SUBJECTIVE IMAGE FIDELITY ASSESSMENT: EFFECT OF THE SPATIAL DISTANCE BETWEEN STIMULI

Steven Le Moan

Massey University
Palmerston North, New Zealand

Marius Pedersen

Norwegian University of Science and Technology
Gjøvik, Norway

ABSTRACT

Understanding how we perceive visual quality is important in a range of applications such as streaming or cross-media reproduction. Despite current perception models showing high correlations with recorded mean opinion scores, the factors influencing visual quality are still not fully understood, particularly when it comes to memory. We designed and carried out a study to compare quality assessment for two different levels of reliance on visual short-term memory. We found that assessments based mostly on memory tend to be more positive for compression, blur or gamut mapping distortions. Our results further highlight the role of memory in subjective quality assessment and visual masking.

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

1. INTRODUCTION

Subjective visual quality (SVQ) research is concerned with how people see and interpret digital images. It is particularly important in such fields as coding and cross-media reproduction. Although most people tend to agree on what defines bad quality (e.g. strong blocking artifacts or large contrast reductions), we do not yet fully comprehend the most advanced parts of our visual system and how they influence our judgment of visual quality. Memory, attention or awareness [1] are so idiosyncratic that their inner workings are difficult to study. While state-of-the-art SVQ models often are visual attention-aware, they rely for the most part on salience. However, salience accounts only for the tip of the iceberg when it comes to attention. It is one thing to recognise what catches our eyes almost unconsciously, but another to predict visual search and inspection strategies for specific tasks [2]. Even though SVQ models can reach linear correlations with mean opinion scores higher than $\rho = 0.95$ [3, 4] on popular benchmarks, there is still a lot to learn about how to properly design these databases, i.e. the experiments they are based on and the interpretation of their results [5]. There are indeed four main questions in visual quality perception:

- What *can* we see? The human eyes filter out a great deal of visual information due to spatial, spectral and temporal bandwidth limitations. We refer to these effects as **low-level**, because they can prevent the conscious perception of visual artefacts, even when the observer knows where they are.

- What do we *expect* to see? While the sensitivity of our eyes is very limited spatially, they can be moved around, guided by visual attention mechanisms. These mechanisms work differently based on the task at hand and they involve both bottom-up and top-down effects. Without priming, they can be quite inefficient in the sense that we can "miss" important visual information even at the centre of gaze [6]. We then refer to these type of perceptual limitations as **high-level**, because they can be overcome by priming.

- What do we *think* we see? Conscious perception of visual information can be further distorted by expectation and emotion. It has been demonstrated [7] that observers can report perceiving differences between two absolutely identical images, even on a display calibrated for spatial homogeneity of brightness.

- What do we *make of* what we think we see? The final step, which lies beyond the visual processing pathway, is the making of a decision.

While the first question is well studied from the perspective of early vision, only a few recent studies have tackled the latter three, like for instance Bosse *et al.*'s investigations of the neural correlates of visual quality [8]. However, with the advent of online streaming and novel imaging technologies (3D, multi-spectral, high dynamic range, light-field), as well as the increasing availability of eye-tracking and brain monitoring devices, a deep understanding of what constitutes visual quality is primordial.

Generally, the main challenge is predicting whether artefacts due to the reproduction process (coding, gamut mapping, noise, blur...) are perceptible to the extent that they influence our overall judgment. One key paradigm that has been exploited in this context is masking, which can be of two kinds [9]: *low-level* masking prevents the perception of differences between stimuli even though we know where they

