Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Subjective Image Quality Assessment: a Method Based on Signal Detection Theory

Yurong He[1,2], Yuming Xuan[1], Wenfeng Chen[1], Xiaolan Fu[1]
[1]State Key Laboratory of Brain and Cognitive Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[2]Graduate University of Chinese Academy of Sciences,
Beijing 100049, China
email: heyr, xuanym, chenwf, fuxl@psych.ac.cn

***Abstract*—In the field of image quality assessment, to develop computerized/objective methods whose evaluations are in close agreement with human judgments becomes a main task. However, accurate evaluation of human's subjective judgments is still a problem. Tradition methods based on mean opinion score (MOS) were not accurate enough, especially for images of minor changes or distortions. The present study tried to apply signal detection theory (SDT) in the field of image quality assessment, since SDT is particularly useful in measuring the way we make decisions under conditions of uncertainty. The results of three psychophysics experiments, in which images of different watermarking strengths were used as stimuli, showed that the SDT-based method was especially useful to detect the small loss of fidelity of images. This conclusion was supported by the higher correlation between the sensitivity score, P(A), with several computerized/objective QA indexes, such as PSNR, VIF and SSIM. Detecting subtle changes of images might involve some unknown implicit mechanisms for participants did not perform well enough in full-reference framework which allowing direct comparisons of the changed image to the original one.**

***Keywords*—Image quality assessment, computerized/objective assessment, human subjective assessment, signal detection theory, watermark**

I. Introduction

Image quality is a characteristic of an image that measures the perceived image degradation (typically, compared to an ideal or perfect image). Most processing (e.g. compression, restoration, resizing or watermarking etc.) may inevitably result in a change of the image quality since distortion or artifacts have to be introduced into the image during the processing [1]. Generally, the output images will undergo an image quality degradation compared to the original ones. That is why reliable and accurate image assessment approaches are necessary [1-5].

The obvious way to measure image quality is to solicit human opinion. This is known as subjective quality assessment (QA) method. However, such evaluations are time-consuming, cumbersome, and expensive to conduct, and the methodology is difficult to embed into real-world applications [2, 4-6]. Thus increase the needs for automatic algorithms for computerized/objective QA that can analyze images and report their quality without human involvement. Considering the fact that no matter what algorithms are used to process images, in most cases we human beings will be the ultimate acceptors of the processed images, to develop methods whose evaluations are in close agreement with human judgments becomes the main tasks. The performance of a computerized/objective assessment method can be judged by comparing the result of computerized/objective QA and that of human subjective QA. The higher correlation between the results of objective and subjective QA is, the better the objective QA method is.

As mentioned above, in a traditional subjective image QA method, a rating scale (usually 5 levels) would be provided to the participants and then they were required to choose a level consistent with their subjective visual observation. Mean opinion scores (MOSs) were computed based on the ratings [2, 6, 7, 8]. Therefore, it may be a problem that average MOSs cannot tell the differences between images of subtle changes or distortions from the original ones, especially when testing a small sample of observers.

To overcome this problem, in the present paper, we tried to develop a new subjective QA method based on signal detection theory (SDT). SDT is used when psychologists want to measure the way we make decisions under conditions of uncertainty. SDT assumes that the decision maker is not a passive receiver of information, but an active decision-maker who makes difficult perceptual judgments under conditions of uncertainty, just like trying to tell whether a mildly distorted image is different from its original image [9]. Therefore, SDT should be a perfect candidate to be used as a method of subjective assessment. Moreover, we believe SDT is a more accurate method than the often-used subjective QA methods based on MOS, since SDT is able to discriminate between the real sensitivity of subjects and their (potential) response biases.

To obtain series of images that change in a small range, we chose watermarked images as experimental stimuli. Evaluating the quality of the watermarked image (or other digital media) is also a very important issue. To protect intellectual property, transparent or invisible but informative watermark need to be embedded into digital media, for instance, digital images. The more informative the watermark, the watermarked image will be more robust to the potential attack meanwhile undergoes more loss of fidelity. These fidelity and robustness constraints often conflict. Researchers need to check out the thresholds of the visibility of the watermark in order to maximize the

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

watermarking strength within the perceptual constraints [11]. Usually, it is quite convenient for a watermarking algorithm to generate series of images of different watermarking strength.

