Similarity of Scenic Bilevel Images

Yuanhao Zhai and David L. Neuhoff

***Abstract*—This paper presents a study of bilevel image similarity, including new objective metrics intended to quantify similarity consistent with human perception, and a subjective experiment to obtain ground truth for judging the performance of the objective similarity metrics. The focus is on scenic bilevel images, which are complex, natural or hand-drawn images, such as landscapes or portraits.**

The ground truth was obtained from ratings by 77 subjects of 44 distorted versions of seven scenic images, using a modified version of the SDSCE testing methodology.

Based on hypotheses about human perception of bilevel images, several new metrics are proposed that outperform existing ones in the sense of attaining significantly higher Pearson and Spearman-rank correlation coefficients with respect to the ground truth from the subjective experiment. The new metrics include Adjusted Percentage Error, Bilevel Gradient Histogram and Connected Components Comparison. Combinations of these metrics are also proposed, which exploit their complementarity to attain even better performance.

These metrics and the ground truth are then used to assess the relative severity of various kinds of distortion and the performance of several lossy bilevel compression methods.

I. Introduction

BILEVEL images have only two intensity levels: 0 (black) and 1 (white). The bilevel images in which we are primarily interested are *scenic* bilevel images, such as those illustrated in Fig. 1, which are complex bilevel images, typically containing natural or hand-drawn scenes, *e.g.*, landscapes and portraits, but which do not include text, line drawings or halftoned images. Silhouettes generally have a simpler form than scenic images.

Objective image similarity metrics that make predictions consistent with human perception are important for many image processing applications. For example, they can be used to assess overall performance of such algorithms, *e.g.*, compression algorithms, or they can play a role in the operation of algorithms, *e.g.*, compression and retrieval algorithms. While a number of objective similarity metrics have been developed for grayscale and color images, with the goal of consistency with human perception, and while a number of bilevel similarity metrics have been developed, there has been almost no development of objective similarity metrics for bilevel images consistent with human perception.

The most common objective similarity metric for bilevel images, is *percentage error* (PE), which for bilevel images is the same as mean-squared error (MSE). Unfortunately, this metric is not always so consistent with human perception, as images with similar percentage error often appear very different to viewers. With applications other than perceptual similarity in mind, many *intensity-based overlap* metrics have been proposed, as reviewed in [1]–[3]. Generally speaking, like PE, these penalize pixel-level disagreements, based on different assumptions about what is important in specific applications. Examples of this kind of metric include those developed by Jaccard [4], Kulczynski [5], Braun-Blanquet [6], Dice [7], and Ochiai [8]. These metrics were first widely used in biology related disciplines to group biotal communities [4] or ecologically related species [9]. Additionally, metrics like Dice [7] were used to quantify bilevel image similarity for medical image processing applications [10], [11]. While these metrics may be good for their intended applications, they were not designed to reflect human judgments of similarity. Hence, it is natural to try to design metrics that better reflect human perception. To the authors' knowledge, the only bilevel metric that attempts to reflect human perception is the SmSIM metric [12] which is based on a Markov random field model. Unlike previous metrics, SmSIM makes use of dependencies among adjacent pixels and measures the similarity of the "smoothness"/"roughness" of two images, as well as their pixel-level similarity.

Fig. 1: Seven scenic images: 'tree', 'woman', 'people', 'boat', 'tools', 'Alc', 'MRF'.

For color and grayscale images, many perceptual similarity metrics have been developed. For example, much recent work has focused on SSIM type metrics [3], [13]–[15]. Moreover, a number of metrics have been proposed just for textured images, including LBP [16], STSIM [17]–[19] and LRI [20], [21]. Such grayscale metrics can provide templates and insight for designing bilevel similarity metrics. Indeed, in some cases, they can be directly applied to bilevel images.

In this paper, several new bilevel similarity metrics are proposed based on hypotheses about human perception. Such new metrics include Adjusted Percentage Error (APE), Bilevel Gradient Histogram (GH), Connected Components Comparison (CC) and combinations of such.

In order to assess the performance of objective image similarity metrics – indeed, to enable their development – it

Yuanhao Zhai (yzhai@umich.edu) and David L. Neuhoff (neuhoff@umich.edu) are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48109 USA.

Portions of this work were presented at ICASSP 2014.

is essential to have ground truth, *i.e.*, a set of distorted images whose perceptual similarity to the corresponding original images (perceived distortion) have been subjectively rated by human viewers.

To develop such ground truth, we follow an approach inspired by studies of how to develop ground truth for grayscale images and video approaches. In particular, ITU-R BT.500-11 [22] made a thorough study of subjective experiment methodologies for videos. Several methods were suggested for different assessment tasks, including double-stimulus continuous quality-scale (DSCQS), double-stimulus impairment scale (DSIS), single-stimulus (SS), and simultaneous double stimulus for continuous evaluation (SDSCE).

Motivated by such, in this paper, we designed and conducted a subjective similarity evaluation of distorted scenic bilevel images using a modified version of the SDSCE methodology. The resulting similarity ratings are then used to assess the performance of the new similarity metrics and to compare them to previous metrics, as well as to assess the relative severity of the various kinds of distortion included in the ground truth database.

The original and distorted images used in the subjective experiments, along with the subjective rating data obtained can be found in "Bilevel Image Similarity Ground Truth Archive" at University of Michigan Deep Blue[1].

With this ground truth, it is found that the newly proposed metrics perform significantly better than previous ones, as assessed by Pearson and Spearman-rank correlation coefficients with respect to ground truth. For example, the best of the new metrics attains Pearson correlation 0.95, in comparison to 0.90 for LBP, which is the best of the previous metrics, and 0.87 for the best of the intensity-based overlap metrics.

Note that metrics for grayscale images have sometimes focused on quality and sometimes on similarity, with the latter referring to quality judged relative to a reference, for example the original image. On the other hand, we assert that it can sometimes be difficult or even impossible to judge the quality of a bilevel image without a reference, due to the fact that many scenic bilevel images are man-made or man-processed and that artists and image processors have different stylistic intentions which to some may appear as distortion, but not to others. As examples, the image 'boat' in Fig. 1 may appear to have low quality despite the fact that it is the original, the image on the left of Fig. 2 might appear to be overly smoothed, while the image on the right might appear to be overly noisy, despite each being the intended result of the ACA segmentation algorithm [23] applied to a grayscale image with parameters set differently due to different intentions. Finally, each "distorted" image in Fig. 17 could conceivably be considered to be an "original". This provides an additional motivation for focusing in this paper on bilevel image similarity metrics, rather than quality metrics.

The remainder of the paper is organized as follows. Section II reviews existing bilevel image similarity metrics. Section III proposes new similarity metrics. Section IV presents details of the subjective experiment that produces ground truth. Results of the subjective experiment are described in Section V. Section VI assesses the performance of the new metrics using the ground truth. Section VII uses the ground truth and metrics to assess the performance of several lossy bilevel compression methods and the relative severity of several types of distortion. Finally, Section VIII concludes the paper.

[1] http://deepblue.lib.umich.edu/handle/2027.42/111059.

Fig. 2: Two original scenic bilevel images.

TABLE I: List of intensity-based overlap metrics.

