HIGH-LEVEL VISUAL MASKING OF IMAGE COMPRESSION ARTEFACTS

Steven Le Moan[†] *Marius Pedersen*[⋆] *Aladine Chetouani*[◇]

[†]Department of Mechanical and Electrical Engineering, Massey University
[⋆]Department of Computer Science, Norwegian University of Science and Technology
[◇] PRISME Laboratory, University of Orléans

ABSTRACT

We present the results of a subjective experiment where we measured detection thresholds for 2°-wide noise targets placed in 23 different natural scenes. Unlike previous studies on visual masking, we focus particularly on dissociating cases of low-level and high-level masking. That is, cases where the target is not perceived predominantly due to limits of either early or late vision. To that end, we exploit the change blindness paradigm and analyse detection rates, times and primed subjective ratings of target visibility. Our results are of significance for developing advanced models of human vision for signal quality/fidelity assessment, particularly in the context of compression.

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

1. INTRODUCTION

Being able to predict what people perceive as a good image quality is an important objective for such applications as compression or cross-media reproduction. However, despite decades of research in understanding the Human Visual System (HVS), the perception and interpretation visual signals by humans is still not well understood. Generally, the problem can be summarised in four main research questions:

- What *can* we see?
- What do we *expect* to see?
- What do we *think* we see?
- What do we *make of* what we think we see?

Most research efforts to date have been towards answering the first question. Indeed, current approaches to predicting subjective assessments of visual quality are mostly based on modelling the eye and the primary visual cortex, i.e. the *early* visual system. However, in a recent series of publications [1, 2], we demonstrated a significant effect of higher-level mechanisms of visual working memory in masking effects. We have proposed that visual masking can be of two distinct kinds: low-level (masking occurs in early vision and the masked target cannot be seen *even though* one knows where it is) and high-level (masking occurs beyond early vision and the masked target cannot be seen *until* the observer knows where it is). The latter can be overcome by focused attention, but not the former.

Even if state-of-the-art models (see e.g. [3, 4]) predict Mean Opinion Scores (MOS) on popular benchmarks with high accuracy, we are still not able to answer the aforementioned questions with certainty, particularly the last three. We believe that understanding what one *can* see is only part of what is required to build a robust model of subjective quality/fidelity assessment. While low-level masking comes from limits of early vision, high-level masking comes from limits in attention and memory. Our limited memory capacity is the main reason why we require attention mechanisms, to select and prioritise visual attributes so they can be processed by decision-making mechanisms. If we had unlimited working memory, one could argue that we would not need attention as all the information in our visual field could be instantly integrated and processed. Instead, it has been proven that limits in our memory capacity renders us incapable of efficiently comparing complex patterns (see literature on crowding [5] and texture masking [6]) suggesting a significant influence of memory in quality assessment, though most recent works have focused only on information encoding and processing in areas V1, V2 and V4 of the visual cortex [7, 8].

Predicting the visibility of artefacts (compression, gamut mapping, noise, blur...) and how they influence our overall judgment remains a challenge. When the artefacts are not salient, we talk about a near-threshold distortion. In such case, observers engage in a visual search to find these artefacts [6]. We previously hypothesised similarities between this kind of strategy in image fidelity assessment and in the change blindness paradigm [9]. Change blindness, as a type of high-level masking, can then be harnessed for example to significantly increase quantisation in complex image regions and save in size, without compromising visual quality on account of high-level masking.

In this paper, we propose a new experimental design to measure the respective effects of high- and low-level visual masking in subjective fidelity assessment for compressed images. We present the results of a subjective experiment in

978-1-7281-6395-6/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Norges Teknisk-Naturvitenskapelige Universitet. Downloaded on February 12,2021 at 12:04:08 UTC from IEEE Xplore. Restrictions apply.

which observers had to locate a single noise target (a compression artefact) in a scene, by comparing it to the pristine image. Our approach is similar to that used by Alam et al. [10]. However, we used a change blindness paradigm, real compression artefacts as noise targets and, more importantly, we measured and analysed the time required for people to notice the target (allowing for a maximum of 30 seconds). Our results demonstrate that detection times are moderately consistent across observers and correlated with primed subjective ratings of the target visibility.

