MEASURING THE EFFECT OF HIGH-LEVEL VISUAL MASKING IN SUBJECTIVE IMAGE QUALITY ASSESSMENT WITH PRIMING

Steven Le Moan

Massey University
Palmerston North, New Zealand

Marius Pedersen

Norwegian University of Science and Technology
Gjøvik, Norway

ABSTRACT

Despite recent advances in subjective image quality research, this topic still calls for many fundamental questions. Though we understand several early vision mechanisms fairly well, little is known about late vision and how the two interact with each other. Here, we look at one particular limitation of our visual system that stems from failures of that interaction: high-level visual masking, best illustrated by the change blindness paradigm. We carried out a user study designed specifically to measure the influence of high-level masking by priming. Results suggest a significant influence of high-level masking in image fidelity assessment at the 95% confidence level for half of the participants, with an average magnitude over three times that of intra-observer variability.

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

1. INTRODUCTION

With the number of digital pictures taken every year running into the trillions [1], the question of image quality has also grown in importance. The last two decades of research have substantially advanced our understanding of what makes the difference between "good" and "bad" image quality, with predictive models boasting linear correlations with average human assessments above 90%. Yet, a number of fundamental questions still need answers, in particular due to the topic's inherent subjectivity. Despite our fairly advanced comprehension of early vision mechanisms such as contrast sensitivity, salience, luminance masking, chromatic adaptation and others, the role played by our brain is still quite unclear. In fact, it is often the subject of rather crude assumptions made for the sake of simplicity. Even though some of these assumptions hold when testing on large databases such as TID [2], the lack of content variety in these benchmarks makes it difficult to draw solid conclusions [3]. Furthermore, the traditional mean opinion scores (MOS) cannot realistically convey the diversity of perceptual abilities from a pool of several hundreds of participants [4]. Factors such as age, gender, cultural background, familiarity with image content, emotion, fatigue, ability to concentrate, and memory, can all affect significantly our response to visual stimuli. Also, with technology constantly evolving towards more pixels, colours, dynamic range, frames per seconds and so on, our perception of quality is changing as well. What we considered to be a good image quality ten years ago would probably be considered unacceptable today.

It is now clear that the keys to understanding image quality lie beyond early vision. In this paper, we propose to focus on the communication channel between early and late vision and how it affects our judgment. Research in cognitive science and psychology have highlighted numerous flaws in that channel [5] and even though the nature and origin of these flaws are still a source of debate, it is clear that the channel is noisy and has a limited bandwidth. In other words, only part of what our eyes can sense is conveyed to the decision-making parts of our brain. A perceptual failure known as *change blindness* [6] illustrates this phenomenon in the most striking way. It basically prevents us from consciously perceiving substantial and often salient differences in image pairs, such as an object re-positioned or re-coloured. Change blindness is the essence of the famous game "Spot the difference". It can affect virtually all types of subjective pair-comparison tasks such as image fidelity assessment. There has been few attempts at exploiting it in order to reduce computational burden in computer graphics and virtual reality environments [7, 8] or for image compression [9]. However, the complexity of the change blindness phenomenon and the paucity of reliable reference data measured with *natural* stimuli do not permit solid conclusions.

Change blindness can also be viewed as a type of *visual masking*, which means it can reduce or suppress the visibility of a "target" stimulus such as a compression artefact, by the presence of another stimulus called a "mask", such as a complex texture. There is a rich literature in image quality assessment that looks at visual masking (see e.g. [10, 11]), yet no one seems to have explicitly studied change blindness. Low-level masking effects such as luminance or contrast masking are well researched but they do not have a direct effect on change blindness. In this paper, we demonstrate that change blindness can influence image fidelity assessment to a significant extent. Consequently, we advocate the need for a reliable model of change blindness and visual awareness to achieve better predictions of subjective judgments.

First, we propose the following definitions of low- and high-level visual masking[1] for clarity.

- *Low-level* masking prevents perception of differences between stimuli **even if** we know where they are. It includes mainly luminance, contrast, texture and structure masking [12, 10, 11]. It is typically associated with early vision mechanisms like luminance adaptation, contrast sensitivity, visual ensembles, etc. Crowding [13] can also be considered as low-level masking but, because it occurs in the peripheral vision, only if the eyes are immobile.

- *High-level* masking prevents perception of differences **unless** we know where they are. In the case of image pairs, high-level masking is effectively a synonym of change blindness. Other perceptual effects such as inattentional blindness [14] also pertain to high-level visual masking, but they are induced by different types of stimuli. Note that texture masking can also be of the high-level kind [9].