Reference	Metric
Jaccard, 1912 [4]	$\frac{a}{a+b+c}$
Kulczynski, 1928 [5]	$\frac{a}{b+c}$
Kulczynski, 1928 [5]	$\frac{1}{2}\times(\frac{a}{a+b}+\frac{a}{a+c})$
Braun-Blanquet, 1932 [6]	$\frac{a}{\max(a+b,a+c)}$
Dice, 1945 [7]	$\frac{2a}{2a+b+c}$
Ochiai, 1957 [8]	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Sokal & Michener, 1958 [24]	$\frac{a+d}{a+b+c+d}$
Simpson, 1960 [25]	$\frac{a}{\min(a+b,a+c)}$
Rogers & Tanimoto, 1960 [26]	$\frac{a+d}{a+d+2(b+c)}$
Sokal & Sneath, 1963 [27]	$\frac{2(a+d)}{2(a+d)+b+c}$
Sokal & Sneath, 1963 [27]	$\frac{a}{a+2b+2c}$

II. Existing Bilevel Image Similarity Metrics

The most commonly used objective metric to quantify bilevel image similarity is the percentage error, which is equivalent to mean-squared error (MSE) in the bilevel case. Even though percentage error is not always consistent with human perception, its simplicity and clear interpretation still make it the most popular metric. Motivated by trying to improve on percentage error, many intensity-based overlap metrics have been developed. These metrics are based on the overlap of the 1 and 0 regions in one image with those of the other image. They give nonnegative values with 0 representing little or no similarity and large values indicating high similarity.

Original image: 512×512 PE: 0.039 Subjective: 0.32 PE: 0.047 Subjective: 0.33 PE: 0.047 Subjective: 0.40 PE: 0.049 Subjective: 0.43

Fig. 3: Scenic bilevel images with different percentage errors (PE) and subjective rating scores.

For all but the second Kulczynski metric, they assign 1 to identical images. A comprehensive review can be found in [1]–[3]. Given two bilevel images, the value assigned by any of these metrics can be expressed in terms of the following four overlap counts: a, respectively d, denotes the number of pixels that have intensity of 1, respectively 0, in both images; b, respectively c, denotes the number of pixels that have intensity 1 only in the first, respectively second image. The most popular of the many metrics of this sort are shown in Table I. One can see that all metrics are symmetric with respect to the two input images, which may not be suitable for applications that focus more on one image than the other, such as compression, where one image is the original and the other a distorted reproduction. Among these metrics, the Dice [7] is most commonly used, especially in medical image processing. Its value is easily seen to equal

$$\frac{2}{1+\frac{|B_1 \cup B_2|}{|B_1 \cap B_2|}},$$

where B_i is the set of pixels where image i has intensity 1 and $|C|$ denotes the number of pixels in set C. One can see from this that the metric depends strongly on the overlap of the regions with intensity 1 in both images; metric value 1 implies a perfect match, and 0 implies total mismatch. Note that like one or two other metrics, it does not depend on d, which implicitly presumes that 1's are more important than 0's. In Section VI, these metrics will be compared to the newly proposed similarity metrics.

More recently another bilevel image similarity metric, SmSIM [12], was proposed based on the idea that not only should the metric assess pixel similarity (as in PE), but to reflect human judgments, it should also assess the similarity of the smoothness of the boundaries between black and white in the two images. The smoothness measure incorporated in SmSIM was based on a bilevel Markov random field model. This metric will be included among those tested in Section VI.

III. New Bilevel Image Similarity Metrics

This section proposes several new bilevel image similarity metrics, all calculated within $n \times n$ windows sliding across the image, for example, $n = 32$. This sliding-window structure is motivated, to a large degree, by the hypothesis that if the window size is of the order of foveal vision, which is the approximately two-degree-wide region[2] of clearest vision [28, p. 7], then what happens outside the window cannot mask errors within the window, whereas masking of errors can be caused by the contents of the window itself. If the window were chosen to be larger than fovial vision, then it could happen that the metric predicts masking that does not actually occur. On the other hand, if the window were chosen smaller than fovial vision, then the metric will be unable to take into account masking effects that occur outside the window but within foveal vision. We find that the hypothesis that the window's size should be of the order of fovial vision is supported by the experimental results in Section VI. Once the window size is specified, one must also specify the horizontal and vertical steps with which the window will slide across each image, which determine the *window overlapping rate*. In Section VI, we choose the overlapping rate based on experiments.

After computing the metric values $M(X_i, Y_i)$ at all window locations i in images X and Y, the final metric value $M(X, Y)$ is the average of all $M(X_i, Y_i)$:

$$M(X,Y) = \frac{1}{N_{\text{win}}} \sum_i M(X_i, Y_i),$$

where N_{win} is the total number of window locations.

In this paper, we consider percentage error (PE) to be the *baseline metric*. Though PE treats all errors in all windows equally, in fact, error visibility depends significantly on the surrounding content. For example, an error can be masked if the surrounding content is "busy" in the sense that there are many nearby black-white transitions, *i.e.*, adjacent pairs of pixels with one being black and the other white. Hence, PE can be improved by taking these effects into account. Figure 3 illustrates some shortcomings of PE. It shows an original scenic bilevel image together with four distorted versions. For each, both PE and subjective rating score are presented. Subjective rating scores, which range from 0 to 1, with 1 meaning identical, are obtained from the subjective experiment described in Section IV. Based on PE, the first distorted image is the most similar to the original, and the others are approximately equally dissimilar. However, the subjective rating scores indicate that human observers found the first two

[2]When viewing a computer monitor at 20 inches, two degrees is approximately 0.7 inches, or 70 pixels at 100 dpi.

distorted images to be significantly less similar to the original than the last two.

Each of the metrics proposed in this section is motivated by some particular hypothesis about human perception, and attempts to outperform PE in measuring bilevel image similarity.

A. Adjusted Percentage Error

The first new metric is motivated by the hypothesis that when more pixels within a window, or adjacent to it, have one color than the other, then errors in (*i.e.*, changes to) the pixels with the minority color are more visible than errors in the pixels with the majority color. Moreover, the visibility of errors in minority pixels increases as their proportion decreases. From now on we refer to the pixels having the minority color as the *foreground* F and the remaining pixels as the *background* B.

Based on this hypothesis, we define the Adjusted Percentage Error (APE) as follows. Suppose the window is $n \times n$, the size of foreground is $|F|$, the size of background is $|B| = n^2 - |F|$, the number of foreground errors is e_F, and the number of background errors is e_B. Then APE is the average of *foreground error rate* $\frac{e_F}{|F|}$ and *background error rate* $\frac{e_B}{|B|}$:

$$\text{APE} \triangleq \frac{1}{2} \times \frac{e_F}{|F|} + \frac{1}{2} \times \frac{e_B}{|B|},$$

which takes value in $[0, 1]$. Since $|F| \leq |B|$, individual foreground errors are given more weight than background errors. When $|F| = |B|$, APE = PE. For a given e_F and e_B, as $|F|$ shrinks, APE increases, consistent with the hypothesis that foreground errors become more significant as the size of foreground becomes smaller. One may also view PE as an average of background and foreground error rates:

$$\text{PE} \triangleq \frac{e_F + e_B}{|F| + |B|} = \frac{|F|}{|F| + |B|} \times \frac{e_F}{|F|} + \frac{|B|}{|F| + |B|} \times \frac{e_B}{|B|},$$

from which we see how PE emphasizes background error rate more than foreground error rate.

