STUDYING THE ADDED VALUE OF VISUAL ATTENTION IN OBJECTIVE IMAGE QUALITY METRICS BASED ON EYE MOVEMENT DATA

Hantao Liu [1] *and Ingrid Heynderickx* [1, 2]

[1] Department of Mediamatics, Delft University of Technology, Delft, The Netherlands
[2] Group Visual Experiences, Philips Research Laboratories, Eindhoven, The Netherlands

ABSTRACT

Current research on image quality assessment tends to include visual attention in objective metrics to further enhance their performance. A variety of computational models of visual attention are implemented in different metrics, but their accuracy in representing human visual attention is not fully proved yet. Thus, to provide more accurate evidence on whether and to what extent visual attention can be beneficial for objective quality prediction, the use of "ground truth" visual attention data is highly desired. In this paper, the data of an eye-tracking experiment are integrated in two objective metrics well-known in literature. Experimental results demonstrate that there is indeed a gain in performance including visual attention in objective metrics. The amount of gain in performance tends to depend on the type of objective metric and image distortion.

Index Terms— Visual attention, eye tracking, natural scene saliency, distortion metric, image quality assessment

1. INTRODUCTION

Objective metrics are aimed at predicting perceived image quality aspects consistent with subjective human evaluation [1]. Metrics based on the human visual system (HVS) are potentially more reliable for accurate quality prediction [2]. It has been demonstrated that incorporating some lower level aspects of the HVS, e.g. frequency sensitivity, luminance masking and texture masking improves the performance of an objective metric (see e.g. in [2], [3]). Studies evaluating whether also higher level aspects of the HVS, such as visual attention, are beneficial for objective quality prediction, and if so, how to apply them in metric design are still limited, but recently have emerged as an active research area [4]–[8].

Intuitively one may expect that a distortion occurring in an area that gets the viewer's attention is more annoying than in any other area. This idea is recently exploited in e.g. [4]–[6], in which the performance of a metric is improved by weighting the measured local distortions with the local saliency. The essential concept behind the design of these metrics is that saliency driven by the original image content (i.e. referred to as *natural scene saliency*) and saliency driven by image distortions are taken into account separately, and they are combined to determine the overall quality score. Obviously, the latter saliency is kind of addressed by the distortion metric itself, and therefore, only the former saliency needs to be explicitly considered.

Before developing an attention based metric, it is worthwhile to know exactly whether and to what extent including visual attention can improve existing distortion metrics, since in real-life implementations the measured gain in metric performance should be balanced against the additional costs needed for the rather complex attention model. The investigation of the added value of saliency largely depends on the reliability of the visual attention data used in the metric. Computational attention models are available [4]–[6], but they are either specifically designed or chosen for a specific domain, or their accuracy in predicting human visual attention is not fully proved yet. Therefore, we decided to use "ground truth" visual attention data for the evaluation of their added value in objective metrics.

A similar approach was adopted in [7]; they also used eye-tracking data to investigate the added value of visual attention in objective metrics. Their results, however, were inconsistent with those found in [4]–[6], i.e. no clear improvement in the performance of the objective metric was found by weighting the local distortions with the local saliency. It should, however, been noted that their eye-tracking data were collected during quality assessment. As such, each original image content (having various distorted versions) was viewed several times by each observer. This might have affected the recorded saliency in the sense that it might have been more affected by the image distortions than by the natural scene saliency, as was discussed in [8]. As a consequence, the visibility of distortions may have been overestimated in the approach taken in [7], and that possibly explains the difference with the conclusions in [4]–[6].

In this paper, we further rely on the approach of [4]–[6] and use the natural scene saliency in the design of an attention based metric. However, instead of using a computational model for visual attention, we performed an eye-tracking experiment to obtain "ground truth" visual attention data. The validation process was carried out with two well-known objective metrics, and for the entire LIVE image quality assessment database [9].

2. VISUAL ATTENTION DATA

It is generally agreed that under normal circumstances human eye movements are tightly coupled to visual attention [10]. Thus, eye movement recording is so far the most reliable means for studying the human visual attention. To obtain data of natural scene saliency, an eye-tracking experiment with unimpaired images under natural viewing conditions was conducted.