The main difference between the two types of masking is the fact that the former is permanent while the latter is only temporary. Incidentally, the notions of ”visibility threshold” and ”just noticeable distortion”, often used in the image quality literature [15], have different meanings in the two cases. For low-level masking, the visibility threshold is invariant over time whereas for high-level masking, the threshold is changed once artefacts are noticed. In the latter case, it would be more natural to define the ”threshold” as the time needed to initially detect the artefact.

In this paper, we advocate the importance to consider low- and high-level masking two fundamentally different perceptual failures in image fidelity assessment. We propose an experimental design based on visual priming to isolate the influence of high-level masking. Priming is a technique which consists of influencing a person’s response to a stimulus via prior exposure to another stimulus. It can effectively eliminate high-level masking [16]. We report that the effect of high-level masking alone was significant for 12 out of 24 participants, with a magnitude on averge 3.5 times larger than intra-observer variability.

2. USER STUDY

2.1. Participants

A total of 24 people participated in the study (17 in NZ, 7 in Norway). They all passed a Ishihara test in order to ensure that they had colour-normal vision. Those who needed glasses or contact lenses were asked to wear them. Ages ranged between 21 and 56, 80% were male and various cultural backgrounds were represented. None of them was given any indications as to the goals of the experiment prior to it. A standard screening [17] revealed that all participants were valid.

2.2. Stimuli

Stimuli were selected from the CID:IQ database [18] and were displayed in pairs. Each pair was made of A) one of 21 different pristine images and B) one of three JPEG-distorted versions of the same image at different compression levels. We will refer to these levels as *low*, *medium* and *high*[2], with high compression level corresponding to the lowest quality.

2.3. Methodology

Participants were asked to ”evaluate the difference between each pair of stimuli displayed on the screen, in terms of quality”. A standard 5-level scale was provided (”Not perceptible”, ”Perceptible, but not annoying”, ”Slightly annoying”, ”Annoying” and ”Very Annoying”. All participants were shown two examples of image pairs prior to the experiment, in an effort to reduce the effect of training during the first session. Examples were the same for all participants and included one pair with nearly no perceptible differences and one with, on the contrary, large artefacts. The whole study lasted about 25 minutes on average per participant.

Session 1 (Figure 1) was a side-by-side pair comparison (original/reproduction). The order and left/right positions of stimuli were systematically randomised. Only the ’low’ and medium’ compression levels were used in this session.

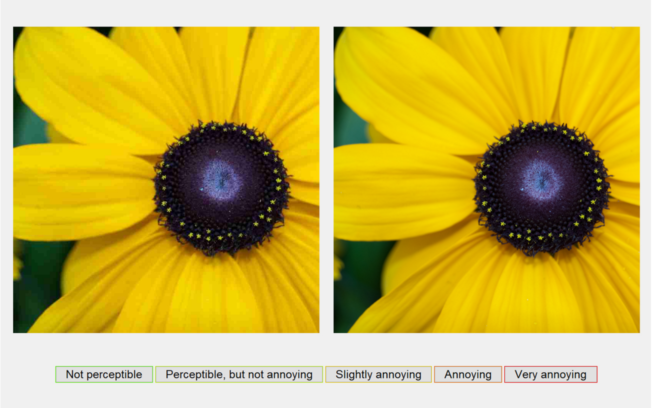

Fig. 1. Stimuli presentation in Sessions 1 & 3.

Session 2 (Figure 2) was dedicated to visual priming. We used the flicker paradigm to make the differences between original and reproduced scenes readily accessible to conscious perception [19]. Participants could alternate between the two scenes as many times as they wanted and they were instructed

[1] Induced by side-by-side image comparison.

[2] These levels respectively correspond to levels 3, 4 and 5 in CID:IQ.

to ”thoroughly examine” the difference between them. Note that instructions for this session were given only after Session 1 to avoid bias. Via priming, this session implicitly trained the participants to pay attention to image regions potentially sensitive to compression artefacts, that they might have missed in Session 1.

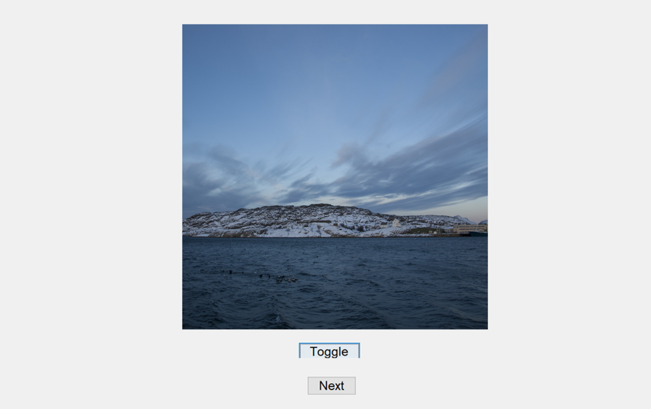

Fig. 2. Stimuli presentation in Session 2. By pressing the ”Toggle” button, participants were able to alternate between the two stimuli without disruption.