To link APE with intensity-based overlap metrics described in Section II, let us formulate APE using the overlap counts a, b, c and d. One may observe that in Section II, all intensity-based overlap metrics are symmetric with respect to image 1 and 2. In other words, b and c always play the same role in metrics. In contrast, APE is an asymmetric metric that focuses more on the original image, say image 1, than the distorted one (image 2). Since the foreground size $|F| = \min(a + b, c + d)$, APE can be rewritten as

$$\text{APE} = \frac{1}{2} \times \frac{b}{a + b} + \frac{1}{2} \times \frac{c}{c + d}.$$

Note that if image 2 were considered the original, then

$$\text{APE} = \frac{1}{2} \times \frac{c}{a + c} + \frac{1}{2} \times \frac{b}{b + d}.$$

from which it becomes clear that the metric is asymmetric.

We also consider two slight variations of APE:

$$\text{APE}' \triangleq \frac{1}{2} \times \frac{e_{F'}}{|F'|} + \frac{1}{2} \times \frac{e_{B'}}{|B'|}, \qquad \text{APE}'' \triangleq \frac{e_F + e_B}{|F|},$$

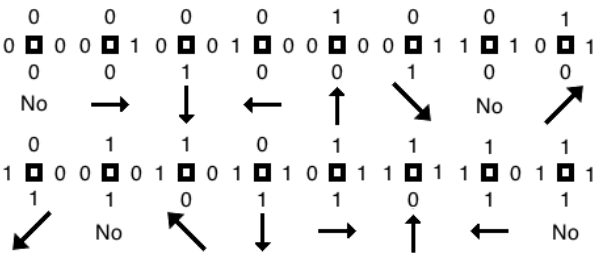

Fig. 4: Bilevel Gradient.

where F' is the one-step dilation of F using a 3×3 all ones structure element matrix, e'_F is the number of errors within F', $B' = W - F'$ denotes the remainder of the window W, and e'_B is the number of errors in B'. The hypothesis behind APE$'$ is that errors adjacent to F are as significant as foreground errors and therefore should be counted in the first term, which has the smaller denominator, rather than the second term, which has the larger denominator. APE$''$ is the ratio of total number of errors to the size of foreground. In this event, foreground and background errors are treated equally in APE$''$, just as in PE. However, the weight of errors within a window is inversely proportional to the size of foreground, so that errors within a window with small foreground count more than those within a window with large foreground.

B. Bilevel Gradient Histogram

For bilevel images, the contours between black and white regions contain most of the information. Hence, as considered in SmSIM [12], similar bilevel images should have similar contour smoothness, roughness and directionality. For grayscale images, a gradient histogram is a feature that captures such information. This is also true for bilevel images. However, a new definition of gradient is needed. With such, the similarity of bilevel gradient histograms of the original and distorted images becomes a good candidate for measuring similarity.

As the *bilevel gradient* at pixel $X(u, v)$, we propose $BG_{u,v} \triangleq \text{angle}(\underline{V}_{u,v})$, where $\underline{V}_{u,v}$ is the complex number

$$\underline{V}_{u,v} \triangleq X(u, v+1) - X(u, v-1) + j(X(u-1, v) - X(u+1, v))$$

provided this number is not zero. When $\underline{V}_{u,v}$ is zero, for example when $X(u, v)$ lies in a monotone region, there is no direction at pixel $X(u, v)$, and $BG_{u,v}$ is not defined. It follows that $BG_{u,v}$ has the eight possible values illustrated in Fig. 4, and consequently, the gradient histogram for a given window position consists of eight values $C = \{C(1), \ldots, C(8)\}$.

Clearly, the proposed bilevel gradient histogram can distinguish different directional contours. Its ability to measure contour smoothness and roughness can be seen from the example shown in Fig. 5. The left image has a smooth contour, so that all pixels along the edge have the same gradient direction, while the rough contour in the right image causes a distinctly different gradient distribution.

To measure the similarity $S(C, D)$ of the histograms C and D corresponding to the original and distorted images, respectively, at a given window location, we propose three

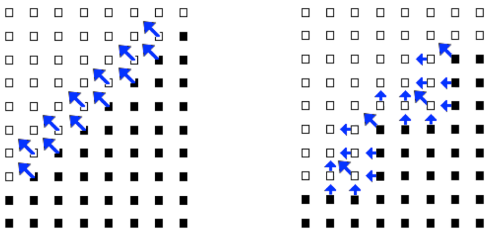

Fig. 5: Exmples of smooth and rough contours.

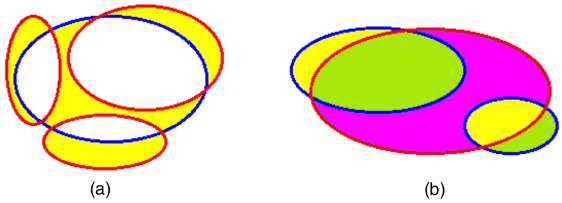

Fig. 6: Examples of CC^2 calculation.

methods. In each, a small value indicates high similarity, and to avoid singularities, we increase any zero histogram value to one.

1. $$S^1(C,D) \triangleq 1 - \prod_{k=1}^{8} \frac{2C(k)D(k)}{C^2(k)+D^2(k)}\,.$$

As each term in the product is the ratio of a geometric average to an arithmetic average (as commonly used for example in [3], [13]–[15], [17]–[20]), it is less than or equal to one, making $S^1(C,D)$ non-negative. By multiplicatively combining eight terms, we tacitly assume that a distorted image has high similarity only when all eight values are similar to the original. Hence, this is a strict measure of histogram similarity, which may over penalize some histogram differences.

2. $$S^2(C,D) \triangleq \sum_{k=1}^{8} c(k)\log\frac{c(k)}{d(k)}\,,$$

where c and d denote C and D normalized so as to sum to one. This is the Kullback-Leibler divergence [29] of probability mass function d with respect to c.

3. $$S^3(C,D) \triangleq \Big(\sum_{k=1}^{8} c(k)\log\frac{c(k)}{d(k)}\Big) \times \frac{\max(\|C\|_1,\|D\|_1)}{\min(\|C\|_1,\|D\|_1)}\,.$$

In addition to the divergence of d with respect to c, this method also considers the similarities between the L_1 norms of C and D, which approximates the total number of pixels along edges within an image window.

We denote the Gradient Histogram metric with these three similarity methods as GH^1, GH^2 and GH^3, respectively.

We also experimented with a definition of bilevel gradient that depended on the eight nearest neighbors, rather than four, yielding 8 gradient directions, and another definition yielding 16. Since the improvements in the experiments of Section VI resulting from these enhanced gradients were small, from now on we assume the gradient definition given previously.

C. Connected Components Comparison

The concept of connected components is useful in bilevel image analysis. Here, we hypothesize that distorted images should preserve the connected components of the foreground of the original. The simplest way to use this hypothesis is to compare the number of connected components in the original and distorted image windows. However, to avoid a small isolated dot adjacent to a large component from being counted as a new connected component, we do a one-step dilation with a 3×3 all ones structuring element before counting. Dilation helps connect isolated dots and islands that are close to some big connected components. We propose two methods to assess similarity using connected components.

The first compares the *effective number* of foreground connected components in a window W of the original and distorted images, where the effective number of connected components in a window W with N connected components is

$$N_W \triangleq \sum_{k=1}^{N} \min\Big(1, \frac{|cc_k|}{T_V}\Big)\,,$$

where $|cc_k|$ is the size of the k^{th} connected component and T_V is a threshold greater than 1, which increases robustness by reducing the effect of small connected components, *e.g.*, isolated dots. In this paper, $T_V = 10$. All connected components with size less than T_V contribute less than one to N_W. Now, if X is the original and Y is the distorted image at window W, the metric value is

$$CC^1 \triangleq 1 - \frac{\min(N_{W,X}, N_{W,Y})}{\max(N_{W,X}, N_{W,Y})}\,.$$

CC^1 takes values between 0 and 1, with 0 meaning identical.

