SHORT-TERM MEMORY EFFECTS IN SUBJECTIVE IMAGE QUALITY ASSESSMENT OF NATURAL IMAGES

Tanzima Habib and Marius Pedersen

Department of Computer Science
Norwegian University of Science and Technology

ABSTRACT

In this paper we discuss the influence of short-term memory in judging image quality of natural images. For this, a category judgement experiment with 15 observers has been carried out on a set of natural images. To analyze the presence of memory effect the autocorrelation and correlation between the ratings given by each observer have been studied. Many different orders of the ratings have been considered to draw a meaningful conclusion on the presence of short-term memory effect while evaluating the quality of natural images.

Index Terms— Memory effect, subjective quality assessment, autocorrelation, category judgement

1. INTRODUCTION

Psychophysical experiments are done to measure perceptual magnitude given a physical stimulus. For a long time psychophysics has been limited to studying the relationship between the perceptual magnitude and the physical stimulus and without taking into account the influence of cognitive areas such as learning and memory [1]. Bjorkman et al. has suggested , "A generalized psychophysics..., has to take into account magnitudes of two subjective continua (spaces): the perceptual (immediate experiences) and the memory continuum (past experiences)."[2]. This has given rise to the area of investigation called memory psychophysics.

In the case of image quality assessment a psychophysical experiment provides the subjective evaluation of a visual stimuli i.e. image which can be used as the ground truth to compare objective quality metrics. It has been shown in the literature that different aspects related to the experimental setup and short-term memory will influence the ratings given by observers [3, 4, 5, 6].

In this paper we investigate the influence of short-term memory while evaluating the image quality of natural images i.e. the influence in rating of images due to the rating of previous images. This evaluation is important in order to account for the effect of memory magnitude in image quality assessment. Our results are also compared with the results obtained for X-ray images by Landre et al. [7].

The paper is organized as follows: first we present the background, then the methodology, followed by results and discussion, at last we conclude and propose future work.

2. BACKGROUND

Psychophysical experiments are a set of responses registered w.r.t. one or more varying physical properties of a stimulus over time. As such the responses can be treated as a time series. It is well known that in time series experiments sequential effects occur. The simplest form of memory effect that has been studied in various fields is the sequence effect where the effect of an earlier event can have an effect on later events. The existence of sequence effect in identification of loudness has been studied as early as 1948 by Helson [8] and substantial work was done by Lockhead et al. whose analysis found that sequential effects appear to extend over as many as five trials [9]. Schiferstein et al. carried out sequence effect experiments on hedonic judgements of taste stimuli and found that there is a negative correlation between the judgement of the current stimuli w.r.t the preceding stimulus [10]. DeCarlo et al. has shown that the positive dependency between response and previous stimulus intensity disappears when the inter-stimulus interval is increased from 2 or 6 sec to 15 or 20 sec [11]. But investigating sequential effect for judging quality of images is fairly new. Le Moan et al. [5] worked on understanding the role of short-term memory in subjective image quality assessment. In their work they carried out a paired comparison experiment, where stimuli were shown side-by-side versus stimuli shown one after the other. In the latter, observers would need to rely on short-term memory when making the judgement. Their results suggest that there is a significant chance that observers will make different quality assessments depending on the experimental setup. Landre et al. [7] worked on understanding the memory effect in X-ray images. Three x-ray images with different dose were presented in a category judgement experiment to 20 observers. Their results indicate a memory effect, showing a correlation between the rating of an image and the ratings given in previously judged images. Chang et al. [12] studied how assimilative sequential effect exists even when sequential judgements are made solely based on ones subjective feeling.

3. METHODOLOGY

In this experiment, we use double stimulus category judgement method i.e. we will ask a participant to grade a reproduced image compared to a reference based on a given scale. We chose categorical judgement as the observations obtained can be used as time series to evaluate memory effect.

3.1. Viewing conditions, data and observers

The experiment was carried out under a controlled environment where the ambient illumination was set to 60 cd/m^2 and the display monitor was calibrated to a colour temperature of 6500K, resolution of 1920x1080, luminous intensity of 100cd/m^2 and gamma value of 2.2.

