THE INFLUENCE OF SHORT-TERM MEMORY IN SUBJECTIVE IMAGE QUALITY ASSESSMENT

Steven Le Moan, Marius Pedersen, Ivar Farup, Jana Blahová

Faculty of Computer Science and Media Technology
NTNU - Norwegian University of Science and Technology
Gjøvik, Norway.

ABSTRACT

Aiming at understanding the role of short-term memory in subjective image quality assessment, we report and compare results from two pair-comparison methods: stimuli shown *side-by-side* versus stimuli shown *one after the other*. Our results suggest that there is a significant chance that an observer will make different quality assessments in the two setups.

Index Terms— Image quality assessment, Pair-comparison, Perception, Memory.

1. INTRODUCTION

The way we perceive differences between visual stimuli has been vastly studied in the context of Image-Quality Assessment (IQA). Given two versions of the same digital image (e.g. before and after compression), the purpose of objective IQA is basically to rate the difference between them, similarly to how a standard human observer would. Having a model that correlates with human judgment is particularly important in automatic quality-aware image processing for enhancement, compression, cross-media reproduction, etc. Most IQA models (see e.g. [1] for a recent review) are therefore designed to estimate what information (in the general sense) is conveyed from a given visual stimulus to the decision-making parts of the brain. In that matter, the notion of *internal representations*, i.e. how visual information is stored and processed in our brains [2], is of the essence. The way these representations are generated, their nature and the level of details that they contain depend, however, just as much on the light signal entering our eye and reaching our brain, than on factors like our expectations with respect to the scene/task, fatigue, memory, awareness or even cultural background. Despite the fact that popular models such as the Structural SIMilarity index [3] or the recent Visual Saliency Index [4] generate quality scores which correlate to a large extent to recorded subjective scores, very little is known about the role played by these high-level factors.

Phenomena such as change blindness [5] have suggested that internal representations are somewhat flawed [6], in that only partial information about a stimulus can be recovered (i.e. remembered) after observation. The game "spot the difference", in which relatively large differences between two visual stimuli can go unnoticed for significant periods of time [7], is a good example of this phenomenon. It implies that observers can only evaluate a limited set of image attributes at a time and therefore that some of these attributes are prioritised over others [8]. Attributes which are considered as important are checked first, resulting in a faster detection of modifications affecting the gist of a scene [9]. Low-level saliency can partly model this process [10, 11], but only to some extent [12], as attention does not necessarily imply awareness. When the two images to compare are not displayed at the same time, the phenomenon is even more likely to occur as one can then only rely on an internal representation (i.e. a memory) of the first stimulus when comparing it with the second one.

Recently, the notion of image memorability [13, 14] was introduced as a means to understand why some visual stimuli have a higher chance to be remembered than others, but only in the context of comparing different kinds of scenes. In this paper, we are interested in memorability but from the point of view of image quality. We aim indeed at assessing whether some image quality features are more memorable than others and how reliable internal representations actually are for subjective IQA. Let us take the example of a pair-comparison setup, in which two versions of the same image are shown side-by-side on a display and where an observer is asked to select the one that they believe has the highest overall quality. First of all, it is important to keep in mind that detailed vision is only available in a small portion of the visual field, roughly of the size of a thumbnail at arm's length. In other words, the observer cannot scrutinise both images simultaneously, as looking at one of them will automatically place the other one in peripheral vision field (i.e. substantially less detailed [15]). Therefore, short-term memory plays an important role here. It having a limited capacity, only a certain quantity of information from the two images can be compared at a time [8].

This work was partially funded by the Research Council of Norway through project number 221073: HyPerCept - Color and Quality in Higher Dimensions. Please address correspondences to: steven.lemoan@gmail.com.