The second method considers not only the number of connected components, but also errors inside or adjacent to each connected component in the original image. The hypothesis here is that a good reconstruction should not only preserve the number of connected components, but also their shapes. Suppose for some window W, the foreground connected components for the original and distorted images are $[cc_1, cc_2, \ldots, cc_{N_1}]$ and $[cc_1^d, cc_2^d, \ldots, cc_{N_2}^d]$, respectively. As explained below, the CC^2 metric value is, basically, the summation of individual metrics, CC_i^2, one for each connected component cc_i in the original.

If $N_1 > 0$, let $[cc_{i,1}^d, cc_{i,2}^d, \ldots, cc_{i,k_i}^d]$ denote all connected components in the distorted image that overlap cc_i, and define

$$CC_i^2 \triangleq \Big| cc_i \;\triangle \bigcup_{t=1}^{k_i} cc_{i,t}^d \Big| \times (|k_i - 1| + 1)^p\,.$$

where $A \triangle B$ denotes the symmetric difference between sets A and B. The term above within *size brackets* measures the total number of errors between cc_i and the union of $cc_{i,t}^d$, $t \in \{1,2,\ldots,k_i\}$. The second term penalizes the lack of any overlapping connected components ($k_i = 0$) or multiple connected components overlapping cc_i ($k_i > 1$). Parameter p, which we choose to equal to 1, controls the severity of the penalty. Figure 6(a) gives an example. The region enclosed by the blue curve is cc_i, and the distorted image has three connected components, enclosed by red curves, overlapping

cc_i. Hence $k_i = 3$, and the size of the yellow region represents the first term in the formula above. Finally, we have

$$\mathrm{CC}^2 \triangleq \sum_{i=1}^{N_1} \mathrm{CC}_i^2 + \sum_{t=1}^{N_2} \delta\left[\left|cc_t^d \cap \left(\bigcup_{r=1}^{N_1} cc_r\right)\right|\right] \times \left|cc_t^d\right|,$$

where $\delta[0] = 1$ and $\delta[n] = 0, \forall n \neq 0$. The second summation above represents the penalty for having connected components in the distorted image that are disjoint with all connected components in the original. This term is important if the distorted image has many new connected components.

Note that if the original image window is monotone, *i.e.*, it contains only background, then $N_1 = 0$, and CC^2 reduces to

$$\mathrm{CC}^2 \triangleq \sum_{t=1}^{N_2} \left|cc_t^d\right|.$$

Note also that CC^2 is closely related to PE. If for all cc_i, $k_i = 1$, and each cc_i^d overlaps only one cc_j for some j, then $\mathrm{CC}^2 = \mathrm{PE}$. However, when there are missing or split connected components, *e.g.*, Fig. $6(a)$, CC^2 will penalize appropriately.

The false connection of two or more connected components is another interesting case. As illustrated in Fig. $6(b)$, two connected components, cc_i and cc_j, enclosed by blue curves, become one connected component in the distorted image, enclosed by the red curve. The yellow region is penalized in CC_i^2 and the green region is penalized in CC_j^2. The purple region, however, is penalized in both CC_i^2 and CC_j^2. Thus, we see that a false connection is penalized multiple times.

IV. Subjective Evaluation Experiment

In previous sections, we reviewed existing bilevel image similarity metrics and proposed several new metrics. In order to compare their performance, ground truth is needed. This ground truth should consist of a collection of distorted scenic bilevel images with subjective similarity ratings to their original. This section describes a subjective experiment designed to obtain such ground truth using a modified version of simultaneous double stimulus for continuous evaluation (SDSCE) suggested in ITU-R BT.500-11 [22], as described next.

A. Experiment Design

In our experiments, each distorted image, called a *test image*, is shown simultaneously side by side with its original. Figure 7 shows an example of the screen that each subject saw during the experiment. 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 [30]. As benchmarks to help subjects make good ratings, the scale is divided into five equal portions, labeled "Bad", "Poor", "Fair", "Good" and "Excellent". Each rating is then rounded to the nearest integer between 0 and 100. 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. To make sure the recorded time information is as accurate as possible, a "Pause" button is

Fig. 7: A sample screen of the subjective experiment.

added so that subjects could take rests during the experiment without influencing the rating times. Finally, since the number of test images is large, to prevent subjects from becoming impatient, we divide the 315 test images into 15 groups and display to subjects the number of remaining groups, instead of the number of remaining images, during the whole experiment. The grouping does not influence the data processing.

During the experiment, 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 come from the same original.

The database of test images is developed from the seven scenic images shown in Fig. 1, each with size 512×512. The first six images are recognizable scenes and the last one, 'MRF', is typical of an Ising Markov random field model, which has been proposed as a model for scenic images [31], [32]. Seven kinds of distortions are created, resulting in 44 distorted images for each original:

i) Finite State Automata Coding (FSA) [33] with nine error rate factors: $[1, 100, 150, 200, 300, 400, 500, 700, 1000]$.
ii) Lossy Cutset Coding (LCC) [31], [32] with eight grid sizes: $[2, 4, 6, 8, 10, 12, 14, 16]$.
iii) Lossy Cutset Coding with Connection Bits (LCC-CB) [31], [32] with the same eight grid sizes as LCC.
iv) Hierarchical LCC (HC) [34] with eight MSE thresholds for block splitting: $[0, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 1]$.
v) Random bit flipping with five different probabilities: $[0.01, 0.03, 0.05, 0.10, 0.15]$.
vi) Dilation with 1, 2 and 3 iterations using a 3×3 all ones structuring element.
vii) Erosion with 1, 2 and 3 iterations using a 3×3 all ones structuring element.

Figure 8 shows the seven test images, each with a randomly selected distortion. Besides these distorted images, every 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.

Before participating, each subject was given an explanation of the purpose of the experiment and a description of the

Fig. 8: Seven randomly selected distorted images in the database, one for each original image.

procedure. In addition, several training images, similar to actual test images, are shown to subjects. These training images roughly cover the whole similarity range in the database.

B. Data Processing

1) Scaling the ratings: In all, 77 subjects, all non-experts, completed a session in which they rated all 315 distorted images. For each subject, raw rating data, test image order and rating times were recorded. As in [35], the raw rating data, $\text{Raw}(i,j)$, 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.

2) Subject screening: Subject screening, such as in [22], [30], which is designed to rule out abnormal subjects and those who are just randomly rating, helps improve the quality of the ground truth. In this experiment, a subject is rejected if at least two of the following criteria are satisfied:

i) Total rating time is less than 10 minutes.
ii) More than 33 outlier ratings. (Described later.)
iii) At least two ratings of original images are outliers.
iv) Average of the scaled ratings for the seven original images is less than 0.5.
v) The "monotonicity test" is failed. (Described later.)

The motivation for criteria ii) and iii) is that the presence of 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 [30], 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 v) 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, $P(i) > 19$, subject i fails the monotonicity test and is screened out of the experiment.

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

TABLE II: Average rating time in seconds for each image.

image	'tree'	'MRF'	'woman'	'Alc'	'tools'	'people'	'boat'
time	4.00	4.10	4.21	4.56	4.64	4.72	4.93

V. Subjective Evaluation Results

A. Rating Time Analysis

For the 70 subjects retained, the average rating time was 23.4 minutes, with standard deviation 8.2. Table II 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 9 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 increases 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.

