IEEE DOI: 10.1109/ICASSP.2014.6853577

SUBJECTIVE SIMILARITY EVALUATION FOR SCENIC BILEVEL IMAGES

Yuanhao Zhai[1], *David L. Neuhoff*[1], *Thrasyvoulos N. Pappas*[2]

[1]EECS Dept., Univ. of Michigan, [2]EECS Dept., Northwestern Univ.

ABSTRACT

In order to provide ground truth for subjectively comparing compression methods for scenic bilevel images, as well as for judging objective similarity metrics, this paper describes the subjective similarity rating of a collection of distorted scenic bilevel images. Unlike text, line drawings, and silhouettes, scenic bilevel images contain natural scenes, e.g., landscapes and portraits. Seven scenic images were each distorted in forty-four ways, including random bit flipping, dilation, erosion and lossy compression. To produce subjective similarity ratings, the distorted images were each viewed by 77 subjects. These are then used to compare the performance of four compression algorithms and to assess how well percentage error and SmSIM work as bilevel image similarity metrics. These subjective ratings can also provide ground truth for future tests of objective bilevel image similarity metrics.

Index Terms— bilevel image similarity, image quality, subjective evaluation

1. INTRODUCTION

Bilevel images are images with only two intensity levels: black and white. Different from text, line drawings and silhouettes, the bilevel images in which we are interested are scenic images, which are complex images containing natural scenes, e.g., landscapes and portraits, but are not halftoned. Several lossy compression algorithms have been developed to compress such images, e.g., Lossy Cutset Coding (LCC) [1,2] and Hierarchical Cutset-MRF (HC) [3], and other coding algorithms, not specifically developed for such, can be applied as well, e.g., Finite State Automata (FSA) [4] and JBIG2 [5]. Traditionally, percentage error is used as a metric to quantify the similarity of distorted scenic bilevel images. This is equivalent to mean-squared error (MSE) in the bilevel case. Unfortunately, this metric is not always consistent with human perception, as images with similar percentage error often appear very different to viewers. Accordingly, there is need to develop better objective bilevel image similarity metrics. However, the only such metric of which we are aware is SmSIM proposed in [6], which, as will be seen, does not perform especially well. In order to develop and assess new metrics, it is essential to have ground truth, i.e., a set of distorted images whose perceptual similarity (perceived distortion) have been subjectively rated by human viewers.

In this work, we conducted a subjective similarity evaluation of distorted scenic bilevel images. The subjective ratings are used to provide ground truth in assessing the performance of four compression algorithms and to judge how well percentage error and SmSIM work as bilevel image similarity metrics. In addition, the subjective ratings can provide ground truth for future designs and tests of objective bilevel image similarity metrics, as in the companion paper [7].

There has been considerable work in designing quality metrics for grayscale and color images and videos. Correspondingly there has been considerable work in developing subjective evaluations to obtain ground truth for such. In ITU-R BT.500-11 [8], a thorough study of subjective evaluation methodologies has been conducted for television pictures. Several methods are suggested based on different purposes. For example, double-stimulus continuous quality-scale (DSCQS) method is suggested for measuring the quality of systems relative to a reference. This method has been applied by the Video Quality Expert Group (VQEG) [9] and has been claimed to have the least contextual effects. However, due to the large number of sample images to be viewed, DSCQS may be too complex, in that observers might not have time to rate enough images. Hence, a single-stimulus (SS) methodology is often used instead. For example, it was used in [10] to judge the quality of color images. Motivated by previous work, in this paper, we design our evaluation using a modified version of the SDSCE methodology suggested in [8], as described in the next section.

The remainder of the paper is organized as follows. Section 2 presents the experiment design. Data processing methods are described in Section 3. In Section 4, the experimental results are analyzed. Section 5 applies the subjective ratings to assessing compression algorithms and two simple similarity metrics. Finally, Section 6 concludes the paper.