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

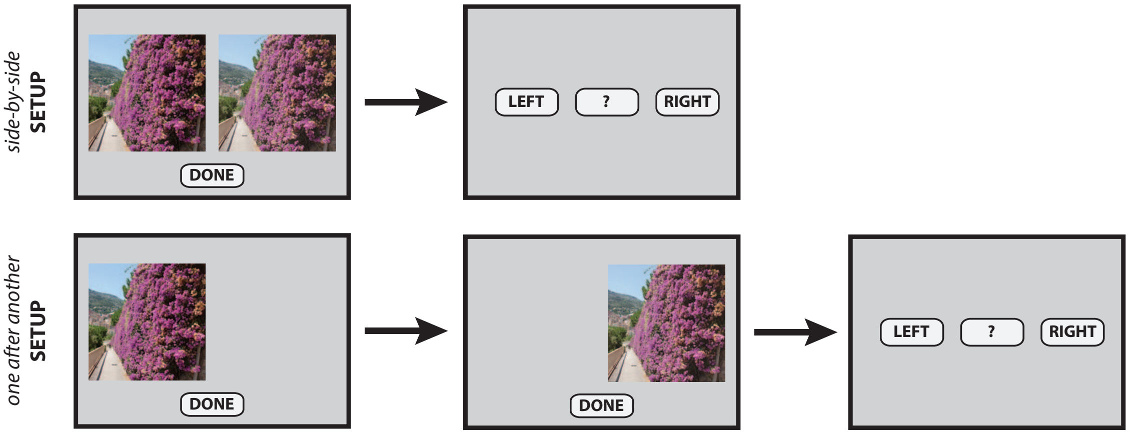

Fig. 1. Description of the two setups in the experiment. In the *side-by-side* setup, observers have access to both stimuli, whereas in the *one after another* setup, they have access to only one at a time.

However, since they are both displayed at the same time, the content of short-term memory can be "updated" at will until reaching a verdict as to the difference of quality between the stimuli. Now, if the stimuli are shown one *after* the other, the possibility of an "update" is removed. Is it then possible for observers to reach the same verdict? In a sense, this question pertains to assessing whether a *full-reference*[1] subjective IQA is equivalent to a *reduced-reference*[1] IQA where the reduced-reference is an internal representation.

In an attempt to answer this question, we report and analyse results from a pair-comparison experiment comprising two sessions: one using a *side-by-side* setup (as it is usually done in the literature), and one using a *one after another* setup (see Figure 1). In both cases, observers were asked to select the image which they believed had the highest quality, with the possibility of *tie* scores (meaning that the images were believed to be indistinguishable in terms of quality).

2. EXPERIMENTAL SETUP

2.1. Viewing conditions

We used an Eizo ColorEdge CG246W display (24.1" - 61cm), calibrated with an EyeOne software for a colour temperature of 6500K, a gamma of 2.2 and a luminous intensity of 80cd/m^2. The experiment was carried out in a dark room. A chin rest was used in order to ensure a viewing distance of 50cm for all observers.

2.2. Data

For our experiments, we selected a subset of 120 images from the CID:IQ database [16]: 10 scenes, 4 types of distortions (JPEG, Poisson noise, Gaussian blur and SGCK gamut mapping) and 3 out of 5 levels of distortions (levels 1, 3 and 5, the latter corresponding to the strongest distortion). These levels were chosen so as to avoid too small differences between stimuli. The undistorted images are depicted in Figure 2. Please refer to the original paper [16] for more details.

2.3. Observers

A total of 18 colour-normal observers participated to the experiment, 15 of which were considered as experts in image processing and/or colour science. Ages ranged between 22 and 53 and various cultural backgrounds were represented. Finally, seven observers stated that they were somewhat familiar with the CID:IQ database. None of them was given any indications as to the actual goals of the experiment. A screening based to the method described in [17] revealed that all observers were valid.

2.4. Task

As previously explained, the experiment comprised two sessions: one during which images were displayed *side-by-side*, i.e. simultaneously, and one during which they were displayed *one after another*, i.e. successively. Note that the same positions were used in both setups (i.e. one image on the left, one on the right), in order not only to be as consistent as possible, but also to avoid immediate change detection during the switch. Figure 1 illustrates the setups.

In the *side-by-side* session, observers could analyse both pictures as much as they wanted until reaching a verdict, at

[1] Note that there are different frameworks for IQA, depending on the availability of a reference image or not: the *full-reference* framework assumes that both stimuli are available, whereas the *reduced-reference* framework assumes that one stimulus is only partially available.