Another interesting result is the average rating time, over all sessions, for the nth displayed test image, as function of n. (Recall that the order of images is randomized for each test session.) As shown in Fig. 10, the average rating time decreases from almost 30 seconds for the first test image to 4 or 5 seconds after rating around 50 test images. The decline of average rating time indicates increasing familiarity with the experiment as the test session proceeds. On average, it takes about 50 images for a subject to be fully familiar with the experiment.

B. Contextual Effects Analysis

As discussed in [22], contextual effects occur when the subjective rating of a test image is influenced by prior images presented to the subject, especially the previous test image. To check whether our testing procedure suffers from strong

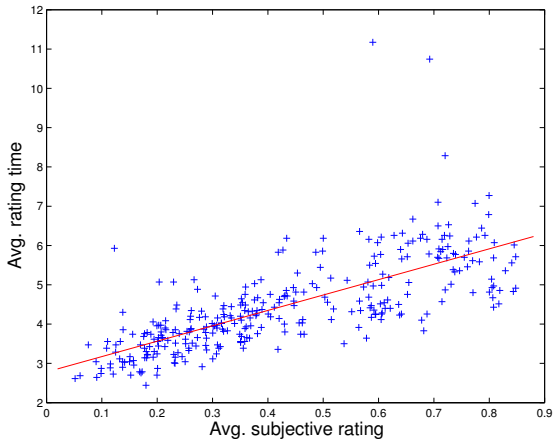

Fig. 9: Average rating time in seconds vs. average subjective rating score for the 315 test images. Regression function: avg. rating time $= 3.92 \times$ avg. subjective rating $+ 2.78$.

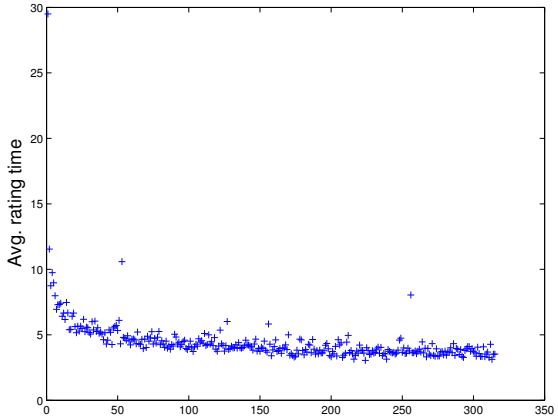

Fig. 10: Average rating time as function of n.

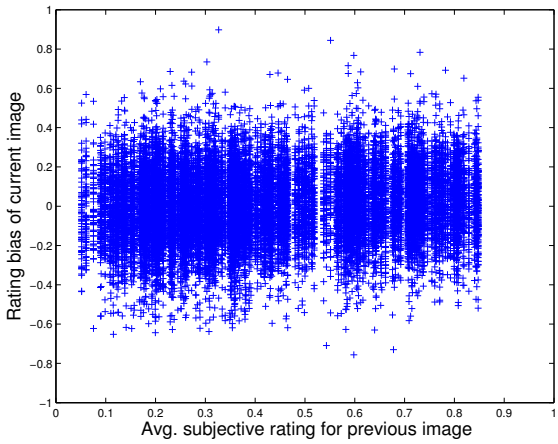

Fig. 11: Results of contextual effects test.

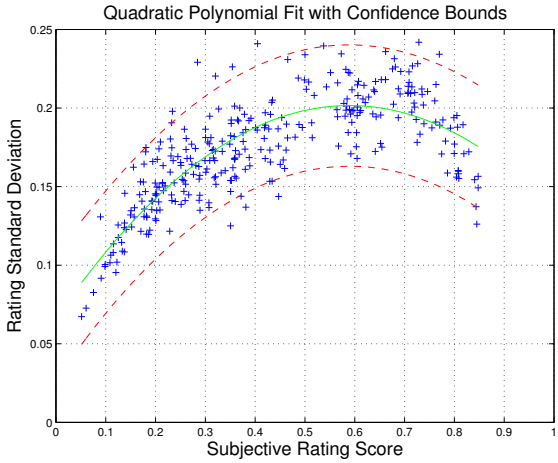

Fig. 12: Standard deviation of ratings. Regression function: std. dev. $= -0.39 \times$ avg.$^2 + 0.46 \times$ avg. $+ 0.07$.

contextual effects, the following analysis is conducted. For each test image in each test session, we plot the relationship between:

i) The average rating score (over all sessions) of the previous test image in this session.
ii) 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, no matter the average rating score of the previous test image. The plot in Fig. 11 supports the hypothesis that the testing procedure is free from strong contextual effects.

C. Standard Deviation of Rating Scores

The ratings of different images have different standard deviations. Figure 12 presents a scatter plot showing the standard deviation of the scaled rating scores for each distorted image vs. its average rating score. The green solid line shows a quadratic regression fit, with two red dashed lines giving 2σ confidence bounds. 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.

Notice that since neither the lowest nor highest average rating scores are near zero or one, respectively, it does not appear that the standard deviation estimates are affected significantly by ceiling effects [36, p. 21].

D. Insensitivity of the 'MRF' Image to Distortion

As mentioned earlier, among the seven original test images, the first six contain recognizable scenes while the last, 'MRF', does not. From the experimental results, we found that human observers are fairly sensitive to the amounts of distortion added to the first six images, but are not so sensitive to the amounts of distortion added to the 'MRF' image. Figure 13 illustrates this finding by showing the subjective rating scores of both

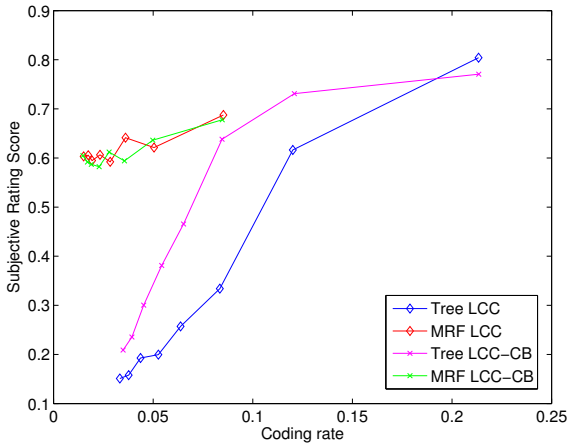

Fig. 13: Subjective rating scores for 'tree' and 'MRF' coded with LCC and LCC-CB.

the 'tree' and 'MRF' images coded with LCC and LCC-CB, respectively. As can be seen, the subjective rating scores of the 'tree' images increase monotonically with coding rate. However, this is not the case for 'MRF' images. One possible reason is that when viewing an image with a recognizable scene, observers have a "ground truth" in mind with which to compare. Hence, it is relatively easy for them to observe the effects of increasing or decreasing distortion. However, if the image contains unfamiliar or abstract content, *e.g.*, the 'MRF' image, observers may have a hard time observing changes to the distortion. For this reason, the 'MRF' image is not used in the tests of the next two sections.