In the present study, three experiments were conducted to explore the feasibility of the SDT–based subjective QA method and its two variants. Traditional subjective QA method based on MOSs was also tested in each experiment to compare to the results of SDT method. Since three experiments were quite similar in their procedures, we would explain the method of experiment 1 in detail then described experiments 2 and 3 in brief. Each experiment would last for about 50-60 minutes. Results of all three experiments would be shown together after the description of the experimental methods.

II. Methods

Experiment 1

Participants

Twenty university students (9 males and 11 females) for payment with normal or corrected-to-normal visual acuity and normal color vision participated in the study. They were all novel to the test. Six of twenty participants were rejected for using only one criterion in the SDT test (which made P(A) become incomputable).

Apparatus and Stimuli

The experiments were run on a Pentium-IV PC with a 17-inch monitor at a 1280×1024 resolution. E-Prime 1.2 was used to control the stimuli presentation and response recording [10]. The luminance was constant and moderate in the testing laboratory. Participants sat approximately 50 cm from the screen and used a standard mouse as the response device. The identical apparatus were used in all experiments.

Full-reference framework was adopted in experiment 1. Two gray-scale pictures, Lena and Peppers, were used as the reference pictures. The original Lena and Peppers image were then watermarked with five levels of strengths as shown in Table I (for the watermarking algorithm, please refer to the paper will be presented in the coming ICIP in this October [12]). Each level of watermarking strength for Lena and Peppers would be a bit different to obtain the same PSNR (peak signal noise ratio, table I).

Procedure

Participants performed a MOS test first and then a SDT test.

For the MOS test, each trial began with a fixation cross at the center of the screen for 1000 ms. A pair of images at full size (512×512 pixels), were simultaneously displayed side by side against a completely black background on the screen, subtending 33.8° ×13.5° visual angle. The left image of the pair was always the original Lena or Peppers image, while the right image was always a watermarked image of left one. For each pairs of image, participants were instructed to rate how strongly he/she felt that the right image had lost its fidelity compared to the left original image and to use the mouse to click on the corresponding rating displayed on the bottom of the screen. The 5 levels of ratings were "5-Imperceptible", "4-Perceptible, but not annoying", "3-Slightly annoying", "2-Annoying" and "1-Very annoying" respectively [8]. The pair of images would stay on the screen until participants made responses. Then a silver screen with 1000 ms duration would follow as masking.

Participants rated each watermarked images for three times. Since there were 5 levels of watermarked images based on the original Lena or Peppers image, altogether there were 30 trials in the MOSs test.

The SDT test consisted of a learning stage and a test stage. In the learning stage, sequences of 6 images would be presented on the center of the screen. Participants were told that each sequence began with the original Lena or Peppers image, following by its watermarked images, and the watermarking strength increased gradually one by one. Each image in the sequence stayed for 3 seconds. Participants watched the Lena sequence first, and then the Peppers sequence and then watched the two sequences again.

In the SDT test stage, the stimuli were presented in a similar way as the MOSs test except the following differences. For SDT test trials, the original image (Lena or Peppers) would be always presented on the left side of the screen. The right image might be the same original image as the left one (we called this pair as a Noise pair) or watermarked image of the left one (we called this pair as a Signal pair). For each pairs of image, participants were instructed to determine how strongly he/she was confident that the two images were different and to use the mouse to click on one of five buttons indicating the confidence levels below the images. The confidence levels were 0% (definitely same), 25%, 50%, 75%, 100% (definitely different) respectively [9].

For simplicity, we called the pair of images presented in the MOS and SDT test as Lena pair if the left image was Lena, and similarly we called a pair of images as Peppers pair if the left image was Peppers.