The viewing distance from the monitor was approximately 50 cm. Though no restriction was posed on the participants regarding viewing distance.

Twelve natural images were carefully selected from the CID:IQ image quality database [13]. Such that varied range of low frequency and high frequency information is available in the natural images. Each image was processed five times with two sharpening levels in an attempt to create two good quality images and three blur levels to produce three bad quality images. Therefore, there were 12 × 5 i.e. 60 reproduced images. The higher sharpening level is called $Sharp2$, lower sharpening level is called $Sharp1$, the lowest blur level is called $Blur1$, the middle blur level is called $Blur2$ and the highest blur level is called $Blur3$. The sharpening and blurring level is image dependent and a short trial with observers for each image was carried out to choose the appropriate sharpening and blurring levels such that there is a just noticeable difference between the two consecutive levels (Sharp2-Sharp1-Original-Blur1-Blur2-Blur3).

Fifteen observers took part in the experiment. The age range was between 21-42. 11 observers had experience with evaluating images although the opinion of quality based on pleasing to the eye varied among the observers.

3.2. Psychometric experiment

The psychometric experiment carried out was a stimulus comparison category judgement method. Twelve different reference images were chosen. Each image had five reproductions. Therefore, five comparisons were done against each reference image. Let's call the reference image and its five reproductions together an image set. The experiment is conducted by carrying out the five pair comparisons of one image set and then proceeding to the next set until twelve image sets were completed. The comparison pair (a reproduction and a reference) of an image set is displayed randomly within an image set. For every observer two rounds were carried out where the image sets were displayed in the same order. Therefore, 2x5x12 = 120 comparisons were made by each observer. The observers were asked to evaluate the quality of the reproduced image w.r.t the reference image subjectively and assign a category to the reproduced image based on a scale discussed next.

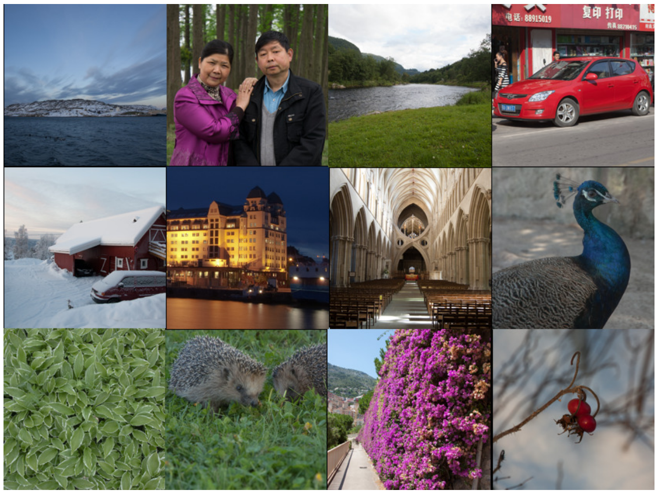

Fig. 1. The twelve images selected from the CID:IQ database.

In this experiment a new nine-category scale was created similar to the ITU-R seven-grade scale. The scale follows as bad (-4), very poor (-3), poor (-1), same (0), fair (1), good (2), very good (3) and excellent (4).

Although, the number of reproduced images are only five, a nine grade scale is chosen to give the participant a certain amount of freedom to decide the category. Also, both the adjectival and the numerical scales were given to the participant.

3.3. Data processing

Correlation in statistics tells the dependence or association between two observations and most commonly states if there is a linear relationship between them.

The most common linear correlation is the Pearson's correlation. Pearsons correlation C_ρ is described as the ratio between the covariance of two observations X and Y and the product of their respective standard deviation σ_x and σ_y.

Spearman's correlation is the non-parametric version of the Pearson's correlation and measures the strength and direction of association between two ranked variables. Spearmans correlation C_s is described as below:

$$C_s = 1 - \frac{\Sigma_{i=1}^{n}(R_i - Q_i)^2}{n^3 - n} \quad (1)$$

where R_i and Q_i are observations of two vectors of rank R and Q obtained on a sample size of n.