2. USER STUDY

2.1. Participants

A total of 10 people with normal or corrected-to-normal vision participated in the study in Palmerston North, New Zealand. Colour vision was tested for each participant with a Ishihara test. Ages ranged between 22 and 34, 80% were male and various cultural backgrounds were represented. No one was given any indications as to the research goals of the experiment before it. A standard screening [11] revealed that all participants were valid (no outliers based on detection times). Participants all signed a consent form and were compensated with a supermarket voucher worth 10 NZD.

2.2. Stimuli

Stimuli were selected from the CID:IQ database [12] and were displayed in pairs. Original stimuli were 800×800 pixels in size. Following the change blindness paradigm, each pair was seemingly identical, but in fact there was a significant change: a single noise target (a compression artefact) was blended into one of them. The artefacts were sampled from heavily distorted (level 5 in CID:IQ) JPEG and JPEG2000-compressed images, over a circular region with a diameter corresponding to about 2° of the visual field. The spatial location of the centre of the circular region was sampled randomly from the final feature map of the MSiCID fidelity measure [13]. On this map, higher pixel values correspond to larger perceived differences across five scales (as in the MS-SSIM index) and in terms of brightness, contrast, structure, chroma, hue, chroma-constrast and chroma-structure. Regions of large estimated difference were then more likely to be selected for artefact sampling.

The targets were then blended into the pristine image by means of a Gaussian kernel to create a seamless blending. The kernel bandwidth was set to 1° of the visual field (see example in Figure 1). For each scene, four artefact-ridden versions were created: two with JPEG artefacts (at different random locations), two with JPEG2000 artefacts (also at different random locations). These four stimuli were the same for each observers. A total of 92 stimuli were created (23 scenes).

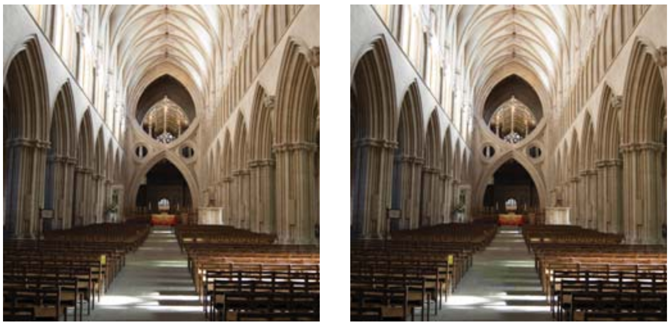

Fig. 1. Pristine (left) and artefact-ridden (right) images. The target (a JPEG artefact) is in the alley, in the lower central region. The reader may need to magnify this figure in order to notice it.

2.3. Apparatus and viewing conditions

We used an Eizo ColorEdge CG2420 display, of size 61cm/24.1” and calibrated with an X-Rite Eye One spectrometer for a colour temperature of 6500K, a gamma of 2.2 and a luminous intensity of 80cd/m^2. All stimuli were encoded in standard RGB. The experiments were carried out in a dark room. The distance to the screen was set to 60 cm.

2.4. Methodology

Participants were asked to locate the artefact and to click on it as soon as they saw it. They were informed that they could click only once and that they had 30 s to do so for each pair. If they clicked within the 2°-diameter circle, we recorded a true positive (TP) case. If they clicked outside, we recorded a false positive (FP) case. Otherwise (no click), we recorded a negative (N) case. The pristine and artefact-ridden stimuli were shown one at a time, with a flicker paradigm (0.8 s on, 0.1 s off). Once a click was recorded or 30 s passed, the solution was displayed by removing the 0.1 s blank interval and highlighting the location of the target with a red circle. At that point, observers were asked to rate the visibility of the target as either ’Clearly visible’, ’Barely visible’ or ’Invisible’. This allowed us to measure the extent of low- and high-level visual masking: if a target is rated as ’clearly visible’ even though it was not detected within 30 s, it indicates a strong high-level masking effect. On the other hand, if a target is rated as ’invisible’ (and was not detected within 30 s), it indicates a low-level masking effect.

