IEEE DOI: 10.1109/ICASSP.2014.6854109

OBJECTIVE SIMILARITY METRICS FOR SCENIC BILEVEL IMAGES

Yuanhao Zhai and David L. Neuhoff

EECS Department, University of Michigan

ABSTRACT

This paper proposes new objective similarity metrics for scenic bilevel images, which are images containing natural scenes such as landscapes and portraits. Though percentage error is the most commonly used similarity metric for bilevel images, it is not always consistent with human perception. Based on hypotheses about human perception of bilevel images, this paper proposes new metrics that outperform percentage error in the sense of attaining significantly higher Pearson and Spearman-rank correlation coefficients with respect to subjective ratings. The new metrics include Adjusted Percentage Error, Bilevel Gradient Histogram and Connected Components Comparison. The subjective ratings come from similarity evaluations described in a companion paper. Combinations of these metrics are also proposed, which exploit their complementarity to attain even better performance.

Index Terms— image similarity, objective metrics

1. INTRODUCTION

Objective image quality/similarity metrics, that make predictions consistent with human perception, are important for many applications. They can be used to assess overall performance of image processing algorithms, and they can play a role in the operation of such algorithms. For color and grayscale images, many quality/similarity metrics have been developed, *e.g.*, SSIM [1]. CW-SSIM [2], RFSIM [3] and FSIM [4]. Moreover, a number of metrics have been proposed just for textured images, including LBP [5], STSIM [6–8] and LRI [9]. In perceptual image coding, similarity metrics play an important role in preserving perceptual quality while minimizing coding rate. For example, in [10], a perceptual masking model is used to determine quantization step sizes in sub-band coding. And in Matched Texture Coding (MTC) [11], STSIM2 [7, 8] and LRI [9] help reduce coding rate by measuring structural similarity between different image blocks. Such grayscale metrics can also provide insight for designing bilevel similarity metrics.

On the one hand, metrics for grayscale images have sometimes focused on quality and sometimes on similarity, with the latter referring to quality judged relative to a reference. On the other hand, as discussed in [12], it is generally impossible to judge the quality of a bilevel image without a reference. Accordingly, in this paper we focus only on bilevel similarity metrics, rather than quality metrics. While there have been many proposed grayscale similarity metrics, the only bilevel metrics of which we are aware are the widely used percentage error (PE), which is actually mean-squared error (MSE) in the bilevel case, and the recently proposed SmSIM [13]. A companion paper [12] has developed a database of distorted images labelled with subjective similarity ratings with respect to the originals and evaluated the performance of PE and SmSIM via Pearson and Spearman-rank correlation coefficients.

In this paper, new bilevel similarity metrics are proposed based on hypotheses about human perception. The new metrics perform significantly better than previous ones, as assessed by Pearson and Spearman-rank correlation coefficients with respect to the ground truth developed in [12].

In the remainder of the paper, Sec. 2 proposes the metrics. Sec. 3 discusses performance and Sec. 4 concludes the paper.

2. BILEVEL IMAGE SIMILARITY METRICS

This section proposes several bilevel image similarity metrics, all calculated within $n \times n$ windows sliding across the image, for example, $n = 32$. The average of all window metric values gives the final similarity score.

The baseline metric is percentage error (PE). Though PE treats all errors in all image windows equally, in fact, the visibility of errors depends significantly on their surroundings. For example, an error can be masked if the surroundings are "busy" in the sense that there are many nearby black-white transitions. Each of the metrics proposed below is motivated by some particular hypothesis about human perception. However, the basic metric structure of averaging sliding window metric values is motivated, to a large degree, by the hypothesis that if the window size is of the order of foveal vision, which is typically around two degrees [14, p. 7], then what happens outside the window cannot mask errors within the window. In practice, we generally find that somewhat smaller window sizes, *e.g.*, 32×32, are more effective.

2.1. Adjusted Percentage Error

The first metric is motivated by the hypothesis that within a window, errors inside or adjacent to the *foreground* are more visible than *background* errors, where the foreground is considered to be the set of all black or all white pixels, whichever is smaller, and the background is the remaining window pix-

els. Moreover, it is hypothesized that foreground errors become more visible as the size of foreground decreases.