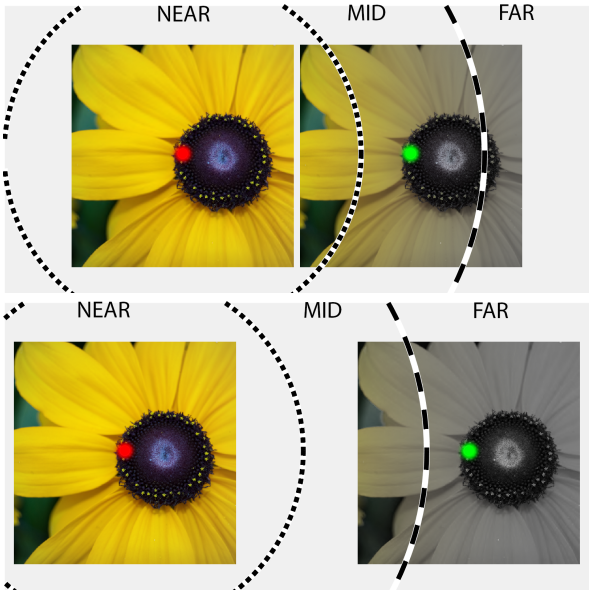

Fig. 1. Peripheral vision in a pair comparison experiment for a viewing distance of 50 cm (top: short distance between stimuli, bottom: large distance). The red dot corresponds to the fixation point in the left stimulus, whereas the green dot corresponds to the same location in the other (unattended) stimulus. Approximate boundaries between the near- mid and far-peripheral regions of the visual field are marked with the dashed circles (at 30° and 60°). The de-saturation illustrates the loss of perception in peripheral vision. In the case of a small distance (23°), the unattended dot lies within mid-peripheral vision and even a part of the unattended image is within near-peripheral vision. In the case of the large distance (34°), the unattended dot lies within far-peripheral vision, where shape, texture and colour perception is the poorest.

are, while *high-level* masking prevents the perception of differences until we know where they are. Here again, the latter is far less understood, as state-of-the-art models of memory, internal representations and visual search strategies have only marginally been applied to SVQ so far.

Here, we look at a particular branch of SVQ: image *fidelity*, otherwise known as *full-reference* (FR-SVQ). In this context, quality is measured by means of a comparison between two reproductions of the same image (typically, one is a pristine version and the other is distorted). In practice, this scenario can occur for instance in printing where soft and hard copies are typically compared side-by-side, when transferring a picture taken with a smartphone to a computer, or to analyse a network in term of how it changes the quality of the input (e.g. for video streaming). A typical setup for FR-SVQ user studies is pair comparison: the two stimuli are displayed simultaneously on a monitor and observers are asked to rate the difference between them, in terms of quality. There are several factors that can influence these ratings. While the effect of most physical (viewing distance, ambient luminosity, screen resolution, etc) and some psycho-physical (contrast sensitivity, chromatic adaptation, etc) factors are well studied, little is known about how the highest levels of visual processing affect our perception of visual quality, and in particular about encoding, storage and access in visual short-term memory. Proven limitations to our ability to compare complex patterns (see literature on texture masking [10]) hint at the influence of memory in FR-SVQ, though most recent works have focused only on information encoding and processing in areas V1, V2 and V4 of the visual cortex [11, 12].

In a recent study [13], we compared results obtained when stimuli are shown side-by-side on the display (standard case) to those obtained when they are shown one after the other, but not at the same time. In the latter case, observers had to rely exclusively on visual memory to compare the stimuli, which resulted in significantly higher quality ratings overall. While low-level effects such as luminance adaptation can also justify this result, we argue that it is mostly due to a conscious decision to commit only some attributes/statistics of the scene to short-term memory during the first display. If not primed, this decision is likely to be sub-optimal, leading to the observer missing important visual differences. Psychological factors such as alertness or confidence (see Box 2 in [1] or the work of Levin *et al.* on change blindness *blindness* [14]) are likely to contribute marginally to the difference in ratings, though we do not discuss them here.