After each sequence of 5 pairs of stimuli, participants were given the opportunity to take a short break in order to reduce visual fatigue, particularly due to the flickering. They just had to press a key to carry on at their convenience. The sessions were supervised and lasted about 46 minutes on average (including breaks). The sequence of stimuli was completely randomised for each observer. The very first image seen by each observer was considered a trial and removed

Authorized licensed use limited to: Norges Teknisk-Naturvitenskapelige Universitet. Downloaded on February 12,2021 at 12:04:08 UTC from IEEE Xplore. Restrictions apply.

from the results.

3. RESULTS

3.1. Features

We extracted 11 features from each stimulus pair in order to facilitate the analysis of the collected data. These features are noted F1 to F11 and are defined as follows:

- F1: Entropy of pixel values in the CIELAB colourspace (three-way entropy),
- F2 and F3: Achromatic edge energy (global and local, respectively), calculated on the $\mathrm{L}^\star$ channel of CIELAB as the average number of zero-crossings after filtering with a Laplacian of Gaussian filter,
- F4 and F5: S_3 sharpness [14] (global and local, respectively),
- F6 and F7: Local salience, with the IKN [15] and SDSP [16] models, respectively,
- F8: MSiCID [13] local predicted similarity across scales and features (MSiCID is a colour image difference measure based on the SSIM index, with 7 features: lightness, contrast, structure, hue, chroma, chroma-contrast and chroma-structure),
- F9: Salience imbalance (between pristine and artefact-ridden stimuli) calculated as the Hamming distance between signature maps as in [17] and [18],
- F10: Salience imbalance calculated as RMSE between SDSP maps,
- F11: Deep learning-based prediction of visual masking thresholds from a modified pre-trained VGG16 architecture [19]. The model predicts the detection thresholds for patches of size 80x80x3 pixels. In this study, we calculated the predictions in patches centered at the center of the target.

These features pertain to the complexity of the mask as well as the salience and magnitude of the target. The local features were extracted using only pixels within the 2°-wide circle around the target. Note also that the results we obtained revealed no significant correlation between the position of the target and detection times/accuracy.

3.2. Inter-observer variability

Looking at the proportion of observers who detected the target in a given stimulus s, noted $\mathrm{p}_d(s)$, we found that it varied significantly across stimuli. On average, 51.3% of participants detected it (mode: 73.2%, min: 0%, max: 100%). For 10.9% of the stimuli, all observers found the target whereas, in 13.0% of the cases, none of them found it. This indicates that there was no variability in terms of detection accuracy for 23.9% of the stimuli. On the other hand, this variability was highest for 17.4% of the stimuli where $\mathrm{p}_d(s)$ was between 40% and 60%, with 50% being the case of highest variability where only half of the observers found the target. We also determined that $\mathrm{p}_d(s)$ correlates positively with F8, F10 and F11 (linear correlation coefficients: 0.50, 0.54 and 0.58 respectively). This shows that these three features predict well the probability of an observer noticing the target.

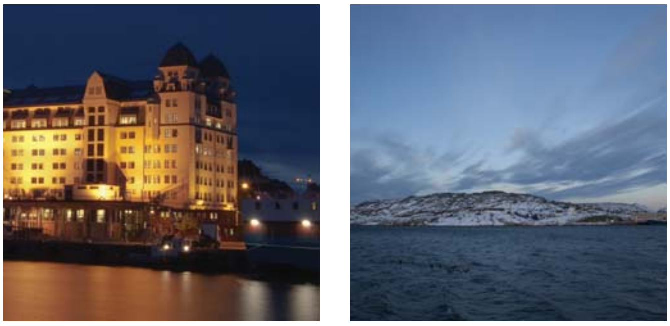

Fig. 2. Scenes with slowest (left) and fastest (right) average detection times.