Session 3 was a repetition of Session 1, with a different sequence randomisation.

In order to estimate intra-observer variability and do a more robust significance analysis on the results, we displayed 50% of the test image pairs twice in Sessions 1 and 3.

Based on our definitions of low- and high-level masking priming can theoretically affect only on the latter [16]. However, priming can also induce an undesirable memory bias: any subsequent exposure to the same stimuli is likely to result in the participant recognising the scene and basing their quality assessment on their memory of the priming rather than on the stimuli that are displayed. In order to avoid this effect, we used different a higher compression level (lower quality) for priming. In Sessions 1 and 3, we used the low and medium levels (in random order) while in Session 2, we used the high level. Additionally, we used a large number of different pristine images (21 in total) to reduce the chance of people becoming too familiar with the benchmark.

If quality assessments made with and without priming are significantly different, the only cause is the disappearance of high-level masking. Therefore, by comparing results from Sessions 1 and 3, we can then effectively estimate the magnitude of high-level masking in our image fidelity assessment benchmark.

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. The distance to the screen was set to approximately 50cm (without chin rest).

3. RESULTS

3.1. Intra-observer variability

The maximum variability over all participants was estimated at about 6.7% of the assessment scale in both sessions, while the overall average was 2.9% (±2.2%) in Session 1 and 2.6% (±1.9%) in Session 3. This indicates no significant difference of intra-observer variability between before and after priming: participants were fairly consistent with their judgment over the course of the whole experiment.

3.2. Did priming influence quality assessments?

A sign test at the 95% confidence level revealed that 12 out of 24 observers had their assessments significantly influenced by priming and for all of them, priming led to more severe ratings over the whole benchmark. Considering each of these 12 observers individually, the effect of priming (the difference of assessment between Sessions 1 and 3) was 3.5 times larger than intra-observer variability, and 13% of the assessment scale on average. Interestingly, all five female participants were in this group.

Ratings from Sessions 1 & 3 have a Spearman rank order correlation coefficient of 0.54 on average over all observers (standard deviation: 0.18) and never larger than 0.80. This clearly demonstrate that priming does not affect all scenes and distortion levels equally. Among the 21 pristine scenes, 10 were significantly affected by priming for all observers (see Figures 3 and 4). The most striking difference between those two subsets of images is in terms of their visual complexity. Scenes with more objects, less overall symmetry and uniformity are naturally more likely to induce change blindness. On the other hand, it is noteworthy that scenes with some of the finest textures (e.g. the grass in the hedgehogs scene or the peacock’s feathers, Figure 4) were not significantly affected by priming, which suggests that these textures did not induce high-level masking.

All observers rated the difference between image pairs as ”Not perceptible” less often after priming. On average, 24% of the stimuli led to a ”Not perceptible” rating in Session 1, but only 10% in Session 3.

From these results, we deduce that change blindness can have a significant effect on our assessment with a magnitude of 10% of the assessment scale.

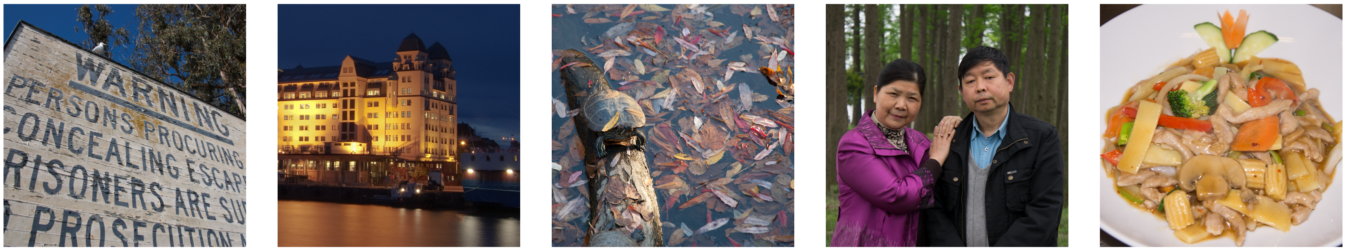

Fig. 3. Example of scenes that were assessed differently before and after priming.