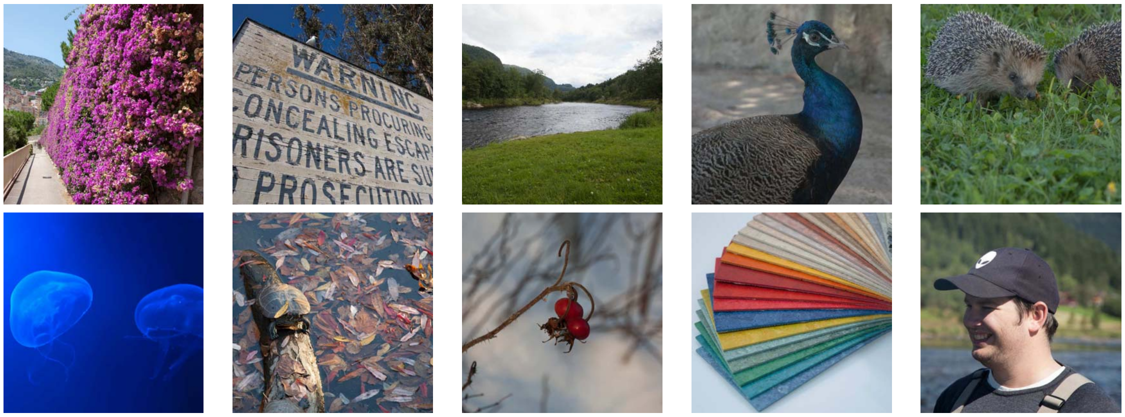

Fig. 2. The scenes used in our visual experiment, numbered from top left (scene 1) to bottom right (scene 10). The scene with the hedgehogs and the one with the turtles were considered as the most difficult ones to judge by our panel of observers, on account of their high level of details.

which point they were asked to click on a button leading to a "decision screen" where they were given the chance to chose between "left" (i.e. the image on the left had the highest quality), "right" (i.e. the image on the right had the highest quality) or "?" (i.e. it was not possible to select either one or the other image). In the *one after another* session, the images were displayed at the same locations but not at the same time. Observers were given as much time as they deemed necessary to look at the first image before moving to the second one. It is particularly noteworthy that in both sessions the observer's assessment was asked once the stimuli were no longer accessible and that there was no time limit. Note that image sequences were generated randomly and differently for each observer and each session. In particular, in the *one after another* session, the choice of which image was shown first (original or reproduction) was also random.

All observers participated to both sessions, with at least 24 hours between them. A randomly selected group of 9 observers started with the *side-by-side* session (Group 1) while the remaining 9 started with the *one after another* session (Group 2). The exact instructions given to the observers were as follows: *Decide which image has the highest quality. Once you make a decision click on "Done" and then specify "right" or "left". If you don't know, click on the question mark.* Finally, time was monitored during the experiment, and observers were informed of this fact.

3. RESULTS

In order to account for intra- and inter-observer variability, we computed for each image pair and in each session the mode of decisions (i.e. the decision taken by the majority of observers). We refer to the result as the decisions of a *representative observer* in the remainder of this section.

3.1. Did the two setups lead to different quality assessments?

To answer this question, let us first consider two types of discrepancies in terms of quality assessments between the two setups for a given image pair and for the representative observer:

- D1: different quality assessments in the two setups, one of them is a tie. This case implies that observers felt less confident about their ability to make an accurate quality assessment in one of the sessions than in the other.

- D2: different quality assessments in the two setups, neither is a tie. This case implies that for a pair of images A and B, if e.g. A was found of higher quality in the *side-by-side* session, then B was found of higher quality in the *one after another* session. Assuming that the assessment made in the *side-by-side* setup is the most accurate, then a D2-type discrepancy essentially means that the observers *over-estimated* their ability to make an accurate quality assessment in the other setup.

We applied the two-sample binomial test at 95% confidence [18] to evaluate whether the number of occurrences of each of these cases was statistically different from zero in our experimental data and found that D1-type discrepancies are not statistically significant. However, there is a significant difference between the proportions of tie scores given in the *side-by-side* and *one after another* sessions, for Group 1 (5% and 7%, respectively) and for Group 2 as well (7% and 10%, respectively). Incidentally, observers from Group 2 found significantly more ties than those from Group 1 in the *one after another* session. This is consistent with the intuitive idea that observers should be more confident in the *one after another* session if they first carried out the other one.

Table 1. Percentages of D2-type observer judgment discrepancies for the representative observer. All values are statistically different from zero according to the two-sample binomial test at 95% confidence. The star (*) indicates where the results from the two groups are significantly different.

	Group 1	Group 2
Overall	42%	33%
JPEG*	**50%**	20%
Poisson noise	30%	30%
Gaussian blur	33%	**47%**
SGCK gamut mapping	**57%**	37%

Table 1 reports the results for D2-type judgment discrepancies. These results show that for a given pair of images A and B, if A was found of higher quality in the *side-by-side* setup, then there is a significant chance that image B was found of higher quality in the *one after another* setup. This seems to stand true particularly for Group 1 in the case of JPEG and Gamut-mapping distortions (50% and 57% D2-type discrepancies, respectively) and for Group 2 in the case of blurred images (47%). We found that these proportions are relatively consistent across scenes, with the exception of scene 10 (human face), for which 67% of image pairs led to a D2-type discrepancy for Group 1, against 25% for Group 2. They are also consistent across distortion levels, with two noteworthy exceptions: gamut mapping level 2 vs level 3 led to proportions of 70% for both groups, and noise level 1 vs level 2 led to proportions not significantly different from zero, also for both groups. This essentially means that gamut mapping distortions are substantially more difficult to assess than Gaussian noise distortions if the original and distorted images are not shown side-by-side. Furthermore, the proportion of D2-type discrepancies is statistically different from zero for all ten scenes for Group 1, but only for six of them for Group 2 (all but scenes 2, 5, 8 and 10). Finally, we found a statistical difference between the two groups only in the case of JPEG distortion (all scenes considered) and for scene 10 (all distortions considered). These results are also consistent with the idea that observers should be more confident in the *one after another* session if they first carried out the other one.