Autocorrelation is the similarity between observations as a function of time lag or delay between the observations. Autocorrelation has been used as a metric to find repeating patterns and can be used to measure memory effect in a series. Autocorrelation function is defined as:

$$r_k = 1 - \frac{c_k}{c_0} \quad (2)$$

where k is the lag, c_0 is the sample variance and c_k is defined below.

$$c_k = \frac{1}{T-1}\Sigma_{t=1}^{T-k}(y_t - \overline{y})(y_{t+k} - \overline{y}) \quad (3)$$

where T is the total number of lag and y is the rating of the response at t and $\overline{y}$ is the mean rating of the responses.

The standardized score rating is calculated as follows:

$$RC_{ij} = \frac{R_{ij} - \frac{1}{N}\Sigma_{j=1}^{N}R_{ij}}{\sqrt{\frac{1}{N-1}\Sigma_{j=1}^{N}|R_{ij} - (\frac{1}{N}\Sigma_j R_{ij})|^2}} \quad (4)$$

where RC is the standardized score rating, R is the original score given by an observer i and for image j and N is the total number of observers.

4. RESULTS AND DISCUSSION

Various metrics such as the Pearson's correlation, Spearman's correlation and most important autocorrelation have been used for analysis. First we will discuss the results obtained and then we will compare the obtained result with the results obtained for X-ray images by Landre et al. [7].

4.1. Results achieved with natural images

The ratings given by the observers to all the images cover the whole nine-grade category scale from *bad* to $excellent$. Figure 2 shows the histogram of the ratings according to the category scale. 65.61% of the images are rated between *bad* to $fair$ with the highest number of images being rated as *poor*. This was expected as there are three blurred reproductions while only two sharpened reproductions were included. Also the sharpened images were not necessarily more pleasing than the original. Accordingly, 26.05% of the images have been rated between *slightly good* to *excellent* while only 9 images were rated $excellent$. 150 out of 1800 images have been rated as *same* while no repetitions of the original images were used. The reproductions had just noticeable difference or more from each other and the original image and although difference were noticeable, as the image pairs were being evaluated side by side the difference in some cases might not have been as apparent as in the case of images being compared by flipping as an overlay.

Figure 3 shows the count of different category ratings w.r.t. the level of sharpening/blur applied on the images. Images belonging to $Sharp2$ were mostly rated as *slightly good* and *good*. This was unexpected because to achieve a second level of sharpness which will produce better image than the original is very difficult. But the results show that the observers mostly rated these higher sharpness images as better than the original. This might be due to the fact that when compared to a preceding image that was blurry the sharpened image will be considered better and secondly the observers might have got trained to rating the sharper images as better and blur images worse as the scale of judgement instead of rating according to the appearance of the image that is pleasing to them quality wise when compared to the original. Similarly, images belonging to $Sharp1$ were mostly rated as *same* or *slightly good*, this is also because the sharpness level was quite small. In the case of blurred images, the difference in quality was apparent and also the quality of the image being considered as worse than the original was coherent among the observers. Images belonging to $Blur1$ were rated mostly as *poor* and then $fair$, images belonging to $Blur2$ were mostly rated as *poor* and next as *very poor* and lastly, images belonging to $Blur3$ were rated mostly as *bad* and then *very poor*.

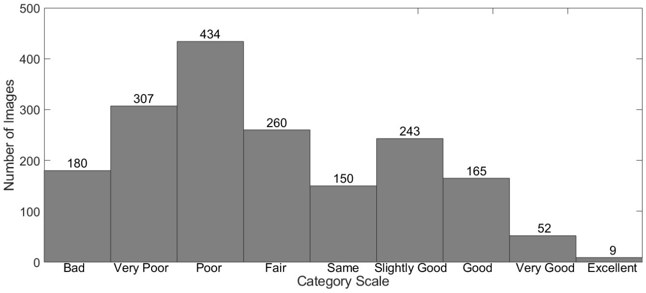

Fig. 2. Histogram of ratings given by the observers.