There were 5 blocks of trials in SDT test stage, with 60 trials in each block. In each block, half trials were Lena pairs and half were Peppers pairs. For both Lena and Peppers pairs, half of them were Signal pairs and half were Noise pairs. The watermarked images in Lena and Peppers Signal pairs had the same PSNR value in each block. To prevent that the same pair of images was displayed on two consecutive trials, Lena pairs and Peppers pairs were displayed by turns, i.e., a Lena pair was always followed by a Peppers pair, and vice versa. But whether a pair was a Noise or a Signal pair was random on each trial. The order of the blocks was random across participants and participants could take a rest between blocks.

Participants had 20 practice trials before the formal test trials. All participants completed 320 trials in the SDT test.

In the end of the experiment, participants were interviewed by the experimenter with the following questions: 1. which images are more easily to tell the traces of watermarking, Lena or Peppers? 2. Do you feel tired after finishing the test?

Experiment 2

Participants

Fifteen new participants for payment (6 male and 8 female university students) with normal or corrected-to-normal visual acuity and normal color vision attended experiment 2. One of the fifteen participants were rejected for using only one criterion in the SDT test

Procedure

The procedure of experiment 2 was quite similar with that of experiment 1 except the following differences.

1. The primary differences was that in both MOS and SDT test, only a watermarked image was presented on the center of the screen on each trial, subtending 13.5° ×13.5° visual angle.
2. In the MOS test, participants were instructed to rate the quality of the image presented on five levels, i.e., “5-Excellent”, “4-Good”, “3-Fair”, “2-Poor”, and “1-Unsatisfactory” [8].
3. In the SDT test, participants were instructed to determine how strongly he/she felt the image presented was different from the original image ever presented in the SDT learning stage.

Experiment 3

Participants

Seventeen new participants for payment (7 male and 10 female university students) with normal or corrected-to-normal visual acuity and normal color vision attended experiment 3.

Procedure

The procedure of experiment 3 was quite similar with that of experiment 2 except the following differences: the test was divided into Lena part and Peppers part. In Lena or Peppers part, only Lena or Peppers images were tested. In both parts, participants would accept the MOSs test, the SDT learning stage and the SDT test stage in order. Eight participants did the Lena part first then the Peppers part and nine participants completed the experiment in a reverse order.

III. Results

For the calculation of MOSs, first the raw scores were converted into Z-scores. The Z-scores were then re-scaled to 1 – 100 range before being averaged across subjects to give the MOSs [7]. Higher MOSs corresponded to lower image quality.

For the results of SDT test, for each picture of Lena and Paper, we calculated the hit rate and false alarm rate in percentage for each participate, and converted both rates into Z-scores and then computed the average Z-scores of hit rate and false alarm rate. After that, the average Z-scours of hit rate and false alarm rate were reconverted to percentage and the Receiver-Operating Characteristic (ROC) curves for each image were obtained. The proportion of area under the ROC curves, i.e., P(A), was computed as participants’ sensitivity score for distinguishing a specific watermarked image from its original image [9]. A higher P(A) meant higher sensitivity.

Table I showed the average MOSs and P(A) for each watermarked Lena and Peppers in all three experiments. A 3 (Experiments: exp1, exp2, exp3) × 2 (Images: Lena, Peppers) × 5 (PSNR: 40, 42, 44, 46, 48db) ANOVA analysis was conducted on MOSs. The results showed that the main effect of Images was significant, $F(1, 42) = 4.60$, $p < 0.05$, indicating generally participants perceived that watermarked Lena images (Mean MOSs = 60.60) had lost more fidelity than watermarked Peppers images (Mean MOSs = 54.12) even though they had the same PSNR (figure 1a). The main effects of PSNR was significant, $F(4, 168) = 5.00$, $p < 0.001$. The higher the PSNR, the higher the MOSs were given to the watermarked images ($\text{MOSs}|_{40db} = 52.09$, $\text{MOSs}|_{42db} = 55.92$, $\text{MOSs}|_{44db} = 56.93$, $\text{MOSs}|_{46db} = 60.42$, $\text{MOSs}|_{48db} = 61.44$), suggesting in general MOSs could reflect the image fidelity. The interaction of PSNR and Images was significant, $F(4, 168) = 3.24$, $p < 0.05$. Simple effect analysis showed that the effect of Images was significant on all levels of PSNR, $F\text{s}(1,44) = 15.76, 13.73, 20.98, 24.73, 25.75$, respectively, all $p\text{s} < 0.001$. The main effect of Experiments, and all the interactions between Experiments and other independent variables were not significant, all $p\text{s} > 0.05$, suggesting that participants in three experiments had given about the same MOSs to the watermarked images.