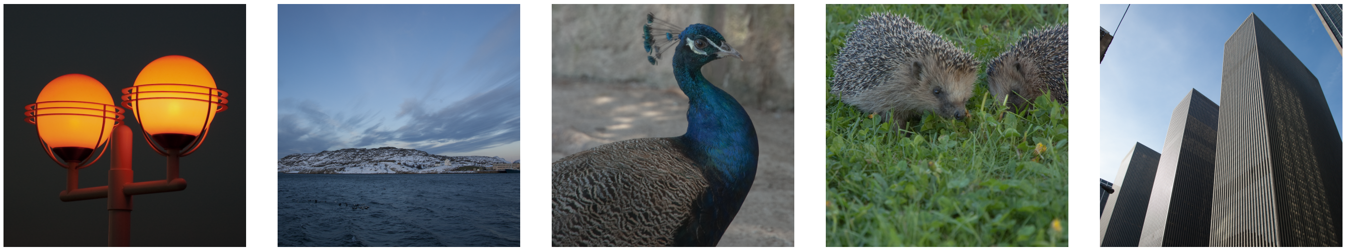

Fig. 4. Example of scenes that were NOT assessed differently before and after priming.

3.3. Can current visual masking models predict change blindness?

In order to further demonstrate the lack of existing models to predict high-level visual masking in image fidelity assessment, we tested the ability of several near- and supra-threshold predictive models (so-called ”image quality metrics”) to predict results from the two sessions. Results are reported in Table 1.

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. Top three metrics are supra-threshold while the others are near-threshold.

		Session 1	Session 2
Supra.	FSIMc [20]	**0.799**	0.806
	MS-iCID [21]	0.384	0.443
	VIF [22]	0.542	0.579
Near.	MAD* [12]	**0.805**	**0.891**
	PSNR-HA [23]	0.352	0.377
	HDR-VDP2.2 [24]	0.113	0.079

A metric that accounts for change blindness should achieve a better correlation with results from Session 1, which is not the case for any metric tested here. Both supra- and near-threshold models yielded a correlation with primed results that is at least as high a correlation as without priming, and even significantly larger for the Most Apparent Distortion metric (MAD). Incidentally, the latter also achieves the best predictions overall. These results indicate that state-of the art metrics perform poorly at predicting high-level masking and that further efforts are necessary in order to achieve higher prediction accuracy.

4. CONCLUSIONS

Despite recent advances in subjective image quality research, that topic is still the source of many fundamental questions. We looked at one particular limitation of our visual system that originates in that channel: high-level visual masking, which is probably best illustrated by the change blindness paradigm. We carried out a user study designed specifically to measure its influence and separate it from that of other types of visual masking and subjective biases. Results suggest a significant influence of high-level masking in image fidelity assessment at the 95% confidence level for half of the participants, with a magnitude over three times that of intra-observer variations.

These results are particularly important for future research in visual quality assessment, be it for images, video or computer graphics.

5. REFERENCES

[1] C. Cakebread, “People will take 1.2 trillion digital photos this year thanks to smartphones,” *Business Insider*, 2017.

[2] 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.

[3] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017.

[4] R. Streijl, S. Winkler, and D. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[5] 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.

[6] 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.

[7] K. Cater, A. Chalmers, and C. Dalton, "Varying rendering fidelity by exploiting human change blindness," in *Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, Melbourne, Australia, February 2003, ACM, pp. 39–46.

[8] E. Suma, S. Clark, D. Krum, S. Finkelstein, M. Bolas, and Z. Warte, "Leveraging change blindness for redirection in virtual environments," in *Virtual Reality Conference (VR), 2011 IEEE*. IEEE, 2011, pp. 159–166.

[9] 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.

[10] M. Alam, K. Vilankar, D. Field, and D. Chandler, "Local masking in natural images: A database and analysis," *Journal of vision*, vol. 14, no. 8, pp. 22–22, 2014.

[11] Y. Zhang, M. Alam, and D. Chandler, "Visually lossless perceptual image coding based on natural-scene masking models," in *Recent Advances in Image and Video Coding*. InTech, 2016.

[12] E. Larson and D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011006, 2010.

[13] J. Freeman and E. Simoncelli, "Metamers of the ventral stream," *Nat. Neurosci.*, vol. 14, no. 9, pp. 1195–1201, 2011.

[14] 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.

[15] Y. Liu and J. Allebach, "Near-threshold perceptual distortion prediction based on optimal structure classification," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 106–110.

[16] Á. Kristjánsson, "Priming of visual search facilitates attention shifts: Evidence from object-substitution masking," *Perception*, vol. 45, no. 3, pp. 255–264, 2016.

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

[18] X. Liu, M. Pedersen, and J.Y. Hardeberg, "CID:IQ - A New Image Quality Database," in *Image and Signal Processing*, pp. 193–202. Springer, 2014.

[19] S. Le Moan and M. Pedersen, "Evidence of change blindness in subjective image fidelity assessment," in *International Conference on Image Processing*. IEEE, 2017.

[20] 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.

[21] 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.

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

[23] 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.

[24] Manish Narwaria, Rafal K Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet, "Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *J. Electron. Imaging*, vol. 24, no. 1, pp. 010501–010501, 2015.