2.1. Eye-Tracking Experiment

The eye-tracking experiment was carried out in the Experience Lab of the Delft University of Technology. Eye movements were recorded with an infrared video-based tracking system (iView X RED, SensoMotoric Instruments). It has a sampling rate of 50 Hz, a spatial resolution of 0.1°, and a gaze position accuracy of 0.5°-1.0°. Since the system can compensate for head movements within a certain range, a chin rest was sufficient to reduce head movements and ensure a constant viewing distance of 70 cm. The twenty-nine source images of the LIVE image quality assessment database [9] were used as stimuli, and were displayed on a 19-inch CRT monitor with a resolution of 1024x768 pixels and an active screen area of 365x275mm.

Twenty students, being twelve males and eight females, inexperienced with eye-tracking recordings, were recruited as participants. Each participant saw all stimuli in a random order. Each stimulus was shown for 10s followed by a mid-gray screen during 3s. The participants were requested to look at the images in a natural way ("view it as you normally would"). Each session (per subject) was preceded by a 3x3 point grid calibration for the eye-tracking equipment.

2.2. Saliency Map

A human saliency map representative for visual attention is usually derived from the spatial pattern of fixations in the eye-tracking data [10]. To construct this map, each fixation location gives rise to a gray-scale patch whose activity is Gaussian distributed. The width (σ) of the Gaussian patch approximates the size of the fovea (about 2° visual angle). A mean saliency map that takes into account all fixations of all subjects is calculated as follows:

$$S_i(k,l) = \sum_{j=1}^{T} \exp[-\frac{(x_j - k)^2 + (y_j - l)^2}{\sigma^2}] \quad (1)$$

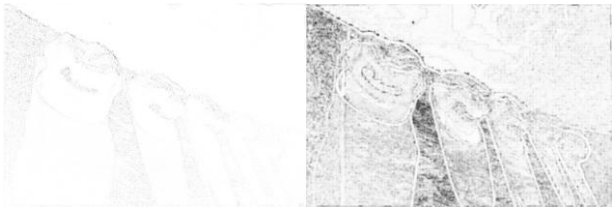
where S_i(k, l) indicates the saliency map for stimulus I_i of size MxN pixels (i.e. kϵ[1, M] and lϵ[1, N]), (x_j, y_j) indicates the spatial coordinates of the *j*th fixation (j=1…T), T is the total number of all fixations over all subjects, and σ indicates the standard deviation of the Gaussian. The intensity of the resulting saliency map is linearly normalized to the range [0, 1]. Figure 1 illustrates the saliency map of one of the images used in our experiment.

(a) (b)

Fig. 1. Illustration of the human saliency map: (a) original image and (b) saliency map.

3. OBJECTIVE DISTORTION METRICS

As a starting point, the added value of including visual attention is evaluated for two well-known and widely used objective metrics: PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index) [1]. Both metrics estimate the image distortion locally, yielding a quantitative distortion map. Figure 2 illustrates the distortion map calculated by PSNR and SSIM, respectively, for the image shown in Fig. 1(a). The intensity value of each pixel in the distortion map indicates the local degree of distortion, i.e. the lower the intensity, the larger the distortion is.



(a) (b)

Fig. 2. Illustration of the distortion map calculated for a JPEG compressed image (bit rate 0.41bbp) with its original shown in Fig 1(a): (a) distortion map of PSNR, and (b) distortion map of SSIM. The lower the intensity, the larger the distortion is.

4. EXPERIMENTAL VALIDATION

The added value of including visual attention in objective metrics is evaluated by comparing the performance of metrics weighted with the saliency map obtained from the eye-tracking experiment to the performance of the same metrics without visual attention.

4.1. Objective Metrics based on Human Saliency Map

The human saliency map is included in the PSNR and SSIM metrics, by locally weighting the corresponding distortion map. It should be noted that the combination strategy used in this paper is still a simple weighting function similar to that in [4]–[7]. More complex combination strategies may further improve the obtained performance (as discussed in [6]), but are not yet investigated here.

