EVIDENCE OF CHANGE BLINDNESS IN SUBJECTIVE IMAGE FIDELITY ASSESSMENT

Steven Le Moan

Massey University
Palmerston North, New Zealand

Marius Pedersen

Norwegian University of Science and Technology
Gjøvik, Norway

ABSTRACT

Change blindness is a striking phenomenon which basically means that we can look without seeing. It originates from a faulty communication between early vision (the eye) and visual working memory (the brain). In this paper, we present evidence that this faulty communication needs to be accounted for in image fidelity assessment (also known as full-reference image quality assessment). We designed a user study to analyse participants opinions based on how much they have to rely on their visual working memory in order to give fidelity score. Results demonstrate that significantly more severe judgments were made when reliance on visual short-term memory was minimal, suggesting limitations in the observers' ability to notice image differences in the typical pairwise comparison setup. Furthermore, a comparison of the efficiency of six state-of-the-art image fidelity assessment models (so-called metrics) reveals that five of them perform significantly better at predicting results obtained when reliance on memory is minimal.

Index Terms— Image Quality Assessment, Perception, Visual Memory, Change Blindness.

1. INTRODUCTION

With the rapid development of digital imaging technologies, understanding how people perceive the quality of images and videos has never been so important. Be it for capture, display or reproduction, an efficient model of visual quality is essential to ensure users satisfaction. There are different ways to approach this problem depending on the application, the context, the availability of reference data, etc. In this paper, we consider the case referred to as *full-reference* image quality assessment, or *image fidelity* assessment (IFA), which concerns principally the reproduction of images (e.g. for compression, gamut mapping, etc). Given two versions of the same image (an original and a reproduction), IFA consists of producing a score that represents the difference of quality between them or, in other words, the loss of quality engendered by the reproduction process. A typical way to obtain reference data for IFA is to perform user studies in which participants have to compare images on a calibrated monitor. These stimuli can be displayed two or three at a time [1] on a monitor and user ratings are typically collected either via pairwise comparison or category judgment methodologies. This type of setup has the advantage of representing a plausible scenario of "real life" IFA, e.g. the comparison of several prints put side-by-side to decide which one is the best. However, one of its main drawbacks is that it compels participants to rely significantly on their visual working memory, which is known to have a limited capacity and bandwidth [2]. The best illustration of this perceptual shortcoming is this striking phenomenon known as *change blindness* [3], which the famous game "Spot the difference" relies on. While the exact origins of this and other associated phenomena such as inattentional blindness [3] or visual crowding [4] are still a source of debate, they are known to come from a faulty communication between early vision (the eye) and visual working memory (the brain). Although two images are displayed at the same time on the monitor, one can only really see sharply a portion of one at a time, roughly corresponding to the size of a thumbnail at arm's length [5]. The way the images are encoded in our brain therefore depends on how our attention is guided throughout the screen [6] but not only, as attention does not imply conscious perception [7, 2]. In other words, the fact that we can *see* does not necessarily mean that we can *notice*, which is something that most existing image quality/fidelity assessment models fail to account for.

These models, also referred to as *image quality metrics* [8] typically involve allegedly human vision-inspired feature extraction and pooling. Yet, despite the fact that popular models such as the multi-scale SSIM index [9] or the recent Visual Saliency Index [10] yield quality scores that correlate to a large extent to subjective opinions, little is truly known about the perceptual mechanisms underlying subjective image quality/fidelity assessment. We argue here that this is partly due to the fact that existing models are essentially meant to simulate early vision, while there are quite a few more things happening further down the visual pathway. Even though visual attention predictors have successfully been used for spatial pooling [10, 11], they cannot fully predict conscious perception. Furthermore, the way image difference features are pooled across frequency bands, orientations and modalities (e.g. lightness-difference, -contrast and -structure in the case of the SSIM index, but also chroma, hue, etc) often lacks bi-

Fig. 1. Selected scenes from TID2013 [1].

ological plausibility. In that regard, findings from fields such as scene understanding [12], and visual semantics [13] should be given more consideration given the importance of image content in the quality assessment task [14].