2. EXPERIMENT DESIGN

Since humans can often evaluate the quality of grayscale and color images without seeing the original image, many methodologies, like SS, ask subjects to rate quality without viewing the original image. However, for bilevel images, without viewing the original image, "quality" does not make much sense. For example, human subjects will not generally agree on the quality of the two images in Figure 1. Clearly, the left one is smoother, and the right one contains more

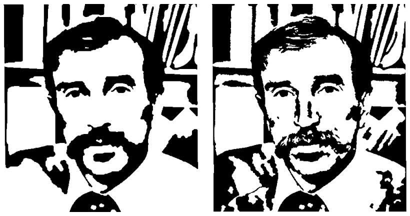

Fig. 1. Two original bilevel images

detail. Hence, instead of attempting to rate the quality of a bilevel image, we will only attempt to rate the similarity of a pair of images. In particular, one will be a distorted copy of the other. Since we are interested similarity rather than quality, in this paper, we use a modified version of simultaneous double stimulus for continuous evaluation (SDSCE) suggested in ITU-R BT.500-11 [8], as described next.

In our experiments, each distorted image, called a *test image*, is shown simultaneously side by side with its original. Subjects are told which is the original and asked to rate the similarity of the distorted image to its original by dragging a slider on a continuous scale as in [10]. As benchmarks to help subjects make good ratings, the scale is divided into five equal portions, labeled “Bad”, “Poor”, “Fair”, “Good” and “Excellent”. In addition, unlike previous work, the rating time for each image by each subject was recorded for screening purposes. However, subjects were not informed of this.

The database of test images is developed from the seven scenic images shown in Figure 2, each with size 512×512. The first six images are natural and the last one, “MRF”, is typical of an Ising Markov random field model, which has been proposed as a model for scenic images [1, 2]. Seven kinds of distortions are created, resulting in 44 distorted images for each original:

1. Finite State Automata Coding (FSA) [4] with nine error rate factors: $[1, 100, 150, 200, 300, 400, 500, 700, 1000]$.
2. Lossy Cutset Coding (LCC) [1,2] with eight grid sizes: $[2, 4, 6, 8, 10, 12, 14, 16]$.
3. Lossy Cutset Coding with Connection Bits (LCC-CB) [1,2] with the same eight grid sizes as LCC.
4. Hierarchical Cutset-MRF (HC) [3] with eight MSE thresholds for block splitting: $[0, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 1]$.
5. Random bit flipping with five different probabilities: $[0.01, 0.03, 0.05, 0.10, 0.15]$.
6. Dilation with 1, 2 and 3 iterations using a 3×3 all ones structuring element.
7. Erosion with 1, 2 and 3 iterations using a 3×3 all ones structuring element.

The original image itself is also included as a “distorted image” in order to verify that, as described later, subjects are making good faith judgments. Thus, since there are seven original images, each subject is asked to rate $45 \times 7 = 315$ images, each displayed side by side with the original at size $4'' \times 4''$. Subjects were asked to view the images from approximately 20 inches. The ordering of test images is independently randomized for each subject to avoid systematic bias that might be caused by some fixed ordering. Moreover, to avoid contextual effects (discussed later), no two successive test images will come from the same original.

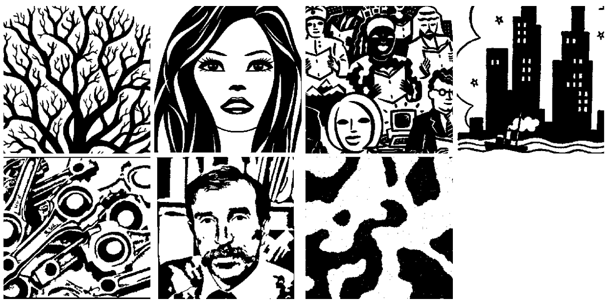

Fig. 2. The seven original images in the database: tree, woman, people, boat, tools, Alc, MRF.

Before participating, each subject was given an explanation of the purpose of the experiment and a description of the procedure. In addition, several training images, similar to actual testing images, are shown to subjects. These roughly cover the whole similarity range in the database.