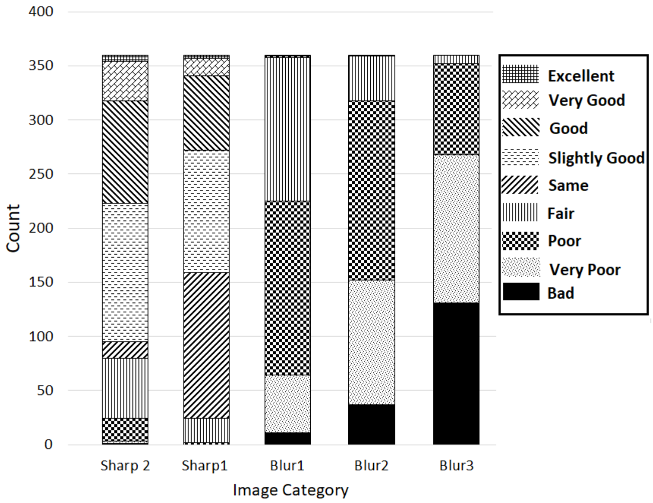

Fig. 3. Count of category scale given by observers for each reproduced level.

The difference between the ratings for each reproduced image and its duplicate i.e. difference in rating in the two rounds were calculated. Figure 4 shows the histogram of the difference in scores between duplicates. The highest count is for 0 difference i.e. 49.8% of the ratings were consistent in the two rounds. While 91.9% ratings differed at most by a difference of 1 in the two rounds for all observers. This tells that the observers were consistent in rating the images.

The spread of the difference is mainly -3 to 3. But there are two images for which the difference in rating is as high as 4 and 6. These two ratings were given by the same observer i.e. observer 9. In the first case, observer 9 rated an image belonging to *Sharp1* as *excellent* and then in the second round rated it as *same* and in the second case the same observer rated an image belonging to *Sharp2* as *excellent* and in the second round rated it as *poor*. This might be due to the time taken in understanding the task to form a subjective opinion. Both images were from the initial image sets 1 and 4 respectively.

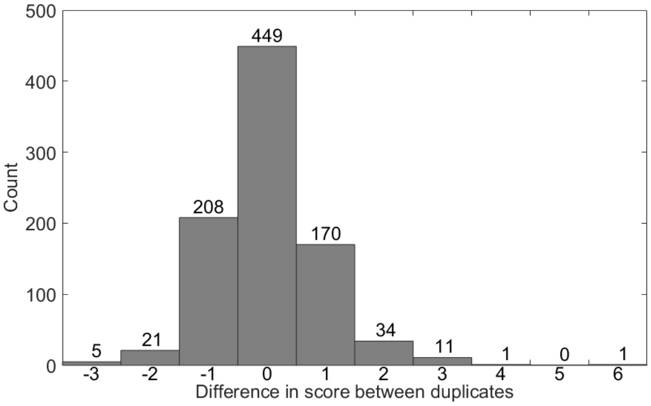

Fig. 4. Difference in score between the duplicates i.e. x-axis represents the difference of ratings given for each reproduced image in the two rounds by an observer and the y-axis is the number of counts.

The linear Pearson correlation between the two rounds of ratings were also calculated for each observer. Except for three observers the linear correlation for all the rest is above 0.85. This again confirms that the observers were consistent in the two rounds. The lowest correlation coefficient is 0.75 which corresponds to observer 9 who also has the two highest difference in ratings as discussed earlier.

Spearman correlation were also calculated for each observers between the two round and the correlation coefficients were all above 0.999 differing only in the fourth decimal. This indicates an almost perfect rank. Therefore, we continue with Pearson correlation results for each observer in this paper.

The autocorrelation for each observer over their whole sequence of raw ratings (category ratings) were calculated. On analysis of average of the autocorrelation graph it was found that all the values after lag 0 were within the 95% bound and were oscillating between negative and positive. The autocorrelation value for lag 1 are mostly negative. We also investigate the autocorrelation for each observer. The average autocorrelation results for all observer and observer 5 with Pearson correlation value 0.95 (highest) are shown as example in figure 5 and figure 6 respectively. These plots do not show any indication of a correlation between the rating given to an image and the succeeding ratings given by observers.