In this paper, we propose a different experimental design to capture the effect of visual memory in FR-SVQ assessment. Specifically, we look at the effect that the spatial distance between stimuli has on subjective ratings. Some of the most widely used FR-SVQ databases such as CSIQ [10] or TID2013 [15] are based on experiments where the spatial distance between stimuli was set in a somewhat *ad hoc* manner. However, this distance (or gap) can influence how much of one stimulus can be seen when looking at the other, as illustrated in Figure 1. For short gaps and relatively low-resolution images, as it is the case for the aforementioned benchmarks, fixating a region in one stimulus means that the corresponding region in the other stimulus roughly lies in near-peripheral vision (above 18° of the visual field). Even though visual acuity and chromatic perception are already very poor and decrease rapidly [16, 17], they get substantially worse in mid-peripheral vision (above 30° of the visual field). Furthermore, due to the distance between stimuli, our eyes need to travel slightly longer, thereby imposing more constraint on memory as it must hold on to the internal representation longer. Mirjalili *et al.* [18] investigated the effect of separation between colour patches on color difference perception. In their experiment, the patches had no separation, 1 pixel, 2 pixels and 140 pixels distance. They found that separation between the patches has an impact on the colour difference. Some results [19] even suggest distinct peripheral sensitivities for the

Table 1. Control and test comparison between groups.

	Group 1	Group 1	
Set 1	short	short	Control
Set 2	long	long	
Set 3	short	long	Test
Set 4	long	short	

blue-yellow and red-green colour opponent channels of our visual system. Freeman *et al.* [12] looked at the transformation between the primary visual cortex and the inferotemporal cortex (ITC), where mid-level attributes such as faces are encoded. The ITC is considered the last component of the ventral stream [11], otherwise known as the "*What*" pathway (as opposed to the dorsal stream, i.e. the "*How/Where*" pathway). Their synthetic metamers, produced based on a bank of oriented bandpass filters, reveal how surprisingly little we perceive in peripheral vision. Therefore, setting up a large gap between stimuli forces observers to rely even less on peripheral vision and more on memory in judging image fidelity.

2. USER STUDY

2.1. Participants

A total of 28 people with normal or corrected-to-normal vision participated in the study (12 in New Zealand, 16 in Norway). Colour vision was tested for each participant with an Ishihara test. Ages ranged between 25, 66% were male and various cultural backgrounds were represented. No one was given any indications as to the goals of the experiment prior to it. A standard screening [20] revealed that all participants were valid.

2.2. Stimuli

Stimuli were selected from the CID:IQ database [21] and were displayed in pairs. Each pair was made of A) one of 22 different pristine images and B) one of six distorted versions of the same image. The distortion types were JPEG2000, Gaussian blur and SGCK gamut mapping. For each type, two near-threshold levels were chosen [1], with higher distortion level corresponding to the lowest quality. Original stimuli were 800×800 pixels in size, but we re-formatted them to 730×730 in order to accommodate distances larger than 30° distance between them on a 1920 pixel-wide display.

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 [20] 5-level scale was provided ("Not perceptible", "Perceptible, but not annoying", "Slightly annoying", "Annoying" and "Very Annoying"). After entering their name and age, each participant was shown two examples of image pairs to familiarise themselves with the task. 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 experiment was then divided into two sessions: one where the spatial distance between stimuli was set to $d_1 \approx 23°$ (centre-to-centre) of the viewing field and for the other, it was set to $d_2 \approx 34°$ (as seen in Figure 1). These values correspond respectively to a 20 and 400 pixel-wide gap between the stimuli. Which session came first was chosen randomly for each observer. Each session featured two sets of five scenes from the remaining 20. Set 1 featured in the "short-distance" session of all observers, while Set 2 featured in all the "long-distance" sessions. The remaining two sets (3 and 4) were organised so that half of the participants were displayed pairs with a gap of d_1 and the other half at d_2. This strategy then allowed us to compare the results of the two groups of observers in the case of an equal distance (control) as a case of different distance (test). Our null hypothesis is that results from the two groups of observers are not significantly different for sets 1 and 2 (case of equal distances) but are for sets 3 and 4, on account of the influence of the distance between stimuli. Figure 1 summarises the distribution. An angle of 30° from the centre of gaze is more that sufficient to induce severe impairment in texture and colour perception [12]. This implies that a 34° gap between 730-pixel wide stimuli may be enough to prevent the perception of e.g. compression artefacts in one stimulus when looking at the other. In view of this, we propose that an increased distance between stimuli compels observers to rely more on their visual working memory and that any difference in judgment between the two topologies (short and long gap) is more due to memory limitations than to decreased peripheral acuity.