It is well known that people tend to over-estimate their visual perception [8] as demonstrated for instance by change blindness experiments. Here we provide evidence that this could also stand true in tasks pertaining to subjective quality assessment. In other words, the process of assessing the quality of an image may be driven not only by perception, but also by assumption. Incidentally, this also suggests that internal representations (i.e. memories) of visual stimuli, are not always precise and/or relevant enough to serve as reliable reference in subjective image quality assessment. In other words, human don't always have the capacity to store in short-term visual memory enough information to reliably compare two images in terms of quality.

3.2. Time analysis

As previously mentioned, time was monitored during the experiment. Table 2 reports the statistics of each group in each session. Note that we considered the time given to rate the very first pair of each session as an outlier.

Table 2. Average times (in seconds). For the *one after another* session, the average time spent on each stimulus is given.

	Group 1	Group 2
side-by-side	**4.1**	**4.0**
one after another: first stimulus	5.0	7.7
one after another: second stimulus	3.3	5.3
***one after another*: total**	**8.4**	**13.1**

In order to assess whether more time was needed in one of the sessions than in the other, we performed the two-sided sign test. We found that, for Group 1, more time was taken in the *one after another* session for about 88% of the pairs, whereas for Group 2, more time was taken in 96% of the cases. Note that there seems to be no significant variations between the different types of distortions. Additionally, a two-sample Kolmogorov-Smirnov test revealed that the proportion of pairs for which a higher decision time was required in the *one after another* session was significantly different between the two groups. This means that if the *side-by-side* session was performed first, observers were faster at finishing the other one.

4. CONCLUSION AND FUTURE WORK

Aiming at understanding the role of short-term memory in subjective image quality assessment, we reported and compared results from two pair-comparison methods: stimuli shown *side-by-side* versus stimuli shown *one after another*. Our main goal was to answer the question: "Can one reach the same verdict in both sessions?". Results suggest that most observers could not. In particular, we found that they tended to significantly over-estimate their ability to make an accurate quality assessment in the *one after another* session, especially if they carried out the *side-by-side* session *a priori*. The significance of these results needs, however, to be ascertained with more experiments of this kind, involving more types and levels of distortions.

5. REFERENCES

[1] Marius Pedersen and Jon Yngve 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.

[2] Timothy F. Brady, Talia Konkle, and George A. Alvarez, "A review of visual memory capacity: Beyond individual items and toward structured representations," *Journal of vision*, vol. 11, no. 4, pp. 1–34, 2011.

[3] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] Lin Zhang, Ying Shen, and Hongyu Li, "VSI: A visual saliency induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[5] Christopher G. Healey and James T. Enns, "Attention and visual memory in visualization and computer graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 7, pp. 1170–1188, 2012.

[6] D. Alexander Varakin, Daniel T. Levin, and Krista M. Collins, "Comparison and representation failures both cause real-world change blindness," *Perception*, vol. 36, no. 5, pp. 737–749, 2007.

[7] Steven Le Moan and Ivar 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.

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

[9] Kirsten Cater, Alan Chalmers, and Colin 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.

[10] Zhou Wang and Qiang Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[11] Chenlei Guo and Liming Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[12] Jonathan A. Stirk and Geoffrey Underwood, "Low-level visual saliency does not predict change detection in natural scenes," *Journal of vision*, vol. 7, no. 10, pp. 3, 2007.

[13] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "What makes an image memorable?," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 145–152.

[14] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Memorability of image regions," in *Advances in Neural Information Processing Systems*, 2012, pp. 305–313.

[15] Jeremy Freeman and Eero P. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[16] Xinwei Liu, Marius Pedersen, and Jon Yngve Hardeberg, "CID:IQ - A New Image Quality Database," in *Image and Signal Processing*, pp. 193–202. Springer, 2014.

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

[18] Lawrence Brown and Xuefeng Li, "Confidence intervals for two sample binomial distribution," *Journal of Statistical Planning and Inference*, vol. 130, no. 1, pp. 359–375, 2005.