Adding saliency results in two attention based metrics, which are referred to as WPSNR and WSSIM, respectively. The metric WPSNR is defined as follows:

$$WPSNR = 10\log_{10}(\frac{MAX^2}{WMSE}) \quad (2)$$

where

$$WMSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\{[D_{i,k}(x,y) - I_i(x,y)]^2 \cdot S_i(x,y)\}}{\sum_{x=1}^{M}\sum_{y=1}^{N} S_i(x,y)} \quad (3)$$

and *MAX* is the maximum pixel value of the image (*MAX*=1 in our experiments), $D_{i,k}$ indicates the distorted image, I_i indicates the original image, and S_i indicates the corresponding saliency map derived from the eye-tracking experiment. The metric WSSIM is defined as:

$$WSSIM = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[ssim\_map(x,y) \cdot S_i(x,y)]}{\sum_{x=1}^{M}\sum_{y=1}^{N} S_i(x,y)} \quad (4)$$

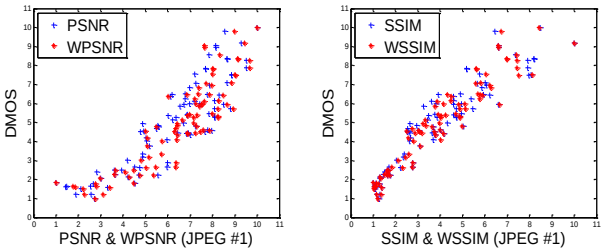
where *ssim_map* is calculated between the distorted image $D_{i,k}$ and its original image I_i, using the SSIM metric.

4.2. Experimental Results

The experiment was conducted for the entire LIVE database. It consists of 779 images distorted with JPEG compression, JPEG2000 compression, white noise, Gaussian blur, and simulated fast fading Rayleigh (wireless) channel. A difference mean opinion score (DMOS) was derived for each distorted image by an extensive subjective quality assessment study [9].

The four metrics PSNR, SSIM, WPSNR and WSSIM are applied to the LIVE database. Figure 3 shows the scatter plots of the DMOS versus each of the four metrics for the different distortion types. The metrics' performance is also quantified by the Pearson and Spearman correlation coefficients between the DMOS and the predictions of the objective metrics, as prescribed by the VQEG [11]. As suggested in [11], one may use a nonlinear fitting of the metrics' predictions to the DMOS before computing the correlation coefficients. Indeed, the image quality community is more accustomed to e.g. a logistic function, to fit the metric's predictions to the DMOS. It may, for example, account for a possible saturation effect at high qualities. A non-linear fitting usually results in higher correlation coefficients in absolute terms, while generally keeping the relative differences between the metrics [2]. On the other hand, without a sophisticated non-linear fitting (often including various parameters) the correlation coefficients cannot mask a bad performance of the metric itself, as discussed in [6]. Therefore, the non-linear fitting is omitted here, and the correlation coefficients are directly computed between the DMOS and the metrics' predictions.

Figure 4 gives the corresponding correlation coefficients. It demonstrates that there is indeed a gain in performance including visual attention in the objective metrics PSNR and SSIM, independent of the metric used and of the image distortion type tested. There is, however, a difference in the amount of gain in performance dependent on the metric. The gain of WPSNR over PSNR corresponds to an average increase in the Pearson correlation coefficient (over all distortion types for the LIVE database) from 0.88 to 0.90 (i.e. ΔP=**2%**) and in the Spearman correlation coefficient from 0.87 to 0.89 (i.e. ΔS=**2%**). The gain of WSSIM over SSIM is ΔP=**3%** (from 0.91 to 0.94) and ΔS=**3%** (from 0.92 to 0.95). Furthermore, the amount of gain in performance also depends on the distortion type. The gain of WSSIM over SSIM for Gaussian blur is ΔP=**7%** (from 0.85 to 0.92) and ΔS=**5%** (from 0.89 to 0.94), but for white noise is ΔP=**1%** (from 0.96 to 0.97) and ΔS=**1%** (from 0.96 to 0.97). Analogously, the gain of WPSNR over PSNR yields a ΔP=**2%** (from 0.77 to 0.79) and a ΔS=**3%** (from 0.78 to 0.81) for Gaussian blur, compared to a ΔP=**0.01%** (from 0.9792 to 0.9793) and a ΔS=**0%** (from 0.985 to 0.985) for white noise.