The study lasted about 24 minutes on average.

2.4. Apparatus and 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. All stimuli were encoded in sRGB. Both experiments were carried out in a dark room. The distance to the screen was set to approximately 50cm (without chin rest). Additionally, we measured eye movements for all observers in New Zealand, with a Gazepoint GH3 HD at 60Hz.

3. RESULTS

3.1. Observer variability

Twenty percent of the stimuli in each session were repeated once for control. We found that the average intra-observer variability was 0.18 (i.e. about 3.6% of the assessment scale)

[1] These levels respectively correspond to levels 1 and 2 in CID:IQ.

in both sessions ($\sigma^2 = 0.17$, $\max = 0.58$ in session 1 and $\sigma^2 = 0.15$, $\max = 0.58$ in session 2).

Inter-observer variability (IOV) was measured as the standard deviation over each set of ratings of the same image pair with the same gap between them. On average, we found no significant difference in IOV between the two sessions: 0.82 ($\sigma^2 = 0.23$) in session 1 and 0.80 ($\sigma^2 = 0.25$) in session 2. This amounts to about 16% of the assessment scale.

These measured variabilities are consistent with results obtained in [22].

3.2. Comparison of ratings associated with different distances between stimuli.

Figure 2 shows the difference in average ratings in the case where both groups were shown the stimuli at the same distance (left) and at a different one (right). It can be seen that increasing the gap between stimuli tend to result in systematically higher quality scores (note: a high rating means a low quality). A sign test, a signed rank test and a paired t-test, each at the 95% confidence level, give that the same-distance difference has a median value not significantly different from 0. In other words, the two groups agree in their ratings. The same tests in the different-distance case (sets 3 and 4) reveals a significant shift of the median rating, thus indicating a significant influence of inter-stimulus spatial distance. The JPEG2000 and Gaussian blur distorted images seem to induce the most consistent results across scenes and levels, while the difference observed on gamut-mapped images is less obvious. In fact, the signed rank test performed on each distortion type individually reveals that only the former two are perceived significantly differently based on inter-stimulus distance at the 95% confidence level, but only at the 90% level for gamut mapping distortions. This is consistent with previous observations that orientation changes are harder to perceive than colour changes in the periphery [23].

Our results show that, by relying more on visual memory, observers tend to find less perceptible differences between stimuli, which affects their ratings. Further studies will need to clarify whether this is a problem of encoding, access or comparison and to determine the role of psychophysical effects like local adaptation (luminance-wise and chromatic), confidence and alertness. For example, it may be that luminance adaptation at a fixated region reduces the probability of detecting a reduction in contrast based on the remembered region. That would be a case of low-level masking since luminance adaptation is not modulated by attention at a given fixation point. However, a probably more significant reason for the different ratings is the access to a limited amount of reference information from memory to compare the attended region with. Under this view, full-reference SVQ is, in fact, a case of reduced-reference SVQ. Finally, it is also likely that reliance on memory renders people less confident in their judgment and less likely to believe. Therefore, our results further demonstrate the effect of high-level masking in image fidelity assessment.

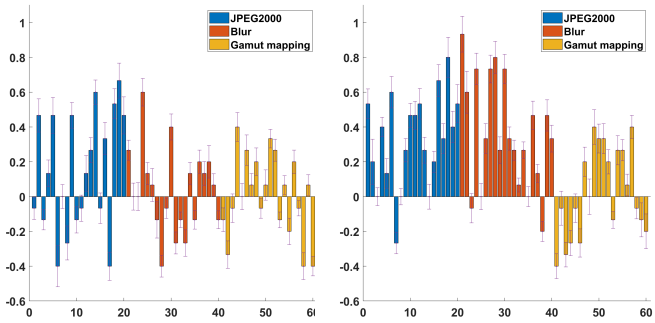

Fig. 2. Left: Difference of average ratings between groups 2 and 1 for sets 1 and 2 (case of same gaps). Right: Difference of average ratings between session 2 and session 1, for sets 3 and 4 (case of different gaps). Confidence intervals are given at 95%.