In order to demonstrate the importance of accounting for visual working memory in state-of-the-art IFA models, we previously carried out a user study in which the need for observers to rely on their visual short-term memory was maximised [15]. Results indicated a significant influence of the latter in the IFA task. Here, we designed a new user study involving image pairs meant to provoke change blindness. The experiment was carried out in two locations, in Norway and in New Zealand and consisted of two sessions: first, images were displayed next to another in pairs, whereas in the second session, the same pairs were displayed one "under" the other, so that observers could see only one at a time, with the possibility to switch from one to the other as many time as desired. The particularity of this second session lies in the fact that reliance on visual working memory was minimal as there was no blank screen when switching between images so that differences were more readily available to conscious perception.

2. USER STUDY

2.1. Methodology

As previously mentioned, the experiment consisted of two sessions, carried out one after the other with a very brief pause in between.

In the first session, for each image pair, both images were displayed at the same time on the monitor, as shown in Figure 2 on the left hand side. Observers had to do a so-called category judgment in that they had to select one of five labels to described how they perceived the difference of quality between the two images:

- "Not perceptible",
- "Perceptible, but not annoying",
- "Slightly annoying",
- "Annoying",
- "Very annoying".

Observers had to confirm that they clearly understood what each of these labels refer to prior to beginning the experiment.

In the second session, stimuli were presented in a different manner. Instead of side-by-side, images were shown one "under" the other, at the exact same position on the screen, as shown in Figure 2 on the right hand side. A button labeled "Toggle" allowed participants to alternate between the two images as many times as they needed to reach a verdict. Note that the switch was direct in that there was no transition screen between stimuli, thus making the visual differences between them more readily available to conscious perception. Observers then had to perform the same task as in session 1, with the exact same five categories.

The order in which pairs were shown as well as the respective positions of original and reproduced images were randomised for each observer and session. Finally, time was monitored during both sessions, without the participants knowing.

2.2. Stimuli

We selected 120 image pairs from the TID2013 database [1]: 5 scenes (see Figure 1), 6 distortions types and 4 distortion levels. These particular scenes were chosen based on their likelihood to provoke change blindness, which was estimated subjectively according to their "complexity" (significant high frequency content, lack of regularities, etc). Change blindness is indeed known to occur when global or local scene statistics are disrupted so that this disruption becomes part of the gist of the altered image (see Figure 2 in [2]). Disruptions of local statistics in complex textures are however difficult to perceive and therefore yield change blindness [16]. The selected distortion types were:

- Additive noise in color components,
- Masked noise,
- Non eccentricity pattern noise,
- Mean shift (intensity shift),
- Contrast change,
- Change of color saturation.

Fig. 2. Screenshots from sessions 1 (left) and 2 (right). Note that these were cropped for better readability.

These were chosen based on similar precepts, but also due to the fact that they tend to be more difficult to predict for current IFA models than other distortions in the database (see Table 2 in [17]).

Finally, distortion levels 1 to 4 (out of 5) were selected, therefore omitting the worst cases of degradations. The reason for this choice is that the higher the distortion level, the less likely the occurrence of change blindness.

2.3. Participants

A total of 25 observers participated to the experiment (15 in NZ, 10 in Norway). They all had to pass a Ishihara test prior to the experiment in order to ensure that they had colour-normal vision. Those who needed glasses or contact lenses were asked to wear them during the experiment. Ages ranged between 24 and 54, about 80% of participants were male and various cultural backgrounds were represented. None of them was given any indications as to the actual goals of the experiment prior to it. A screening based on the method described in [18] revealed that all observers were valid.

2.4. Viewing conditions

We used Eizo ColorEdge displays (CG2420 in New Zealand and CG246W in Norway), both 61cm/24.1" and calibrated with an X-Rite Eye One spectrophotometer for a colour temperature of 6500K, a gamma of 2.2 and a luminous intensity of 80cd/m^2. Both experiments were carried out in a dark room, and participants were given about 20 seconds to adapt to the obscurity after they had been given instructions and the lights had been switched off. The distance to the screen was set to approximately 50cm.

3. RESULTS

Intuitively, we can assume that the second session would lead to more severe ratings (i.e. corresponding to lower image fidelity) as it is meant to emphasise image differences more than in the first one. Additionally, one can expect a smaller intraobserver variability in session two, assuming that change blindness (which occurs only in session 1) can affect a person differently whether it is the first time they observe a particular image pair or not. Finally, as mentioned previously, we have reasons to believe that current IFA models would perform better on the data gathered from session 2 than that from session 1, due to the fact that they fail to account for change blindness. We will now verify whether these assumptions are valid or not.