TABLE I. Mean MOSs and P(A) for Lena or Peppers images of different PSNR, Watermarking Strength (Wm. Str.), VIF, & SSIM in three experiments.

	MOSs			P(A)			PSNR	Wm.	VIF	SSIM
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	(db)	Str.		
Lena	45.36	55.35	53.17	0.69	0.84	0.85	40	0.440	0.822	0.951
	58.87	55.66	59.72	0.68	0.88	0.84	42	0.348	0.865	0.968
	60.82	69.71	64.66	0.60	0.78	0.81	44	0.276	0.900	0.979
	51.35	67.64	66.86	0.54	0.69	0.67	46	0.220	0.928	0.986
	60.08	72.49	65.65	0.54	0.59	0.63	48	0.170	0.951	0.992
Peppers	52.34	55.16	51.13	0.54	0.63	0.62	40	0.470	0.795	0.952
	63.10	45.18	52.85	0.63	0.62	0.59	42	0.380	0.847	0.969
	52.42	45.37	48.69	0.58	0.60	0.56	44	0.300	0.887	0.980
	64.91	50.51	59.97	0.48	0.50	0.51	46	0.230	0.922	0.987
	55.04	60.74	55.08	0.50	0.49	0.47	48	0.180	0.948	0.992

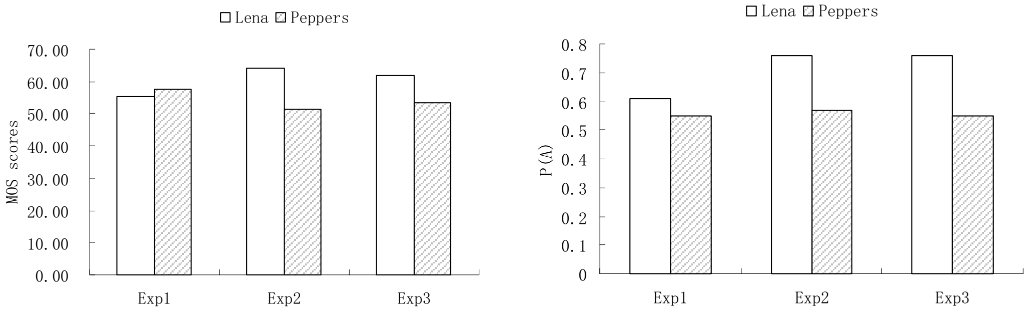

Figure1. Comparison of mean MOSs (a) and P(A) (b) for Lena and Peppers in three experiments.

For P(A), A 3 (Experiments: exp1, exp2, exp3) × 2 (Images: Lena, Peppers) × 5 (PSNR: 40, 42, 44, 46, 48db) ANOVA analysis was also conducted. The results showed that the main effect of Images was significant, $F(1,12) = 45.91$, $p < 0.001$. Participants were more sensitive to watermarked Lena images (Mean P(A) = 0.66) than to watermarked Peppers images (Mean P(A) = 0.66) even though they had the same PSNR. The main effects of PSNR was significant, $F(4, 168) = 19.19$, $p < 0.001$. The lower the PSNR (i.e., the stronger the signal [watermarking] was), the higher sensitivity participants would demonstrated (P(A)$|_{40db}$ = 0.67, P(A)$|_{42db}$ = 0.63, P(A)$|_{44db}$ = 0.59, P(A)$|_{46db}$ = 0.56, P(A)$|_{48db}$ = 0.54), indicating in general P(A) could also reflect the image fidelity. The interaction of Experiments and Images was significant, $F(2,42) = 7.10$, $p < 0.01$. Simple effect analysis showed that the effect of Images was not significant in experiment 1, $p = 0.375$, but was significant in experiments 2 and 3, *F*s(1,12) = 21.47, 41.63 respectively, all *p*s < 0.001. As shown in figure 1b, the P(A) difference between Lena and Peppers in experiments 2 and 3 were greater than that in experiment 1. All other effects were not significant.