In terms of detection times, inter-observer variability was generally significant, even though the time required to gaze at the location of the change can vary greatly from person to person due to the idiosyncratic and fundamentally stochastic nature of eye movements. The largest linear correlation found between detection times of any two participants was 0.61 (min: 0.21, mean: 0.43, mode: 0.52). The largest Spearman rank order correlation was 0.66 (min: 0.21, mean: 0.43, mode: 0.45). Furthermore, for 17% of the stimuli, the standard deviation of detection times was smaller than or equal to 5.0s. Overall, we measured a minimum of 1.1s, an average of 7.6s (mode: 8.8s) and a maximum of 12.0s. We also found that the distribution of detection times is non-normal in 52% of cases (Anderson-Darling test, 95% confidence). This indicates that the sample mean of decision times may not be the most representative statistics to analyse and predict our data. Consequently, we elected to use the mode, rather than the mean, to represent the distribution in the next section as in [18].

3.3. Inter-scene variability

Figure 2 shows the images that resulted in the slowest and fastest detection times overall. Respectively, their average detection time were 25.4s ($\sigma^2 = 4.1$) and 12.0s ($\sigma^2 = 3.2$). We found a maximum linear correlation of 0.85 between the detection times of any two scenes (min: -0.57, mean: 0.09, mode: -0.02). This shows that there is little consistency in detection times based on the scene only. Generally, verbal feedback from participants after the experiments indicated that scenes that were more 'visually complex' made it more dif-

Authorized licensed use limited to: Norges Teknisk-Naturvitenskapelige Universitet. Downloaded on February 12,2021 at 12:04:08 UTC from IEEE Xplore. Restrictions apply.

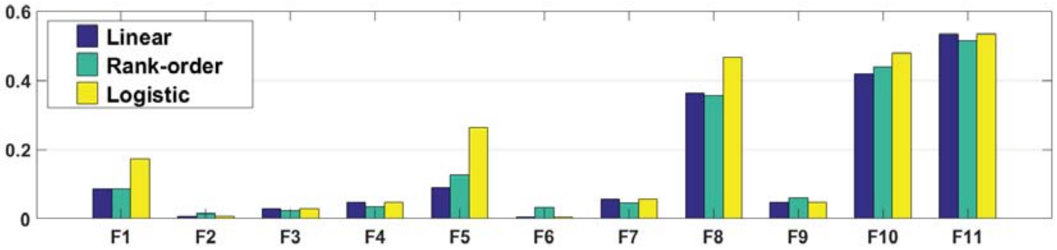

Fig. 3. Correlation of extracted features with mode detection times: Pearson linear, Spearman Rank Order and Pearson linear after logistic mapping of the predictions to the data.

ficult to detect the targets.

3.4. JPEG vs JPEG2000

On average over all observers, 67% of JPEG artefacts were detected as opposed to only 35% of JPEG2000 artefacts. Average time for JPEG artefacts: 15.7 s ($\sigma^2 = 6.6$). Average time for JPEG2000 artefacts: 26.0 s ($\sigma^2 = 5.8$). In assessing these results, one must keep in mind that the CID:IQ database was not designed so that distortion levels are equivalent in terms of visibility across distortion types. In other words, a level 5 JPEG distortion may not have the same visibility than a level 5 JPEG2000 distortion. However, JPEG2000 compression is known to create artefacts that are generally considered less ”annoying” than the blocking and other structural degradation induced by JPEG artefacts. Our results confirm this.

3.5. Subjective evaluation of visibility

The subjective ratings were collected after priming the participant, i.e. after that the solution was shown. Participants were specifically instructed to make their judgement not based on the time it took them to find the target, or whether or not they did find it, but solely on their opinion *after* being primed.