VI. Tests of Bilevel Image Similarity Metrics

In this section, we analyze the performance of existing and new bilevel image similarity metrics using the ground truth obtained from the subjective experiments described above. In addition, we analyze several similarity metrics designed for grayscale images, namely, SSIM [13], LBP [16] and LRI [20]. LBP is computed using the eight surrounding pixels without interpolation. As suggested in [16], only uniform patterns with less than or equal to two 0/1 transitions are labeled. LRI-A is applied with $K = 4$ and $T < 1$, where $T < 1$ guarantees that all 0/1 transitions trigger non-zero LRI-A indices. While these other metrics were not specifically designed for bilevel images, they can obviously be applied. Generally speaking, they have considerably higher computational complexity.

The performance of each metric is evaluated using Pearson and Spearman-rank correlation coefficients to rate its consistency with the ground truth consisting of 44 distorted versions of the six images in Fig. 1 with recognizable scenes. (As mentioned earlier, we decided not to use 'MRF'.) The Pearson correlation is computed after nonlinear transformation of metric values by the 5-parameter logistic model proposed in [30] and shown below, with parameters chosen to maximize correlation for the metric being evaluated. In particular, the logistic model is:

$$Y = \beta_1 \text{logistic}(\beta_2, (X - \beta_3)) + \beta_4 X + \beta_5 \,,$$

where X is a metric value, Y is the transformed metric value and

$$\text{logistic}(\tau, X) = \frac{1}{2} - \frac{1}{1 + \exp(\tau X)} \,.$$

This is the usual strategy that avoids penalizing a metric simply for having a nonlinear relationship to the ground truth.

The next two subsections discuss the influence of window size and overlapping rate, respectively.

A. Window Size Selection

In our experiments, each metric was evaluated with a variety of $n \times n$ window sizes: $n = 8, 16, 32, 64, 128, 256, 512$. Different metrics reacted differently to changes in n. We found that APE gives the best performance with $n = 64$ and 128. For small and large n, the performance decreases. We believe this result is closely related to the size of foveal vision (2 degrees) described in Section III, which under the environment of our subjective experiment is approximately $0.7''$ or 90 pixels. We found that GH performs best for moderate window sizes ($n = 16$ and 32). On the one hand, when the window size is too small, the histogram is not robust. On the other hand, when the window size is greater than 32, the histograms naturally become more similar, even if the original and distorted images do not. The performance of CC decreases monotonically as n decreases, which is not surprising since small windows are not robust to the consideration of connected components. Finally, as a compromise, we choose window size 32×32 for all metrics evaluated in this section. However, this choice is influenced by viewing distance and image resolution, and might not be optimal if the experimental environment changes.

Note that while the intensity-based overlap metrics mentioned in Section II were originally applied globally to images, they can also be applied locally by computing and averaging metric values for windows sliding across both images, and in the results of this section, they are applied with the same 32×32 window as the new metrics.

B. Window Overlapping Rate Selection

On the one hand, if windows are not overlapped, then distortion in an image edge lying on the boundary between two windows could be missed by the metric. On the other hand, a high rate of window overlapping can significantly increase the computational load. In our experiments, we compared overlapping rates of 0%, 25%, 50% and 75% for several metrics; results are shown in Table III. We see that for all chosen metrics, as window overlapping rate increases, the performance measured by Pearson and Spearman-rank correlation coefficients either stays constant or increases by very small amounts. Since the improvements are not significant, we use non-overlapped windows from now on for both the new and existing metrics.

C. Evaluation of Metrics

This subsection compares metric performance by reporting the Pearson and Spearman-rank correlation coefficients with the ground truth using non-overlapped 32×32 windows. In Table IV, the newly proposed similarity metrics are compared to

TABLE III: Experiments with different window overlapping rates.

Overlapping rate	0%	0%	25%	25%	50%	50%	75%	75%
Metric	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
APE	0.87	0.86	0.87	0.86	0.87	0.87	0.87	0.87
GH^1	0.88	0.80	0.88	0.80	0.88	0.80	0.88	0.80
GH^2	0.92	0.88	0.92	0.88	0.92	0.89	0.93	0.89
GH^3	0.91	0.85	0.91	0.86	0.92	0.87	0.92	0.88
CC^1	0.87	0.84	0.85	0.83	0.86	0.85	0.88	0.86
LBP [16]	0.90	0.84	0.90	0.84	0.91	0.85	0.91	0.85
LRI [20]	0.89	0.84	0.89	0.84	0.90	0.84	0.90	0.85

TABLE IV: Metric evaluation (P = Pearson, S = Spearman)

Metric	P	S	Metric	P	S
PE	0.84	0.81	SmSIM [12]	0.81	0.74
Jaccard [4]	0.87	0.81	SSIM [13]	0.77	0.78
Kulczynski [5]	0.75	0.76	LBP [16]	0.90	0.84
Kulczynski [5]	0.86	0.82	LRI [20]	0.89	0.84
Braun-Blanquet [6]	0.85	0.79	APE	0.87	0.86
Dice [7]	0.86	0.81	APE′	0.88	0.80
Ochiai [8]	0.86	0.81	APE″	0.86	0.84
Sokal & Michener [24]	0.86	0.82	GH^1	0.88	0.80
Simpson [25]	0.61	0.54	**GH^2**	**0.92**	**0.88**
Rogers & Tanimoto [26]	0.85	0.82	GH^3	0.91	0.85
Sokal & Sneath [27]	0.86	0.82	CC^1	0.87	0.84
Sokal & Sneath [27]	0.87	0.81	CC^2	0.87	0.83

the existing metrics designed for bilevel images, *e.g.*, intensity-based overlap metrics described in Table I and SmSIM [12], as well as metrics designed for grayscale images, *i.e.*, SSIM [13], LBP [16] and LRI [20].

All but two of the existing intensity-based overlap metrics (the first column in Table IV), have very competitive performance. Compared to these intensity-based overlap metrics and the baseline metric PE, the proposed APE gives better performance, especially in Spearman-rank correlation coefficients. Surprisingly, SmSIM [12] performs a little worse than the baseline metric, PE, meaning it is not a very effective similarity metric. SSIM is designed to measure grayscale image quality. Results in Table IV show that SSIM does not provide satisfactory performance for scenic bilevel images. Although LBP and LRI are designed to measure grayscale texture similarity, results suggest that they are also capable of measuring bilevel image similarity.

All three versions of APE outperform PE, proving that its hypothesis is good. Specifically, the fact that APE and APE′ work better than PE and APE″ indicates that foreground errors are more visible than background errors, and should be penalized harder. The fact that APE outperforms APE′ suggests that dilation of the foreground is not necessary. Among the three versions of bilevel gradient histogram metrics, GH^1 is the worst, suggesting that multiplicatively combining eight terms may cause over-penalization. Both GH^2 and GH^3 provide very good results, suggesting that divergence is suitable for comparing histogram similarity in this application. In addition, GH^2 is the overall best similarity metric. CC^1 and CC^2 give comparable performance to APE. We know CC^2 is closely related to PE. The fact that CC^2 outperforms PE suggests that the consideration of connected components helps predict human judgments on scenic bilevel image similarity.