3. DATA PROCESSING

3.1. Scaling the ratings

In all, 77 subjects, all non-experts, participated in the experiment. For each, raw rating data, test image order and rating times were recorded. As in [9], the raw rating data for the j^{th} image by the i^{th} subject was then scaled to reduce systematic differences in ratings among subjects and to obtain values between 0 and 1, with 1 representing highest similarity.

$$\text{Scaled}(i,j) = \frac{\text{Raw}(i,j) - \min(\text{Raw}(i,k), \forall k)}{\max(\text{Raw}(i,k), \forall k) - \min(\text{Raw}(i,k), \forall k)} .$$

From now on, we will work with scaled rating data.

3.2. Subject screening

Subject screening, such as in [8, 10], which is designed to rule out abnormal subjects and those who are just randomly rating, helps improve the quality of data. In this experiment, a subject is rejected if at least two of the following criteria are satisfied:

1. Total rating time is less than $T = 10$ minutes.
2. More than $R = 33$ outlier ratings. (Described later.)
3. At least two ratings of original images are outliers.
4. Average of the seven ratings for the original images is less than $T_{\text{ref}} = 0.5$.
5. The “monotonicity test” is failed. (Described later.)

The motivation for criteria 2 and 3 are that many outlier ratings, especially for original images, indicate abnormal behavior or careless rating. Hence the corresponding subjects should be screened out. Similar to the approach taken in [10], a scaled rating $\text{Scaled}(i, j)$ is considered an outlier if

$$|\text{Scaled}(i, j) - \text{avg}(j)| > \delta \times \text{std}(j)\,,$$

where $\text{avg}(j)$ and $\text{std}(j)$ are the expectation and standard deviation of scaled rating scores for image j by all subjects. δ is chosen to be 1.96 corresponding to a 95% confidence interval, assuming scaled rating scores are Gaussian.

The "monotonicity test" in criterion 5 is a new idea, based on the property of our database that for each type of distortion, there is a clear monotonicity in the amount of distortion with respect to some parameter, such as bit flipping probability, number of dilation/erosion iterations, and coding rate for compression. Hence, if any subject's rating scores are too far from monotonic, the subject should be screened out. Specifically, for each subject i, a penalty counter $P(i)$ is initialized to zero. Now suppose

$$[\text{Scaled}(i, n_1), \text{Scaled}(i, n_2), \ldots, \text{Scaled}(i, n_k)]$$

are k ratings that should be monotonically non-increasing for reasons such as mentioned above. Then for each $t \in [1, 2, \ldots, k-1]$ such that

$$\text{Scaled}(i, n_{t+1}) > \text{Scaled}(i, n_t)\,,$$

$P(i)$ is increased by $\text{Scaled}(i, n_{t+1}) - \text{Scaled}(i, n_t)$. If, finally, $P(i) > T_{\text{mon}} = 19$, subject i fails the monotonicity test and is screened out.

After screening as described above, seven subjects were removed. From now on, all analyses are based only on the 70 remaining subjects.

4. RESULT ANALYSIS

4.1. Rating time analysis

For the 70 subjects retained, the average rating time was 23.4 minutes, with standard deviation 8.2. Table 1 shows the average rating times for each original image. Generally speaking, average rating time increases with image complexity, which makes sense because people need more time to evaluate a complex image than a simple one.

Figure 3 shows the relationship between subjective rating scores and average rating times. The red line is a linear regression fitting. It shows that average rating time increased with image similarity, which makes sense because it becomes harder to see and evaluate distortion as image similarity increases. This suggests that the subjects made serious efforts.