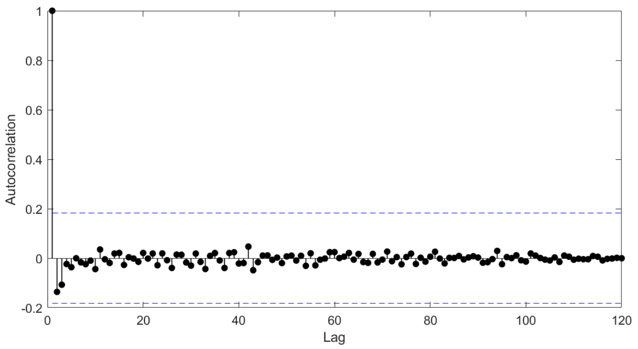

Fig. 5. The average autocorrelation of the original ratings given by all observers. Dotted blue line indicate the 95% confidence bound.

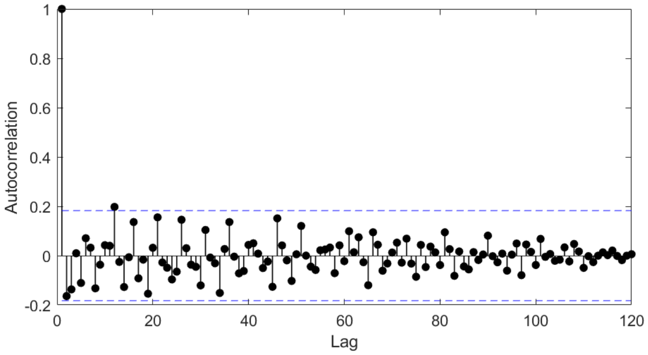

Fig. 6. Autocorrelation of the original ratings given by observer 5. Dotted blue line indicate the 95% confidence bound.

Therefore, the standardized score ratings for each image for each observer were calculated and autocorrelation was calculated over these ratings, following a similar procedure as Landre et al. [7]. Now the autocorrelation value for lag 1 is positive, but still insignificant. Only in few cases the value for lag1 is significant. The average autocorrelation values for both original scores (figure 5) and standardized score (figure 7) ratings were calculated and they were negligible and oscillated around Y=0.

These results indicate that the overall rating of natural images using category judgement were not influenced by memory. So for further investigation, we selected two types of sequences for each observer ratings. In the first case, a pair of ratings were added to the sequence when a preceding image is better than the subsequent image. If one of the images in the pair of good-followed by-bad images is included in the previous turn then the rating for that image is not included again and only its corresponding image's rating is added. Similarly, the second sequence where bad quality follows the good quality images are created as well. We plot the difference between the current stimuli level and the previous stimuli level vs the difference between the current response rating and the previous response rating for the two sequences as shown in figure

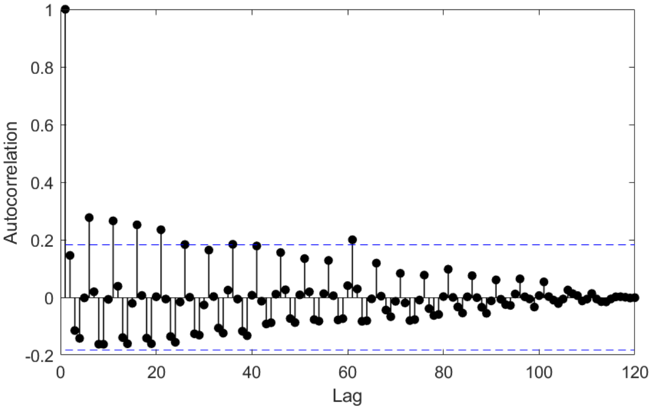

Fig. 7. Autocorrelation of standardized score ratings for all observers.

8 and figure 9.

We see for the two sequences that most observers were able to maintain approximately a similar difference in the ratings as the level difference between the two stimuli. This shows that depending on the two consequent stimuli level difference the observers try to maintain a similar difference for their respective response ratings. This indicates the sequence effect in time series which is well known to be impacted by only stimulus preceding in time sequence by one [8].