3.1. Did the second session yield more severe ratings?

In order to compare ratings from both sessions, we used the sign test for the following null hypothesis (NH): "The difference between ratings from session 1 and session 2 has zero median". Overall, for all observers and stimuli, a one-sided test rejects the null hypothesis at the 98% confidence level. A Wilcoxon signed rank test also rejects (at 99% confidence level) the NH that the difference between ratings from the two sessions comes from a distribution with zero median. Note that, unlike the sign test, the Wilcoxon signed rank test assumes that the difference between ratings follows a symmetrical distribution and is therefore more restrictive. For this reason, we chose to report only results from the sign test (at the 95% confidence level) in the remainder of this section.

Out of the six distortions types considered, results from only three of them (Non eccentricity pattern noise, Mean shift and Contrast change) led to reject the NH. Looking at each observer individually, results from a majority of 16 of them also led to a rejection of the NH. Two of them actually gave significantly *less* severe ratings in session 2, which could be explained by the fact that the way change blindness affected them in session 1 led them to hallucinate image differences, as reported in a recent study [16].

Additionally, we computed the Spearman's Rank-Order Correlation Coefficients (SROCC) between ratings from the two sessions. The average (over all observers) SROCC is as low as 0.364, with a standard deviation σ of 0.136 and a maximal value of 0.672. On the other hand, if we look at the average rating over all observers for each image pair, the resulting mean opinion scores yielded an SROCC of 0.562. These results confirm that the correlation between the results from the two sessions is low.

In conclusion, the second session yielded significantly more severe ratings.

3.2. Intraobserver variability

In order to measure the importance of intraobserver variability during the experiment, we ensured that each observer had to rate 10 randomly selected pairs twice in each session. If, for a given image pair, an observer gave a score of 4 the first time and 3 the second time (in the same session), we estimated a variability of 20% (1 out of 5) for this particular pair. The maximal average variability (over the 10 duplicated image pairs) obtained by an observer was 10% in the first session and 8% in the second. On average over all observers, we obtained 3.52% for both sessions, with standard deviations of 3.23% and 2.33% for sessions 1 and 2 respectively.

These results indicate that there was no significant difference of intraobserver variability between the two sessions. They also suggest that change blindness can affect people similarly whether it is the first time they observe a particular image pair or not, in the context of IFA.

3.3. Comparison with objective scores

We measured the relationship between between objective scores from six state-of-the-art IFA models (metrics) and the subjective ratings obtained in sessions 1 and 2. The models are the Multi-Scale Colour Image Difference (MS-iCID) [17], the Visual Saliency Index (VSI) [10], the Feature Similarity Index with colour component (FSIMc) [19], the PSNR-HA [20], Multi-Scale Structural SIMilarity index [9] and the Visual Information Fidelity index (VIF) [21]. This relationship was measured by means of the SROCC after applying a non-linear regression model to the objective score [21]:

$$f(x) = \theta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\theta_2 (x - \theta_3)}} \right) + \theta_4 X + \theta_5, \qquad (1)$$

where θ_i, $i = 1, 2, 3, 4$, and 5 are the parameters to be be fitted. Initial parameters are *max*(subj. scores), *min*(subj. scores), *median*(obj. scores), 0.1, and 0.1. We found however that results obtained with and without non-linear mapping were identical in the case of VSI, FSIMc, MS-iCID and MS-SSIM.

From the results in Table 1 We observe that the prediction accuracy on subjective ratings from session 2 are higher for all the metrics. Note that the last three metrics (PSNR-HA, MS-SSIM and VIF) discard colour information, which explains that they perform worse than the others. In order to assess whether this difference is significant or not, we used a z-test. Note that comparing Spearman correlation coefficients can be done by treating them as Pearson coefficients and using Fisher's z-transform and subsequent z-test [22].

The significance analysis reveals that all metrics except one (VIF) perform better at predicting the subjective data from session 2. It is noteworthy to remember that that all these metrics were calibrated from subjective data obtained from user studies employing a pairwise comparison setup such as the one from session 1. Consequently, these results are even more compelling.