Based on this hypothesis, we define the Adjusted Percentage Error (APE) as follows. For an $n \times n$ window, suppose the size of foreground is A, the size of background is $B = n^2 - A$, the number of foreground errors is a, and the number of background errors is b. Then,

$$\text{APE} \triangleq \frac{a}{A} + \frac{b}{B},$$

which takes value in $[0, 2]$. Since $A \leq B$, foreground errors are given more weight than background errors. When $A = B$, APE = 2×PE. For a given a and b, as A shrinks, APE increases, consistent with the hypothesis that foreground errors become more significant as the size of foreground becomes smaller. One may also view APE as the sum of *foreground similarity* $\frac{a}{A}$ and *background similarity* $\frac{b}{B}$. PE can also be expressed in such terms as

$$\text{PE} \triangleq \frac{a+b}{A+B} = \frac{a}{A} \times \frac{A}{A+B} + \frac{b}{B} \times \frac{B}{A+B}.$$

From the above, we can see precisely how PE emphasizes background similarity more than foreground similarity.

We also consider two slightly different versions of APE:

$$\text{APE}' \triangleq \frac{a'}{A'} + \frac{b'}{B'}, \qquad \text{APE}'' \triangleq \frac{a+b}{A},$$

where A' is the one-step dilation of A using a 3×3 all ones structure element matrix, a' is the number of errors within A', B' denotes the remainder of the window, and b' is the number of errors in B'. The hypothesis behind APE′ is that errors adjacent to A are equally significant to foreground errors. APE″ is the ratio of total number of errors and the size of foreground. 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.

2.2. Bilevel Gradient Histogram

For bilevel images, the contours between black and white regions contain most of the information. Hence, 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 to measure 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

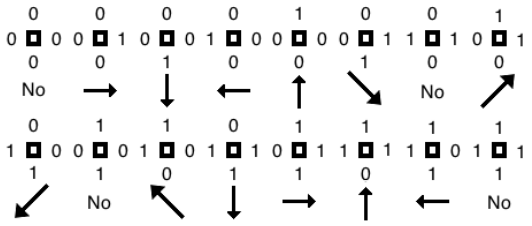

Fig. 1. Bilevel Gradient

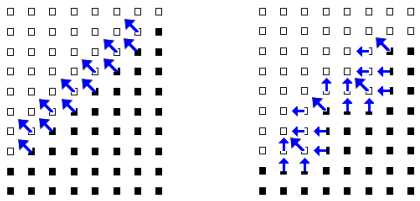

Fig. 2. Exmples of smooth and rough contours

that $BG_{u,v}$ has the eight possible values illustrated in Figure 1, 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 Figure 2. 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 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 [1–4, 6–9]), 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 method, which may over penalize some distortions.

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 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

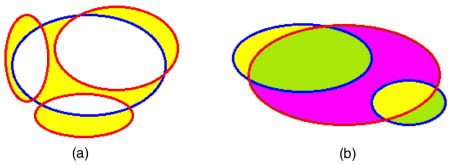

Fig. 3. Examples of CC^2 calculation

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.

2.3. Connected Components Comparison

Here, we hypothesize that distorted images should preserve the connected components 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. We propose two methods to assess similarity using connected components.

The first method compares the *effective number of 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, $T_V = 10$. Now, if X is the original and Y is the distorted image at window W, the metric value is

$$\text{CC}^1 \triangleq 1 - \frac{\min(N_X, N_Y)}{\max(N_X, N_Y)}\,.$$

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 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.

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

$$\text{CC}_i^2 \triangleq \Big|\Big(cc_i \cup \big(\bigcup_{t=1}^{k} cc_{i,t}^d\big)\Big) \setminus \Big(\bigcup_{t=1}^{k} (cc_i \cap cc_{i,t}^d)\Big)\Big| \times (|k-1|+1)^p\,.$$

The term above within *size brackets* measures the total number of errors between cc_i and the union of $cc_{i,t}^d$, $t \in [1, 2, \ldots, k]$. The second term penalizes the lack of any overlapping connected components ($k = 0$) or multiple overlapping connected components ($k > 1$). Parameter p, which we choose to equal to 1, controls the severity of the penalty. Figure $3(a)$ gives an example. The region enclosed by the blue curve is cc_i, and in the distorted image there are three overlapping connected components enclosed by red curves. Hence $k = 3$, and the size of yellow region represents the first term in the formula above. Finally, we have

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

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.

If $N_1 = 0$, then the original image window contains only background. Hence all connected components in the distorted image window should be penalized accordingly. In this case,

$$\text{CC}^2 \triangleq \Big(\sum_{t=1}^{N_2} |cc_t^d|\Big) \times N_2^p\,.$$

Note that CC^2 is closely related to PE. If for all cc_i, $k = 1$, and each cc_i^d only overlaps one cc_j for some j, then $\text{CC}^2 =$ PE. However, when there are missing or split connected components, *e.g.*, Fig. $3(a)$, CC^2 will penalize appropriately.