The higher sensitivity of Lena over Peppers could also be seen from the after-experiment interview. All participants claimed that watermarked Lena images were more easily to perceive from the original image than watermarked Peppers images.

To compare how closely subjective QA scores were consistent with those objective QA indexes, Spearman correlation coefficients were computed. The reason to choose Spearman correlation but not Pearson correlation was that Spearman correlation was more reliable for small samples like our experiments. We added two more objective QA indexes, namely VIF [13] and SSIM [14] to have more comparisons (tables I and II).

As shown in table II, we could see that for Lena or Peppers images in all experiments, Spearman correlations between P(A) and any objective QA index were greater than (sometimes the same as) those between MOSs and the corresponding objective QA index. The differences were more obvious for Peppers images. To be specific, for watermarked Peppers images, Spearman correlations between MOSs and any objective QA scores were no more than 0.60, and did not reach statistical significance, while Spearman correlations between P(A) and any objective QA scores were all −1.00 in experiments 2 and 3.

TABLE II. Spearman correlation between subjective QA scores [MOSs or P(A)] and objective QA scores [PSNR, Watermarking Strength (Wm. Str.), VIF, or SSIM] for Lena and Peppers images in three experiments. * Correlation is significant at the 0.05 level (2-tailed); ** Correlation is significant at the 0.01 level (2-tailed).

		MOSs			P(A)		
		Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
Lena	PSNR	0. 50	0. 90*	0. 90*	−0. 97**	−0. 90*	−1. 00**
	Wm. Str.	−0. 50	−0. 90*	−0. 90*	0. 97**	0. 90*	1. 00**
	VIF	0. 50	0. 90*	0. 90*	−0. 97**	−0. 90*	−1. 00**
	SSIM	0. 50	0. 90*	0. 90*	−0. 97**	−0. 90*	−1. 00**
Peppers	PSNR	0. 50	0. 40	0. 60	−0. 60	−1. 00**	−1. 00**
	Wm. Str.	−0. 50	−0. 40	−0. 60	0. 60	1. 00**	1. 00**
	VIF	0. 50	0. 40	0. 60	−0. 60	−1. 00**	−1. 00**
	SSIM	0. 50	0. 40	0. 60	−0. 60	−1. 00**	−1. 00**

Correlations between MOSs and objective QA index were good (0.90) for Lena images in Experiments 2 and 3, but dropped to a low level for Peppers images; while correlations between P(A) and objective QA index kept at a high level for both Lena and Peppers images in experiments 2 and 3. Bearing these in mind and considering that participants showed lower sensitivity to watermarked Peppers images according to previous analysis, SDT seemed to be a more appropriate method than MOSs was as to detect signals to which people were less sensitive.

IV. Discussion

In the present paper, three psychophysics experiments were conducted to examine a SDT- based subjective QA method. Our results showed that the SDT-based method seemed to be better than MOSs method in which participants could detect subtle changes (signals) better.

Applying SDT to field of image QA is not an easy thing. In fact, there had been some study of image QA which used SDT-alike method, such as the two alternative forced choice procedure [1,7]. But a real SDT method had not been used in QA. For a standard experiment based on SDT, signals and noises are usually separate, and noises are not presented to the participants before test. Participants learn signals stimuli only in the learning stages and try to discriminate the signals out of noises stimuli in the test stages. But for image QA, signals (changes, distortions or watermarks) are embedded in noises (the cover Work, or original Work [11]). Thus in fact stimuli represented to participants are noises + signals. And because SDT is only useful for uncertain conditions, it determines that strong signals cannot be used in SDT experiments. For human participants, detecting weak signals embedded in noises for a long time must be tiresome. Indeed, some participants reported that they were a bit exhausted after the one-hour session.

