**0.92730**	**0.94360**	**0.96960**	**0.93200**
Ao-SSIM	0.96821	0.95000	**0.93663**	**0.94462**	**0.96473**	**0.93320**
At-SSIM	0.96830	0.94840	**0.93580**	**0.94370**	**0.96700**	**0.92990**
Go-SSIM	**0.96990**	**0.95560**	**0.92770**	**0.94660**	**0.96860**	**0.93350**
Gt-SSIM	0.96880	**0.95410**	**0.92550**	**0.94580**	**0.96940**	**0.93300**
Sw-SSIM	0.96388	0.94593	0.89801	0.93302	0.95786	0.91700

References

1 R. Achanta, F. Estrada, P. Wils, and S. Susstrunk. 'Salient region detection and segmentation', In A. Gasteratos et al., editors, *Computer Vision Systems*, volume 5008 of *Lect. Notes in Comp. Science*, pp. 66–75, 2008.

2 S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. on Broadcasting*, 57(2):165 –182, june 2011.

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

4 H. Liu and I. Heynderickx, TUD image quality database: Eye-tracking release 1, 2009.

5 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 Intrn. Conf. on Image Processing (ICIP)*, pp. 3097 –3100, nov. 2009.

6 U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack. GAFFE: A gaze-attentive fixation finding engine.*IEEE Trans. on Image Processing*, 17(4):564 –573, april 2008.

7 H.R. Sheikh, Z.Wang, L. Cormack, and A.C. Bovik. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

8 Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004.

9 A.K. Moorthy and A.C. Bovik, "Visual Importance Pooling for Image Quality Assessment," Selected Topics in Signal Processing, IEEE Journal of, vol.3, no.2, pp.193-201, April 2009.