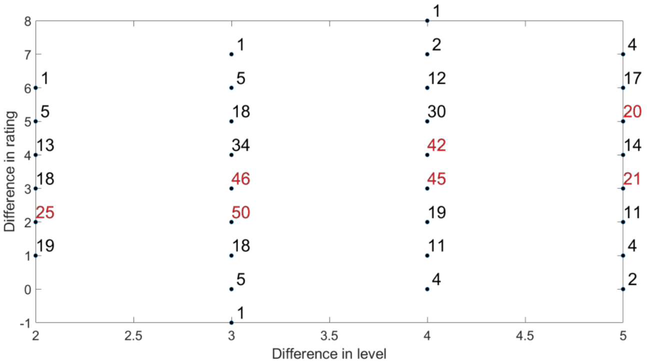

Fig. 8. The count of observations plotted for a stimuli level difference vs rating difference graph where a good quality image preceded the next image calculated for all observers.(x-axis is the difference between the stimuli levels and y-axis is the difference between their responses for good followed by bad image sequence).

Therefore to substantiate the sequential effect with lag one we applied a regression model to the observers data with the following equation:

$$R_i = b_1 R_{i-1} + b_2 S_{i-1} + b_3 S_i + b_4 e \tag{5}$$

where R_{i-1} is the preceding response rating, S_{i-1} preceding stimuli level, S_i is the current stimuli level, R_i is the current response level and e is the error. While b_1, b_2, b_3 and b_4 are the respective coefficients that we calculate by fitting each observer data to the model. For all three predictors R_{i-1}, S_{i-1}, and S_i the mean coefficients are 0.1256 -0.0918 and 0.8905 respectively. The S_i coefficients b_3 are positive with $p - value < 0.001$. The S_{i-1} coefficients b_2 are negative with $p - value = 0.0095$ and the R_{i-1} coefficients are positive with $p - value = 0.0013$ which shows that the current response ratings are in contrast to the previous stimuli and they assimilate to the previous response rating i,e, the current response rating is biased towards the rating given to the previous response but away from the previous stimulus level [14].

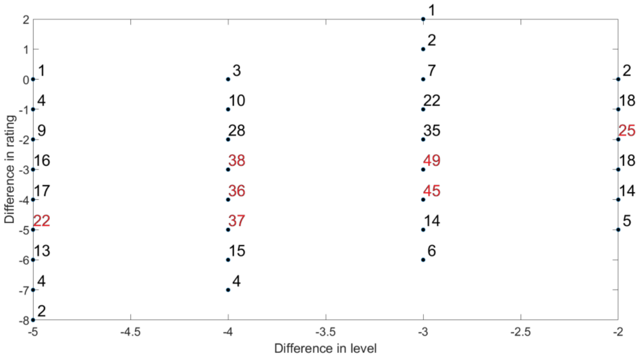

Fig. 9. The count of observations plotted for a stimuli level difference vs rating difference graph where a worse quality image preceded the next image calculated for all observers.x-axis is the difference between the stimuli levels and y-axis is the difference between their responses for bad followed by good image sequence).

4.2. Comparison with results from X-ray images

Landre et al. [7] carried out a similar experiment to gauge the memory effect in the subjective quality assessment of X-ray images. This paper has been an extension of that investigation onto natural images. Landre et al. used three types of X-ray images and three different levels of dose for reproduction of the first two type and four levels for the third image. In total they used 20 images to be rated by 20 radiology students and used a 5-grade scale ranging from -2 to 2 (bad to excellent). The image was judged according to the sharpness of the trabeculae, which are small elements in the form of beams, struts or rods. Their findings show that the observers tended to rate between -1 to 1 and mostly rated as same as the original.This is very different from our case, because the difference in quality is apparent in natural images with the chosen image processing metrics. While in the case of X-ray images they are black and white, the study of judging the goodness of the trabeculae is technical and concentrated, and not as subjective and vast in information as in judging natural images. Also the methods to create the reproductions are different.