Though our SDT method cannot overcome the disadvantage of time-consuming, it is encouraging that SDT method seems to need fewer participants than MOSs method does. For MOSs method, usually 15 participants are needed. In each of our experiments, we had 14 to 17 participants to meet this requirement. But in fact half of them were enough to get the similar SDT results according to our preliminary data analysis in the running of the experiments.

Besides, comparing to P(A) for Lena, the reason why P(A) for Peppers in experiment 1 dropped to a low correlation with objective QA indexes is unknown. According to the data available, participants of our experiments did better in SDT test in experiments 2 and 3 than in experiment 1. It seems that full references framework (like experiment 1) is not quite suitable for detecting subtle signals to which people are less sensitive. To detect subtle signals, a semi-full references framework may be more effective, in which observers somehow can do better by some implicit mechanism.

V. conclusion

The present study tried to apply signal detection theory in the field of image quality assessment. The results of three psychophysics experiments showed that the SDT-based method was especially useful to detect the small loss of fidelity of images. This conclusion was supported by the higher correlation between the sensitivity score, P(A), with several computerized/objective QA indexes, such as PSNR, VIF and SSIM. Detecting weak or subtle change of images might involve some unknown implicit mechanisms for participants did not perform well enough in full-reference framework which allowing direct comparisons of the changed image to the original one. Therefore, the contribution of the present paper is to propose a subjective QA method which can be used to calibrate the performances of objective QA methods especially for images with minor changes or loss of fidelity.

Acknowledgment

This research was supported in part by grants from 973 Program (2006CB303101) and the National Natural Science Foundation of China (90820305, 30500157, 30600182). Correspondence should be emailed to Yuming Xuan (xuanym@psych.ac.cn). We thank Dr. Xiangui Kang, Dr. Jiwu Huang, and Dr. Xiong Zhong in Sun Yat-Sen University for the help of generating the watermarked stimuli.

Reference

[1] H.R Sheikh, A.C. Bovik, Information Theoretic Approaces to Image Quality Assessment. In: Bovik, A.C. Handbook of Image and Video Processing. Elsevier, 2005.

[2] A. M. Eskicioglu, Quality measurement for monochrome compressed images in the past 25 years. Journal of Electronic Imaging, 2001, 10(1): 20–29

[3] T. N. Pappas, R. J. Safranek, Perceptual Criteria for Image Quality Evaluation. Handbook of Image and Video Processing, Al Bovik (ed.), Academic Press, San Diego, 2000: 669–684

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error measurement to structural similarity. IEEE Trans. Image Process, 2004, 13: 600–612

[5] Z. Wang, A. C. Bovik, L. Lu, Why is image quality assessment so difficult? IEEE International Conference on Acoustics, Speech, & Signal Processing, May 2002.

[6] A. Shnayderman, A. Gusev, A. M. Eskicioglu, An SVD-Based Gray-Scale Image Quality Measure for Local and Global Assessment. IEEE Transactions on Image Processing, 2006, 15: 422–429

[7] H. R. Sheikh, Image Quality Assessment Using Natural Scene Statistics. PhD thesis, The University of Texas at Austin, 2004

[8] ITU-R Rec. BT. 500-11. Methodology for the Subjective Assessment of the Quality for Television Pictures, 2002

[9] D. McNicol, A Primer of Signal Detection Theory London: George Allen & Unwin, 1972

[10] W. Schneider, A. Eschman, A. Zuccolotto, E-Prime User's Guide: Psychology Software Tools, 2002

[11] I. J. Cox, M. L. Miller, The First 50 Years of Electronic Watermarking. EURASIP J. of Applied Signal Processing, 2002, 56(2): 126–132,

[12] X. G. Kang, X. Zhong, J. W. Huang, W. J. Zeng. An Efficient Print-scanning Resilient Data Hiding Based on a Novel LPM. Proceedings of the 15th IEEE international conference on image processing 2008, pp. 2080-2083. ICIP 2008, Oct 12-15. 2008, San Diego, CA, USA.

[13] H. R. Sheikh, A. C. Bovik, G. de Veciana, An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics. IEEE Transactions on Image Processing, Dec. 2005, 14(12): 2117–2128

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. Apr. 2004, 13(4): 600–612