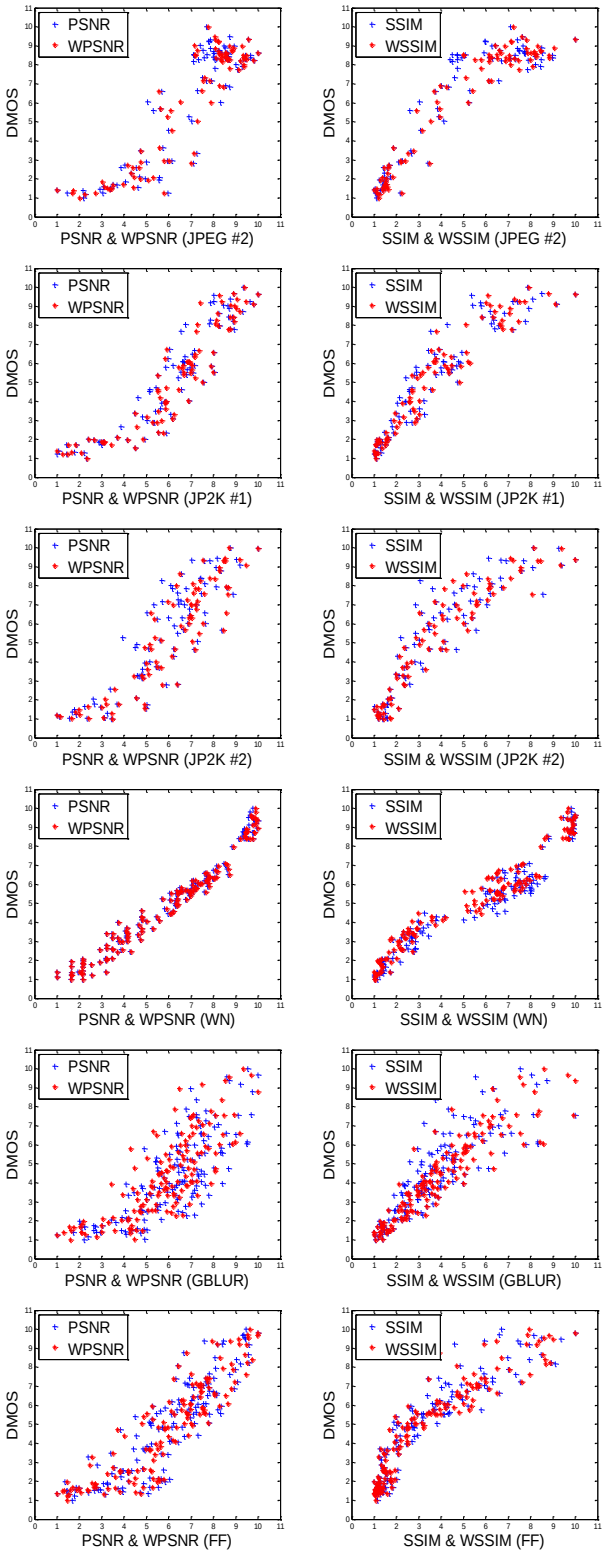


Fig. 3. Scatter plots of DMOS vs. four metrics PSNR, WPSNR, SSIM, and WSSIM for JPEG#1, JPEG#2, JPEG2000#1, JPEG2000#2, white noise (i.e. WN), Gaussian blur (i.e. GBLUR), and fast-fading (i.e. FF), respectively.