4.2. Contextual effects analysis

As discussed in [8], contextual effects occur when the subjective rating of a test image is influenced by previous images presented to the subject, especially the previous test image. To check whether our testing procedure suffers from strong contextual effects, the following analysis is conducted. For each test image in each test session, we plot the relationship between:

Image	tree	MRF	woman	Alc	tools	people	boat
Time	4.00	4.10	4.21	4.56	4.64	4.72	4.93

Table 1. Average rating time in seconds for each image

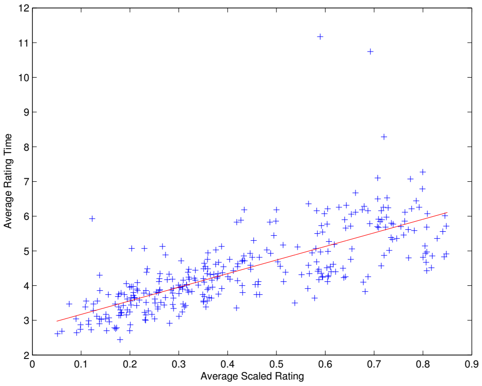

Fig. 3. Rating time analysis. Regression function: Avg. rating time $= 3.92s \times$ Avg. subjective rating $+ 2.78s$

1. The average rating score for the previous test image in the same test session.
2. The difference between the rating score of the current test image in the current test session and the average rating score for the current test image over all test sessions. This difference is called a "rating bias".

If the testing procedure does not suffer from strong contextual effects, the rating bias of the current image should have symmetric distribution around zero. The plot in Figure 4 supports the hypothesis that the testing procedure was free from strong contextual effects.

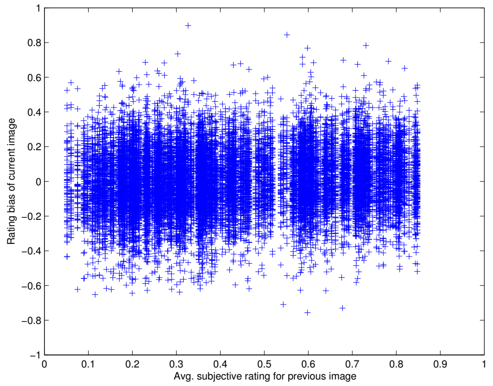

Fig. 4. Contextual Effects Study

4.3. Standard deviation of rating scores

Figure 5 presents a scatter plot showing the standard deviation of the scaled rating scores for each distorted image vs. its average rating score. The red line shows a quadratic regression fit. As one would expect, for low and high similarity images, the standard deviations of rating scores are relatively small, meaning subjects are more consistent with their judgments. However, for images with moderate similarity, the standard deviations of rating scores are relatively large, showing less agreement among subjects.

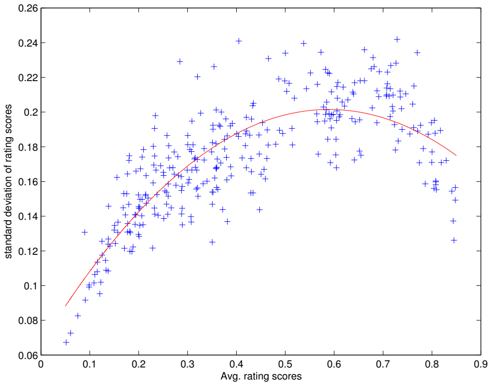

Fig. 5. Standard deviation of ratings. Regression function: Standard deviation $= -0.39 \times \text{Avg.}^2 + 0.46 \times \text{Avg.} + 0.07$.

5. APPLICATIONS

5.1. Comparing the four compression algorithms

Figure 6 shows the average scaled rating of the images produced by the four compression algorithms, plotted vs. coding rate in bits per pixel (bpp). As can be seen, HC [3] has the best performance for all coding rates. The runner-ups are two versions of Lossy Cutset Coding (LCC-CB and LCC) [1, 2]. FSA [4] has the lowest rating. Moreover, the plot for HC suggests that coding at rates between 0.04 and 0.06 bpp is quite attractive, as higher coding rates do not substantially increase subjective rating, while lower rates suffer a significant drop.