3.3. Analysis of the eye-tracking data

Given a fixation point in one stimulus, we estimated that the probability of the next fixation being the same point in the other stimulus was about 24.5% for session 1 and 29.2% for session 2 on average over all observers, thus indicating that the larger gap led to more back-and-forth eye movements. These values were measured by first selecting only saccades with length equal to the inter-stimulus distance (centre-to-centre), with a margin of error at 5°. Among these, saccades with an absolute vertical coordinate larger than 5° were discarded. This result clearly indicate that observers were overall less confident in session 2, as they had to rely more on visual memory.

4. CONCLUSIONS

In this paper, we looked at the influence of the spatial distance between stimuli on subjective image fidelity assessment. We found that increasing the centre-to-centre distance from mid-peripheral to far-peripheral vision led to significantly different quality ratings, with observers finding the difference between stimuli less annoying. Ratings for compressed and blurred image reproductions are particularly affected by inter-stimulus distance, which is in accordance with previous results suggesting that we are more sensitive to colour differences than shape or texture differences in peripheral vision. It also suggests that low-level texture masking is memory-dependent and that the bandwidth of the access to quality features within visual short-term memory is significantly more narrow than that within the ventral stream.

5. REFERENCES

[1] 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.

[2] C. Wloka, I. Kotseruba, and J. Tsotsos, "Saccade sequence prediction: Beyond static saliency maps," *arXiv preprint arXiv:1711.10959*, 2017.

[3] 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.

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[5] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on pattern analysis and machine intelligence*, 2018.

[6] 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.

[7] S. Le Moan, I. Farup, and J. Blahová, "Towards exploiting change blindness for image processing," *Journal of Visual Communication and Image Representation*, vol. 54, pp. 31–38, 2018.

[8] S. Bosse, L. Acqualagna, W. Samek, A.K. Porbadnigk, G. Curio, B. Blankertz, K.-R. Muller, and T. Wiegand, "Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[9] S. Le Moan and M. Pedersen, "Measuring the effect of high-level visual masking in subjective image quality assessment with priming," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3553–3557.

[10] 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.

[11] V Lamme, H. Super, H. Spekreijse, et al., "Feedforward, horizontal, and feedback processing in the visual cortex," *Current opinion in neurobiology*, vol. 8, no. 4, pp. 529–535, 1998.

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

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

[14] D. Levin, N. Momen, S. Drivdahl IV, and D. Simons, "Change blindness blindness: The metacognitive error of overestimating change-detection ability," *Visual Cognition*, vol. 7, no. 1-3, pp. 397–412, 2000.

[15] 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.

[16] J. Besharse and D. Bok, *The retina and its disorders*, Academic Press, 2011.

[17] A. Kvitle, M. Pedersen, and P. Nussbaum, "Quality of color coding in maps for color deficient observers," *Electronic Imaging*, vol. 2016, no. 20, pp. 1–8, 2016.

[18] F. Mirjalili, M. Luo, G. Cui, and J. Morovic, "A parametric colour difference equation to evaluate colour difference magnitude effect for gapless printed stimuli," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2018, vol. 2018, pp. 123–127.

[19] K. T Mullen et al., "Differential distributions of red–green and blue–yellow cone opponency across the visual field," *Visual neuroscience*, vol. 19, no. 1, pp. 109–118, 2002.

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

[21] 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.

[22] S. Le Moan and M. Pedersen, "Evidence of change blindness in subjective image fidelity assessment," in *International Conference on Image Processing*. IEEE, Sep. 2017, pp. 3155–3159.

[23] MPS To, ID Gilchrist, T Troscianko, and DJ Tolhurst, "Discrimination of natural scenes in central and peripheral vision," *Vision Research*, vol. 51, no. 14, pp. 1686–1698, 2011.