TABLE V: Metric combination evaluation

Overlapping rate	0%	0%	75%	75%
Combination	Pearson	Spearman	Pearson	Spearman
APE & GH^2	0.94	0.92	**0.95**	**0.94**
PE & GH^2	0.93	0.90	0.94	0.91
CC^2 & GH^2	0.93	0.90	0.94	0.91
LBP & LRI & GH^2	0.93	0.89	0.94	0.90

D. Combining Different Metrics

Since the different metrics assess complementary aspects, one can expect to attain better performance by combining them. After testing many combinations, the best ones are shown in Table V. The formula for combining metrics $X_i,\ i = 1, 2, \ldots, m$, is

$$Y = \prod_{i=1}^{m} X_i^{p_i} ,$$

where the X_i's are similarity metric values after nonlinear transformation. The best combination we found is APE and GH^2 (with $p_1 = 0.2$ and $p_2 = 0.4$), where APE measures the overall accuracy of the distorted image to the original, while GH^2 quantifies the contour similarity. The motivation behind this combination is similar to that for SmSIM [12]. Similarly, PE and CC^2 also provide accuracy information and are complementary to GH^2. The combination of LBP, LRI and GH^2 also gives comparable performance. However, as the computational load of LBP and LRI is much higher, this combination is not suggested. The fact that all of the best combinations include GH^2 suggests that the bilevel gradient histogram contains information that is important to predicting human perception of scenic bilevel image similarity.

VII. Assessing Bilevel Image Distortion

This section uses the ground truth and new similarity metrics to compare the performance of several lossy bilevel compression methods and to assess the relative severity of several types of distortion, including random bit flipping, dilation and erosion.

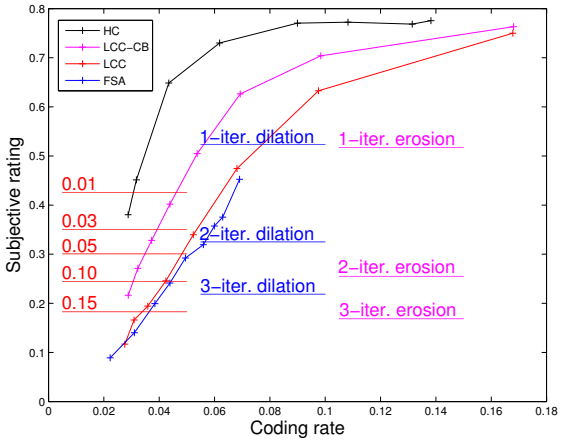

Fig. 14: Experimental results. Red: random bit flipping with different probabilities. Blue: dilation. Purple: erosion.

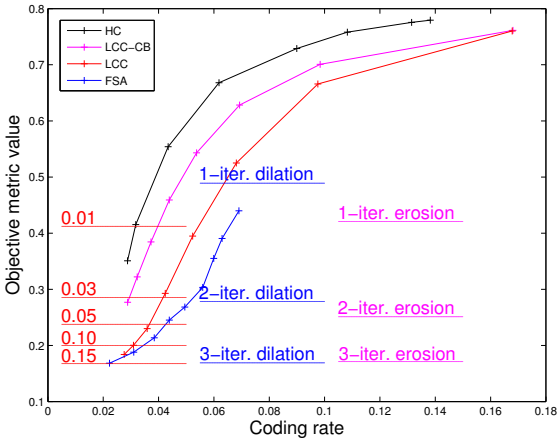

Fig. 15: Objective metric values of distorted bilevel images coded with lossy compression methods.

A. *Comparing Lossy Compression Algorithms*

As mentioned earlier, one important application of similarity metrics is to judge the performance of compression algorithms. In this subsection, we use both the metrics and the ground truth to compare the four lossy compression algorithms used in the subjective experiment, namely, Finite State Automata (FSA) [33], Lossy Cutset Coding (LCC) [31], [32], Lossy Cutset Coding with Connection Bits (LCC-CB) [31], [32] and Hierarchical LCC (HC) [34].

Figure 14 shows the subjective rating scores of the reconstructed images produced by the four lossy bilevel compression algorithms, averaged over the six images with recognizable scenes in Fig.1, and plotted vs. coding rate in bits per pixel (bpp). As can be seen, HC has the best performance at all coding rates. The runner-ups are two versions of Lossy Cutset Coding (LCC-CB and LCC). FSA has the lowest rating scores at each coding rate, although its difference to LCC at low coding rates is negligible. 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 the subjective rating scores, while lower rates suffer a significant drop.

Next, as an application of objective similarity metrics, Fig. 15 compares the same four lossy compression methods on the basis of the new metric with the best performance, namely, the combination of APE and GH^2 (computed using 32×32 windows and 75% window overlapping rate). (The metric values plotted are those obtained after the nonlinear transformation that maximizes the Pearson correlation.)

Compared to the subjective rating scores in Fig. 14, Fig. 15 preserves the relative relationship of the four compression methods. However, one can see the relative sizes of the gains from one coding method to another are not always accurately reflected in the objective metric values. For example, according to the ground truth results in Fig. 14 at rates around 0.04 bpp, the subjective rating score for LCC is a little better than FSA, LCC-CB is considerably better than LCC, and HC is considerably better than LCC-CB. Figure 15 shows the same relationship. However, the advantage of HC with respect to LCC-CB is smaller and the advantage of LCC with respect to FSA is larger.

Figure 16 shows the four 'tree' images coded at rates around 0.04 bpp with different compression algorithms. The corresponding subjective rating scores, objective rating scores and PE are shown below each image. Note that the object metric values and PE are after the non-linear transformation and are supposed to match the subjective rating scores. From this figure, one can observe several things. First, the objective metric values match the subjective rating scores significantly more than PE. Second, while according to the ground truth the FSA and LCC images have nearly the same similarity, the natures of their distortion are different. The FSA image appears noisy, while the LCC images appears to have incomplete structure. Finally, while the HC image is rated significantly higher than the LCC-CB image, the LCC-CB image actually looks quite good when viewed on its own. However, subjects did not rate it nearly as highly as the HC image because when viewed side-by-side with the original, differences can be easily seen in the LCC-CB image, but not in the HC image. We believe the differences mentioned above between the ground truth and objective metric values are partially caused by the specific form of the nonlinear transformation. They might be reduced by better transformations or by future improvements to objective bilevel image similarity metrics.

B. *Comparing the Impact of Different Types of Distortion*

One can also use the ground truth and objective similarity metrics to make judgments on the relative severity of the different types of man-made distortion that were introduced in the test images in the subjective experiment.

Figure 14 overlaps the subjective rating scores due to random bit flipping, dilation and erosion with those due to the four compression algorithms. One can see that all three kinds of man-made distortions seriously impact image similarity, even at their lowest levels, *i.e.*, they give subjective rating scores of 0.55 or less. Random bit flipping with probability only 0.01 has a subjective rating score similar to HC with the lowest coding rate. Morphological transformations, *i.e.*,

Fig. 16: 'Tree' coded at around 0.04 bpp with four different lossy compression algorithms. Starting from the second image: FSA, LCC, LCC-CB, HC. Numbers shown below each image are: subjective rating scores, transformed objective metric values and transformed percentage errors, from left to right.