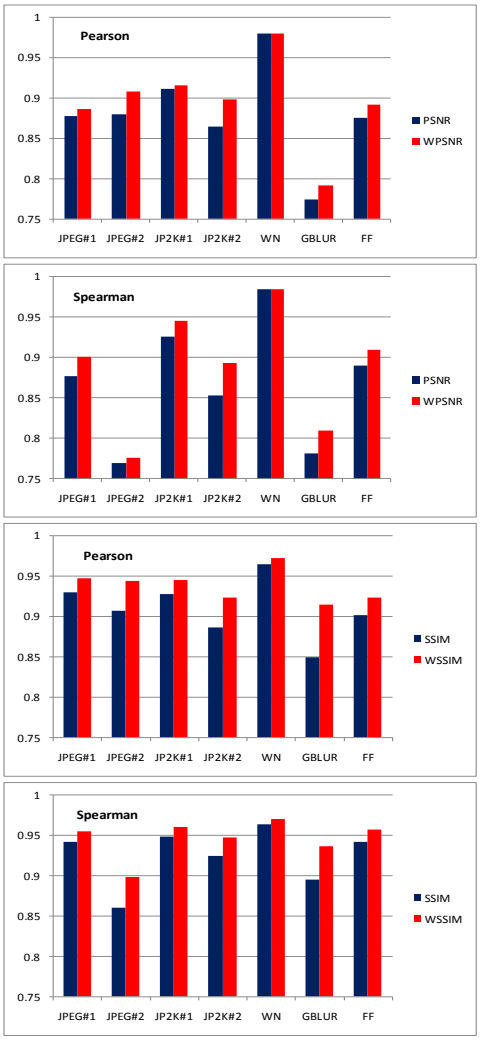


Fig. 4. Correlation coefficients (without nonlinear regression) of four metrics PSNR, WPSNR, SSIM, and WSSIM for JPEG#1, JPEG#2, JPEG2000#1, JPEG2000#2, white noise (i.e. WN), Gaussian blur (i.e. GBLUR), and fast-fading (i.e. FF), respectively.

5. DISCUSSION AND FUTURE WORK

Our results show that adding visual attention improves the performance of the PSNR and SSIM metric. This conclusion is in agreement with the results of [4]–[6], but in contradiction to the conclusions of [7]. As shown with this paper, the contradiction is not a consequence of using saliency from eye-tracking data (as used here and in [7]) instead of from a model (as used in [4]–[6]). More probably, the contradiction results from the different type of visual attention data used: i.e. natural scene saliency here versus saliency during scoring in [7]. This, however, still needs to be proven.

The added value of visual attention in terms of a performance improvement is shown here for the metrics PSNR and SSIM. Although this study is limited in the number of metrics used, we do not expect the result to be different for other objective metrics. It might, however, be that the added value is more limited in case the performance of the metric itself is already high. To better visualize differences in performance we propose (as also already stated in [2], [6]) to avoid any non-linear fitting and to directly use linear correlation between the metric predictions and subjective data.

It should be noted that the combination strategy between the human saliency map and the distortion map of the metric in this paper is still a simple weighting function. This may underestimate the impact of perceived artifacts on image quality, as already discussed in [6], [7]. Therefore, we expect that an advanced combination strategy (e.g. in which the weighting function is adapted to the features of the image distortions) yields a larger gain in performance for some metrics. This should be further investigated using the available visual attention data.

6. CONCLUSIONS

In this paper, we provide, based on eye-tracking data, more accurate quantitative evidence on whether visual attention is beneficial for objective metrics, and if so, to what extent. Our results show that there is indeed a gain in the performance of the PSNR and SSIM metrics. The amount of gain in performance varies between both metrics and for the same metric between different image distortion types.

7. REFERENCES

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers, 2006.

[2] S. Winkler, "Vision Models and Quality Metrics for Image Processing Applications," Ph.D. dissertation, Dept. Elect., EPFL, Lausanne, 2002.

[3] H. Liu and I. Heynderickx, "A No-Reference Perceptual Blockiness Metric," in *Proc. IEEE Int. Conf. ICASSP*, pp. 865-868, March 2008.

[4] R. Barland and A. Saadane, "Blind Quality Metric using a Perceptual Importance Map for JPEG-2000 Compressed Images," in *Proc. IEEE Int. Conf. ICIP*, pp. 2941-2944, Oct. 2006.

[5] N. G. Sadaka, L. J. Karam, R. Ferzli, and G. P. Abousleman, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in *Proc. IEEE Int. Conf. ICIP*, pp. 369-372, Oct. 2008.

[6] A. K. Moorthy and A. C. Bovik, "Perceptually significant spatial spooling techniques for image quality assessment," in *Proc. Electronic Imaging*, 2009

[7] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric," in *Proc. IEEE Int. Conf. ICIP*, pp. 169-172, Oct. 2007.

[8] C. T. Vu, E. C. Larson, and D. M. Chandler, "Visual Fixation Patterns when Judging Image Quality: Effects of Distortion Type, Amount, and Subject Experience," in *Proc. IEEE SSIAI*, pp. 73-76, March. 2008.

[9] 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.

[10] N. Ouerhani, R. V. Wartburg, H. Hugli, and R. Muri, "Empirical Validation of the Saliency-based Model of Visual Attention," *Electronic Letters on Computer Vision and Image Analysis*, 3(1): 13-24, 2004.

[11] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," http://www.vqeg.org.