The minimum and maximum standard deviations of subjective ratings were found to be 0 and 0.83 (27% of the assessment scale), respectively. On average, this variability was found to be 0.44 (15% of the assessment scale), which indicates a significant agreement between observers. For 98% of stimuli, the distribution of subjective ratings was found to be non-normal (Anderson-Darling test, 95% confidence).

The average linear correlation with detection times was 0.52, which indicates a moderate, yet significant correlation between target visibility before and after priming. More specifically, for 11.2% of stimuli with a target rated as ”clearly visible” and including all observers, we recorded either no detection or a false positive detection. In 12.6% of the same cases, we recorded a true positive with a detection time larger than 20s. These cases of strong high-level visual masking then represent a total of 23.8% of the recorded ratings. On the other hand, 17% of targets received an ”invisible” rating, corresponding to low-level masking cases.

3.6. Prediction of detection times

Figure 3 shows the correlation of each of the 11 aforementioned features (F1 to F11) with mode detection times. Three features stand out: the MSiCID local predicted similarity (F8), as well as salience imbalance (F10) and the deep model (F11) with linear correlations of 0.36, 0.42 and 0.53, respectively. Using the Fisher r-to-z transformation, we found that there is no significant difference between these three values at the 95% confidence level. It is also noteworthy that local sharpness (F5) yielded a significant correlation as well, although only after logistic mapping. Other features do not seem to correlate (< 0.2) with the mode decision times.

Furthermore, we previously demonstrated that high-level visual masking in a change blindness paradigm is significantly dependent on the experience and expectations of the observers. As the experiment goes on, participants tend to become better at spotting the target, as they gradually fine-tune their search strategies. This seems to be the case even when the stimuli are presented in a completely randomised sequence. However, in our results, we found no significant correlation between the position of a stimulus in a sequence and detection accuracy or time.

4. CONCLUSIONS

In this paper, we analysed the effect of high-level visual masking of image compression artefacts. We presented a user study design based on the change blindness paradigm, where participants have to locate a single noise target, hidden in the scene. Our results indicate that detection times were moderately consistent across observers. Importantly, we found evidence of strong high-level visual masking for 24% of all stimuli and low-level masking in 17% of cases. Mode detection times correlate moderately with local quality difference (with MSiCID features), salience imbalance, as well as the masking threshold predictions of a deep model trained on subjective data. Future work should focus on improving the prediction accuracy of current image quality/fidelity models and on characterising more finely the distribution of detection times and inter-observer variability

Authorized licensed use limited to: Norges Teknisk-Naturvitenskapelige Universitet. Downloaded on February 12,2021 at 12:04:08 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] 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.

[2] 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.

[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] 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] 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.

[7] 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.

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

[9] Daniel J Simons, Steven L Franconeri, and Rebecca L Reimer, "Change blindness in the absence of a visual disruption," *Perception*, vol. 29, no. 10, pp. 1143–1154, 2000.

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

[12] 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.

[13] 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.

[14] Cuong T Vu, Thien D Phan, and Damon M Chandler, "A spectral and spatial measure of local perceived sharpness in natural images," *IEEE transactions on image processing*, vol. 21, no. 3, pp. 934–945, 2011.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] Lin Zhang, Zhongyi Gu, and Hongyu Li, "Sdsp: A novel saliency detection method by combining simple priors," in *2013 IEEE international conference on image processing*. IEEE, 2013, pp. 171–175.

[17] Xiaodi Hou, Jonathan Harel, and Christof Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, no. 1, pp. 194–201, 2012.

[18] Steven Le Moan and Marius Pedersen, "A three-feature model to predict colour change blindness," *Vision*, vol. 3, no. 4, pp. 61, 2019.

[19] A. Chetouani, M. Pedersen, and S. Le Moan, "Prediction of chromatic visual masking with deep learning," in *(submitted to) International Conference on Image Processing*. IEEE, 2020.

Authorized licensed use limited to: Norges Teknisk-Naturvitenskapelige Universitet. Downloaded on February 12,2021 at 12:04:08 UTC from IEEE Xplore. Restrictions apply.