The difference in score between duplicates in the study by

Landre et al. is similar to the ones obtained in our study of natural images. Both spread mainly between a difference of -3 and 3 while most rating belong to 0 difference signifying that the observers were consistent in the two rounds.

The overall autocorrelation results from Landre et al. obtained for the standardized score rating of all the observers showed fairly high and positive autocorrelation for the first four lags suggesting that memory effect is present for the first four lags and then the impact attenuates. This is different in the case of natural image where the overall sequence doesn't signify the influence of memory while making judgements.

5. CONCLUSION

Unlike in the case of X-ray images studying the memory effect in natural image is very elusive. No significant autocorrelation was found for the overall rating of the images by all observers even for lag 1. When the sequence for each observer were selected according to a good quality image followed by a bad quality image and vice versa, a comparison between the two consequent stimuli level difference and two response rating difference suggest that sequential effect of lag 1 exists as the observers try to maintain the same level difference in most cases. Also in this experiment the images in the next trial appeared right after the previous one without any pause and at the same position, so the difference between the previous and the new stimuli becomes clear to the observer the moment they change which influence the direction of their judgement. The regression method applied also suggests that there is a significant sequential effect of lag 1. Over the period, the observers showed a sign of being trained by the presence of the different levels of the image processing done and by judging just the reproduced levels as good or bad rather than evaluating subjectively the quality of the reproduced image. There is also much more information in a natural image for an observer to consider while judging its quality. In the future, including a pause between the two images should be considered which is known to decrease sequential effect. Also, natural images can be categorized according to the amount of colour information, texture information etc. it holds and can be further investigated to understand the influence of memory according to these categories. Lastly, calculating the sequential effect using regression can be performed for various lags as has been studied in other fields such as identification of loudness, judging taste, guessing price of an object. etc.

References

[1] John Charlton Baird and Elliot Jason Noma, *Fundamentals of scaling and psychophysics*, John Wiley & Sons, 1978.

[2] M Björkman, I Lundberg, and S Tärnblom, "On the relationship between percept and memory; a psychophysical approach," *Scandinavian Journal of Psychology*, vol. 1, no. 1, pp. 136–144, 1960.

[3] Steven Le Moan and Marius Pedersen, "Subjective image fidelity assessment: Effect of the spatial distance between stimuli," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 445–449.

[4] Steven Le Moan and Marius Pedersen, "Evidence of change blindness in subjective image fidelity assessment," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3155–3159.

[5] Steven Le Moan, Marius Pedersen, Ivar Farup, and Jana Blahová, "The influence of short-term memory in subjective image quality assessment," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 91–95.

[6] Steven Le Moan and Marius Pedersen, "Subjective image fidelity assessment: Effect of the spatial distance between stimuli," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

[7] Victor Landre, Marius Pedersen, and Dag Waaler, "Memory effects in subjective quality assessment of x-ray images," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 314–325.

[8] Walt Jesteadt, R Duncan Luce, and David M Green, "Sequential effects in judgments of loudness," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 1, pp. 92, 1977.

[9] Morris K Holland and GR Lockhead, "Sequential effects in absolute judgments of loudness," *Perception & Psychophysics*, vol. 3, no. 6, pp. 409–414, 1968.

[10] Hendrik NJ Schifferstein and W Erno Kuiper, "Sequence effects in hedonic judgments of taste stimuli," *Perception & psychophysics*, vol. 59, no. 6, pp. 900–912, 1997.

[11] Lawrence T DeCarlo, "Intertrial interval and sequential effects in magnitude scaling," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 4, pp. 1080, 1992.

[12] Seah Chang, Chai-Youn Kim, and Yang Seok Cho, "Sequential effects in preference decision: Prior preference assimilates current preference," *PloS one*, vol. 12, no. 8, pp. e0182442, 2017.

[13] Xinwei Liu, Marius Pedersen, and Jon Yngve Hardeberg, "Cid: Iq–a new image quality database," in *International Conference on Image and Signal Processing*. Springer, 2014, pp. 193–202.

[14] William J Matthews and Neil Stewart, "Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks," *Judgment and Decision Making*, vol. 4, no. 1, pp. 64, 2009.