The false connection of two or more connected components is another interesting case. As illustrated in Fig. $3(b)$, 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.

3. METRIC PERFORMANCE EVALUATION

In this section, we analyze the performance of the metrics proposed in the previous section, based on the ground truth from [12]. As suggested in [12], the data for six natural images are used. For each of the six, the subjective ratings for 44 distorted images, created by random bit flipping, dilation, erosion and four compression algorithms, are available as ground truth. Since each similarity metric is calculated within a window sliding across the whole image, window size and overlapping rate are two parameters that affect metric performance. The performance of each metric is evaluated using Pearson and Spearman-rank correlation coefficients. Because one does not wish to penalize a metric simply for having a nonlinear relationship to the ground truth, we adopt the usual strategy, *c.f.* [15], that for each metric, a 5-parameter logistic function is optimized to nonlinearly transform metric values

Table 1. Metric evaluation (P = Pearson, S = Spearman)

Metric	P	S	Metric	P	S
PE	0.84	0.81	APE″	0.86	0.84
SmSIM [13]	0.81	0.74	GH^1	0.88	0.80
LBP [5]	0.90	0.84	**GH^2**	**0.92**	**0.88**
LRI [9]	0.89	0.84	GH^3	0.91	0.85
APE	0.87	0.86	CC^1	0.87	0.84
APE′	0.88	0.80	CC^2	0.87	0.83

to maximize the Pearson coefficient, and this same transformation is used for the Spearman coefficient.

3.1. Window size selection

In our experiments, each metric was evaluated with a variety of window sizes: $n = [8, 16, 32, 64, 128, 256, 512]$. Different metrics reacted differently to the change of n. We found that APE gives the best performance for $n = 64$ and 128. We believe this result is closely related to the size of foveal vision described in Section 2.1, which under the experimental environment in [12] 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. However, this choice is influenced by viewing distance and image resolution, and might not be optimal if the experimental environment changes.

3.2. Window overlapping rate selection

Besides window size, window overlapping rate is another important parameter. 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 will significantly increase computation. In our experiments, we compared overlapping rates of 0%, 25%, 50% and 75%. The results, some of which are given in Table 2, show that all metrics have slightly better performance with higher overlapping rates. Since the improvement is not large, we believe that non-overlapped windows will suffice for most cases.

3.3. Evaluation of different metrics

The Pearson and Spearman coefficients are shown in Table 1 for all metrics, computed with non-overlapped 32×32 windows. SmSIM [13], LBP [5] and LRI [9] are also tested. LBP is computed using the surrounding eight pixels without interpolation. LRI-A is applied with $K = 4$ and $T < 1$.

We see that SmSIM performs worse than our baseline metric, PE. Although LBP and LRI were designed to measure homogeneous texture similarity, the results show that they are also capable of measuring bilevel image similarity.

Table 2. 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

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. The fact that APE outperforms APE″ suggests that dilation of 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.

3.4. 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 2. The formula for combining metrics X_i, $i = 1, 2, \ldots, m$, is $Y = \Pi_{i=1}^{m} X_i^{p_i}$. 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 [13]. Similarly, PE and CC^2 also provide accuracy information and are complementary to GH^2. 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 bilevel similarity.

4. CONCLUSIONS

This paper proposes several objective similarity metrics for scenic bilevel images, including Adjusted Percentage Error, Bilevel Gradient Histogram and Connected Components Comparison. On the ground truth provided by the subjective similarity evaluation in a companion paper, the proposed metrics substantially outperform existing ones, attaining Pearson and Spearman-rank correlation coefficients as high as 95% and 94%, respectively. We anticipate they will be useful, for example, in judging lossy compression methods.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, Jul. 2002.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1945–1948, vol. 3, 1989.

[11] G. Jin, Y. Zhai, T.N. Pappas and D.L. Neuhoff, "Matched-texture coding for structurally lossless compression," *IEEE Intl. Conf. on Image Proc. (ICIP)*, pp. 1065–1068, Oct. 2012.

[12] Y. Zhai, D. Neuhoff and T. Pappas, "Subjective similarity evaluation for scenic bilevel images," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, May. 2014.

[13] 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.

[14] Fairchild, Mark, *Color Appearance Models*. Reading, Mass.: Addison, Wesley, & Longman, ISBN 0-201-63464-3, 1998.

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