Table 1. Spearman rank order correlation coefficients between objective and subjective scores for each metric and session. (*) indicates that the results from both sessions are significantly different according to a z-test at the 95% confidence level.

	Session 1	Session 2
MS-iCID*	0.567	0.753
VSI*	0.559	0.740
FSIMc*	0.358	0.673
PSNR-HA*	0.460	0.637
MS-SSIM*	0.172	0.495
VIF	0.199	0.306
MOS from TID	0.697	0.432

4. CONCLUSIONS AND FUTURE WORK

We provided evidence that the faulty communication between early vision and visual working memory, which gives rise for instance to change blindness, has a significant influence on observers opinion of image quality in a pairwise comparison task. While several existing computational models of image quality assessment have demonstrated excellent prediction abilities, we argue that one of the reasons why they can only provide a partial understanding of the way people perceive image quality is that they fail to account for what happens beyond early vision. We designed and carried out a user study, the results of which demonstrated that observers made significantly more severe judgments when reliance on working memory was minimal, suggesting indeed an effect of change blindness in the typical pairwise comparison setup. Furthermore, a comparison of the efficiency of six state-of-the-art IFA models revealed that five of them perform significantly better at predicting results obtained also when reliance on working memory was minimal. In conclusion, we recommend that visual short-term memory be accounted for in the design of novel image quality metrics.

5. REFERENCES

[1] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *4th European Workshop on Visual Information Processing*, 2013, pp. 106–111.

[2] M. A. Cohen, D. C. Dennett, and N. Kanwisher, "What is the bandwidth of perceptual experience?," *Trends in Cognitive Sciences*, vol. 20, no. 5, pp. 324–335, 2016.

[3] M.S. Jensen, R. Yao, W.N. Street, and D.J. Simons, "Change blindness and inattentional blindness," *Wiley Interdiscip. Rev. Cognit. Sci.*, vol. 2, no. 5, pp. 529–546, 2011.

[4] D. Whitney and D.M. Levi, "Visual crowding: A fundamental limit on conscious perception and object recognition," *Trends in cognitive sciences*, vol. 15, no. 4, pp. 160–168, 2011.

[5] C.G. Healey and J.T. Enns, "Attention and visual memory in visualization and computer graphics," *IEEE Trans. Visual Comput. Graphics*, vol. 18, no. 7, pp. 1170–1188, 2012.

[6] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[7] V. Lamme, "Why visual attention and awareness are different," *Trends in cognitive sciences*, vol. 7, no. 1, pp. 12–18, 2003.

[8] M. Pedersen and J.Y. Hardeberg, "Full-reference image quality metrics: Classification and evaluation," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 1, pp. 1–80, 2012.

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment," in *IEEE Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, 2003, vol. 2, pp. 1398–1402.

[10] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, 2014.

[11] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, 2016.

[12] C.L. Zitnick, R. Vedantam, and D. Parikh, "Adopting abstract images for semantic scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 627–638, 2016.

[13] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Conference on*. IEEE, 2006, vol. 1, pp. 419–426.

[14] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[15] S. Le Moan, M. Pedersen, I. Farup, and J. Blahová, "The influence of short-term memory in subjective image quality assessment," in *Image Processing, 2016 IEEE International Conference on*. IEEE, 2016, pp. 91–95.

[16] S Le Moan and I. Farup, "Exploiting change blindness for image compression," in *11th International Conference on Signal, Image, Technology and Internet Based Systems (SITIS)*, Bangkok, Thailand, November 2015, pp. 1–7, IEEE.

[17] S. Le Moan, J. Preiss, and P. Urban, "Evaluating the Multi-Scale iCID metric," in *Image Quality and System Performance XII*, Mohamed-Chaker Larabi and Sophie Triantaphillidou, Eds., San Francisco, CA, February 2015, vol. 9396, pp. 9096–38, SPIE.

[18] ITU-R BT.500-12, "Recommendation: Methodology for the subjective assessment of the quality of television pictures," November 1993.

[19] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[20] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *CAD Systems in Microelectronics (CADSM), 2011 11th International Conference The Experience of Designing and Application of*. IEEE, 2011, pp. 305–311.

[21] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[22] L. Myers and M.J. Sirois, "Spearman correlation coefficients, differences between," *Wiley StatsRef: Statistics Reference Online*, 2006.