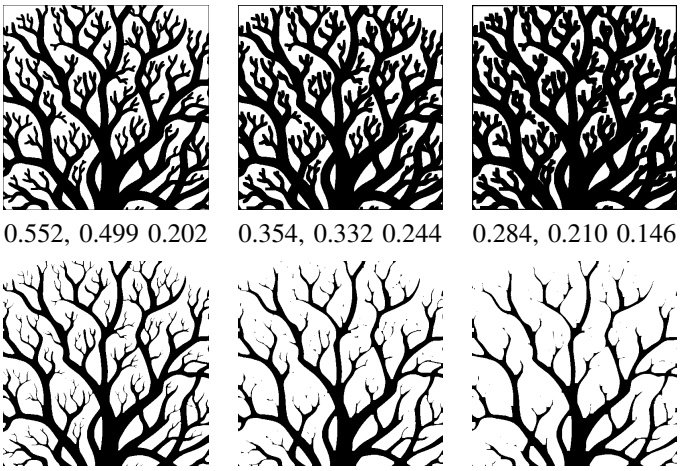

Fig. 17: Dilation and erosion added to 'tree'. First row: dilation with 1, 2 and 3 iterations. Second row: erosion with 1, 2 and 3 iterations. Numbers shown below each image are: subjective rating scores, transformed objective metric values and transformed percentage errors, from left to right.

dilation and erosion, with two or more iterations have very low similarity based on human perception. Also note that there is a large gap between the scores for one and two iterations of the morphological transformations. The gap is illustrated visually in Fig. 17, where one sees that the second iteration of dilation or erosion has a larger visual effect than the first. Another interesting fact is that people are more tolerant of dilation than erosion, which suggests that people have lower tolerance to incomplete structures.

Similarly, in Fig. 15, the objective metric values due to the three types of distortion overlap with those due to the four compression algorithms. The objective metric values, subjective rating scores and PE's are also shown in Fig. 17. One can easily see that the objective metric values match the subjective rating scores much better than PE.

VIII. Conclusions

In this paper, we presented a study of scenic bilevel image similarity, including a subjective experiment to obtain ground truth and development of new objective metrics to quantify bilevel image similarity consistent with human perception.

In the subjective experiment, seven scenic images were each distorted in forty-four ways, including random bit flipping, dilation, erosion and lossy compression. To produce subjective rating scores, the distorted images were each viewed side-by-side with the corresponding original by 77 subjects. The ratings from each subject were normalized. Several screening tests were applied to rule out subjects whose ratings were not sufficiently good. The normalized ratings were then analyzed on the basis of rating time, contextual effects and standard deviation. The result is a set of 264 subjectively rated pairs of images to use as ground truth for testing metrics and other applications.

Based on hypotheses about human perception of bilevel images, we proposed several new objective bilevel image similarity metrics. These include Adjusted Percentage Error (APE), Bilevel Gradient Histogram (GH), Connected Components Comparison (CC) and combinations of such. The performance of these and pre-existing metrics was then assessed in terms of Pearson and Spearman-rank correlation with the ground truth obtained in the subjective experiment. It was found that the GH method outperformed all previous methods, and also, that the overall best performance was achieved by the combination of APE and GH, attaining Pearson and Spearman-rank correlation coefficients as high as 0.95 and 0.94, respectively. These are significantly better than the best of the pre-existing metrics, namely, 0.90 for LBP, and 0.84 for LBP or LRI, respectively.

The ground truth and the best new metric (APE + GH) were then used to compare the performance of four compression algorithms, and to assess the severity of the various kinds of distortion.

We anticipate that the proposed similarity metrics will be useful in a number of other applications, either to judge the performance of some method, or to be used as part of the system, for example in a retrieval or segmentation algorithm.

References

[1] Z. Hubalek, "Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation," *Biol. Rev.*, vol. 57, no. 4, pp. 669-689, 1982.

[2] G.R. Shi, "Multivariate data analysis in palaeoecology and palaeobiogeography - a review," *Palaeogeogr., Palaeoclimatol., Palaeoecol.*, vol. 105, no. 3-4, pp. 199-234, Nov. 1993.

[3] M.P. Sampat, Z. Wang, S. Gupta, A.C. Bovik and M.K. Markey, "Complex wavelet structural similarity: a new image similarity index," *IEEE Trans. Image Proc.*, vol. 18, pp. 2385–2401, Nov. 2009.
[4] P. Jaccard, "The distribution of flora in the Alpine zone," *New Phytol.*, vol. 11, pp. 37-50, 1912.
[5] S. Kulczynski, "Zespoly rslin w pieninach," *Bull. Int. Acad. Pol. Sci. Letters*, vol. 2, pp. 57-203, 1928.
[6] J. Braun-Blanquet, "Plant Sociology: The Study of Plant Communities," New York: McGraw Hill, 1932.
[7] L.R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297-302, 1945.
[8] A. Ochiai, "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions," *Bull. Jpn. Soc. Sci. Fish*, vol. 22, pp. 526-530, 1957.
[9] S.A. Forbes, "On the local distribution of certain Illinois fishes: an essay in statistical ecology," *Bulletin of the Illinois State Laboratory for Natural History*, vol. 7, pp. 273-303, 1907.
[10] A. Zijdenbos, B. Dawant, R. Margolin and A. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 716-724, Apr. 1994.
[11] K. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad. Radiol.*, vol. 11, no. 2, pp. 178-189, Feb. 2004.
[12] M. Reyes, X. Zhao D. Neuhoff and T. Pappas, "Structure-preserving properties of bilevel Image compression," *Human Vision Electr. Im. XIII*, Jan. 2008, *Proc. SPIE*, vol. 6806, pp. 680617-1-12.
[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Proc.*, vol. 13, pp. 600–612, Apr. 2004.
[14] L. Zhang, L. Zhang and X. Mou, "RFSIM: a feature based image quality assessment metric using Riesz transforms," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. 321–324, 2010.
[15] L. Zhang, D. Zhang, X. Mou and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Proc.*, vol. 20, pp. 2378–2386, 2011.
[16] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution grayscale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, Jul. 2002.
[17] X. Zhao, M.G. Reyes, T.N. Pappas and D.L. Neuhoff, "Structural texture similarity metrics for retrieval applications," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. 1196–1199, Oct. 2008.
[18] J. Zujovic, T.N. Pappas and D.L. Neuhoff, "Structural similarity metrics for texture analysis and retrieval," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. 2225–2228, Nov. 2009.
[19] J. Zujovic, T.N. Pappas and D.L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *IEEE Trans. Image Proc.*, vol. 22, pp. 2545–2558, July 2013.
[20] Y. Zhai, D. Neuhoff and T. Pappas, "Local radius index - a new texture similarity feature," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1434–1438, May 2013.
[21] Y. Zhai and D. Neuhoff, "Rotation-invariant local radius index - a compact texture similarity feature for classification," *IEEE Intl. Conf. on Image Proc. (ICIP)*, Oct. 2014.
[22] ITU-R BT. 500-11, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.
[23] T.N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Tr. Sig. Proc.*, vol. 40, pp. 901-914, Apr. 1992.
[24] R. Sokal and C. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kansas Sci. Bull.*, vol. 38, pp. 1409-1438, 1958.
[25] G. Simpson, "Notes on the measurement of faunal resemblance," *Amer. J. Sci.*, vol. 258, pp. 300-311, 1960.
[26] D. Rogers and T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, pp. 1115-1118, 1960.
[27] R. Sokal and P. Sneath, "Principles of Numerical Taxonomy," San Francisco, CA: W. H. Freeman, 1963.
[28] Fairchild, Mark, *Color Appearance Models*. Reading, Mass.: Addison, Wesley, & Longman, ISBN 0-201-63464-3, 1998.
[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79-86, Mar. 1951.
[30] 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.
[31] 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.
[32] M. Reyes, D. Neuhoff and T. Pappas, "Lossy cutset coding of bilevel images based on Markov random fields," *IEEE Trans. Image Proc.*, vol. 23, pp. 1652-1665, Apr. 2014.
[33] 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.
[34] 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.
[35] 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
[36] D. Cramer, D. L. Howitt, "The SAGE Dictionary of Statistics: A Practical Resource for Students in the Social Sciences (Third ed.)", 2005.