5.2. Evaluating scenic bilevel image similarity metrics

As mentioned earlier, one principal motivation for subjective ratings is to provide ground truth for evaluating existing and new bilevel image similarity metrics. To assess the performance of the two existing metrics, percentage error and SmSIM [6], we compute the Pearson and Spearman-rank correlation coefficients, after nonlinear transformation by the 5-parameter logistic function proposed in [10]. Recall that among the seven original images in our database, six are natural and "MRF" is artificially generated. Since subjective ratings of "MRF" had very high standard deviation, we used only the ratings of the six natural images in the evaluation of the two similarity metrics. Table 2 shows that percentage error has higher correlation coefficients than SmSIM, i.e., it performs better. However, neither metric fits the ground truth especially well. Hence, better metrics are needed, as for example proposed in the companion paper [7], which proposes metrics attaining significantly higher correlation coefficients on the database of the present paper.

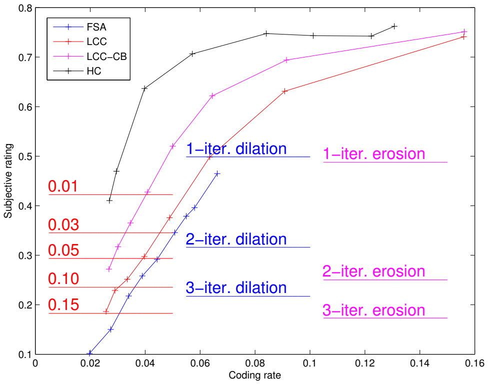

Fig. 6. Experimental results. Red: random bit flipping with different probabilities. Blue: dilation. Purple: erosion.

Table 2. Comparison of metrics

Metric	Pearson	Spearman
Percentage Error	0.84	0.81
SmSIM	0.81	0.74

5.3. Comparing the impact of different types distortions

Besides the four compression algorithms, we also included three kinds of distortions: random bit flipping, dilation and erosion. Figure 6 overlaps the average rating scores of these distortions with results of the four compression algorithms. One can see that all three kinds of distortions impact images similarity seriously. Random bit flipping with probability only 0.01 has a subjective rating score similar to HC with the lowest coding rate. Dilation and erosion with two or more iterations have very low similarity based on human perception. Another interesting fact is that subjects are more tolerant of dilation than erosion.

6. CONCLUSIONS

In this paper, we conducted subjective similarity evaluations of distorted scenic bilevel images. The experimental results are used to provide ground truth for assessing the performance of four compression algorithms, the impacts of three kinds of distortion and the goodness of percentage error and SmSIM as bilevel image similarity metrics. It is anticipated that the subjective ratings will continue to be useful in assessing future bilevel image similarity metrics.

7. REFERENCES

[1] M. Reyes, X. Zhao, D. Neuhoff and T. Pappas, "Lossy compression of bilevel images based on Markov random fields," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. II-373 - II-376, 2007.

[2] M. Reyes, D. Neuhoff and T. Pappas, "Lossy cutset coding of bilevel images based on Markov random fields," to appear in *IEEE Transactions on Image Processing*, 2013.

[3] S. Zha, T. Pappas and D. Neuhoff, "Hierarchical bilevel image compression based on cutset sampling," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. 2517-2520, 2012.

[4] K. Culik, V. Valenta and J. Kari, "Compression of silhouette-like images based on WFA," *Journal of Universal Computer Science*, 3(10):1100-1113, 1997.

[5] "Lossy/lossless coding of bi-level images," ISO/IEC Std. 14492, 2000.

[6] M. Reyes, X. Zhao, D. Neuhoff and T. Pappas, "Structure-preserving properties of bilevel image compression," *Human Vision Electr. Im. XIII*, *Proc. SPIE*, vol. 6806, pp. 680617-1-12, Jan. 2008,

[7] Y. Zhai and D. Neuhoff, "Objective similarity metrics for scenic bilevel images," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, May. 2014.

[8] ITU-R BT. 500-11, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.

[9] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," [Online]. Available: *http://www.vqeg.org*, Aug. 2003

[10] H. Sheikh, M. Sabir and A. Bovik, "A statistical evaluaIon of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, pp. 3440-3451, Nov. 2006.