Effective and Efficient Subjective Testing of Texture Similarity Metrics

Jana Zujovic,[1] Thrasyvoulos N. Pappas,[2,*] David L. Neuhoff,[3] René van Egmond,[4] and Huib de Ridder[4]

[1] *FutureWei Technologies, Santa Clara, CA 95050 USA*

[2] *EECS Department, Northwestern University, Evanston, IL 60208 USA*

[3] *EECS Department, University of Michigan, Ann Arbor, MI 48109 USA*

[4] *Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands*

compiled: March 10, 2015

The development and testing of objective texture similarity metrics that agree with human judgments of texture similarity require, in general, extensive subjective tests. The effectiveness and efficiency of such tests depend on a careful analysis of the abilities of human perception and the application requirements. The focus of this paper is on defining performance requirements and testing procedures for objective texture similarity metrics. We identify three operating domains for evaluating the performance of a similarity metric: the ability to retrieve "identical" textures; the top of the similarity scale, where a monotonic relationship between metric values and subjective scores is desired; and the ability to distinguish between perceptually similar and dissimilar textures. Each domain has different performance goals and requires different testing procedures. For the third domain, we propose ViSiProG, a new Visual Similarity by Progressive Grouping procedure for conducting subjective experiments that organizes a texture database into clusters of visually similar images. The grouping is based on visual blending and greatly simplifies labeling image pairs as similar or dissimilar. ViSiProG collects subjective data in an efficient and effective manner, so that a relatively large database of textures can be accommodated. Experimental results and comparisons with structural texture similarity metrics demonstrate both the effectiveness of the proposed subjective testing procedure and the performance of the metrics.

OCIS codes: (100.2000) Digital image processing; (100.2960) Image analysis; (330.4060) Vision Modeling; (330.7310) Vision; (100.4995) Pattern recognition, metrics; (110.3925) Metrics.

http://dx.doi.org/10.1364/XX.99.099999

1. Introduction

In spite of a large body of research on texture analysis, the development and testing of objective metrics for texture similarity, remain quite open. Further progress depends on a careful examination of the fundamental assumptions about the signal characteristics, the capabilities of human perception, and the requirements of the intended applications. In [1] Zujovic *et al.* presented a new class of *structural texture similarity metrics (STSIMs)* that account for the first two considerations, namely, the (typically) stochastic nature of textures and the ability of the human visual system (HVS) to perceive textures with visible point-by-point differences as similar or essentially identical (in the sense that they could be patches from a large perceptually uniform texture). This paper considers the evaluation of texture similarity metrics based on all three considerations. This entails the establishment of performance requirements for such metrics and the development of effective and efficient subjective testing procedures. The main contributions of this paper are the following: (1) We identify operating domains for similarity metrics, each of which has different performance goals and requires different testing procedures. (2) We then focus on one of these domains (separating similar from dissimilar textures) and propose a new efficient subjective testing procedure for creating ground truth for evaluating the performance of objective texture similarity metrics in this domain. We will show that the first is essential for obtaining meaningful results, and that both result in sizable reductions in the amount of subjective testing.

It is widely agreed that textures are images that are *spatially homogeneous, and that typically contain repeated structures, often with some random variation (e.g., random positions, size, orientations, or colors)* (e.g., see [2]). The statistical characteristics of texture, coupled with human perception, necessitate a different approach for the development of objective similarity metrics for textures that differs from that of traditional image similarity metrics (often referred to as quality metrics). Thus, HVS can perceive textures with visible point-by-point differences as essentially *identical* [1]. This is beyond the traditional threshold of perception, or the *just noticeable distortion (JND)* threshold,

* Corresponding author: pappas@eecs.northwestern.edu

(a) (b) (c) (d) (e)

Fig. 1. Examples of texture pairs: (a) identical; (b,c) similar; (d,e) dissimilar textures

(a) (b) (c) (d) (e)

Fig. 2. Examples of texture distortions: (a) original; (b) warping; (c) JPEG, Q=15; (d) JPEG, Q=30; (e) severe warping

below which two images are perceptually indistinguishable [3, 4]. As we move away from the threshold of perception and the domain of identical textures, differences between two texture signals can take many different forms. For example, the differences can take the form of distortions of an original texture, due to compression, graphical rendering, or the imaging chain. On the other hand, both images may be (mostly) undistorted with different texture attributes, such as scale, directionality, regularity, etc. When such attributes correspond to perceptually important differences, we will refer to them as perceptual dimensions. The goal is then to evaluate the ability of a metric to assess the similarity (or difference) between two textures. The type of difference depends, of course, on the application, for example, image and video compression, computer vision, and content-based retrieval (CBR). Since most of the applications relate to human perception of textures, it is important that metric performance agrees with human judgments.

Based on the application requirements and the capabilities of human perception, in Section 3, we identify three operating domains for evaluating the performance of similarity metrics:

1. *The retrieval of identical textures.* Figure 1(a) shows an example of *identical* textures. As we saw, they are identical in the sense that they can be considered to be patches of the same perceptually uniform texture, even though they have visible point-by-point differences.
2. *The top of the similarity scale, where a monotonic relationship between metric predictions and human judgments is desired.* Figures 2(b)-2(e) show geometric and coding distortions of the original texture in Fig. 2(a). For such distortions human subjects are generally able to provide consistent similarity ratings. It thus makes sense (and is useful) to expect that metrics provide ratings that are consistent with those of the human subjects.
3. *The ability to distinguish between similar and dissimilar textures.* Figures 1(b) and 1(c) show examples of similar textures, while Figs. 1(d) and 1(e) show examples of dissimilar textures.

Each of these domains imposes different performance goals for similarity metrics and requires different metric testing procedures. The first domain was explored in [1, 5, 6]. The second domain is addressed in [7] and a forthcoming paper. The third domain is the focus of Sections 5, 6, and 7 of this paper.

The evaluation of the performance of objective texture similarity metrics is based on comparisons between objective and subjective similarity scores. In general, this requires a large number of subjective tests over a large database (e.g., containing several hundred images). A commonly used approach is to ask subjects to rate the similarity of texture pairs, and then use traditional statistical measures to compare them to metric predictions, such as Pearson's correlation coefficient. However, the number of possible texture pairs grows quadratically with the number of textures in the database. Thus, there is a need to select a subset of pairs as stimuli for the subjective experiments. As we will see in Section 3, there are problems with random selection of texture pairs. On the other hand, our analysis of human perception abilities and application requirements indicates that a large part of the subjective comparisons is irrelevant and unnecessary. The proposed operating domains for texture similarity are essential for identifying the comparisons that are needed for obtaining meaningful results and, at

the same time, result in vast reductions in the amount of subjective testing.

The need to conduct extensive subjective tests is essentially bypassed when one considers the first operating domain, namely, the ability to retrieve textures that are identical to the query texture. This is an important problem, which is a special case of CBR called "known-item search" [1, 5]. All that is needed in this domain is to start with a database consisting of perceptually uniform textures, out of which we can cut (perhaps partially overlapping) pieces, in order to obtain the test database. Any two pieces that come from the same original (perceptually uniform) texture are then considered identical textures. Thus, the ground truth is known and no further subjective tests are required.

In Sections 5–7, we focus on the domain of distinguishing between similar and dissimilar textures, and propose an efficient procedure for conducting subjective tests and creating ground truth for testing metric performance. Our goal is to label a sufficient number of similar and dissimilar pairs of textures in a database for metric testing. A direct approach for this requires a huge number of texture-to-texture comparisons. An alternative approach that considerably reduces the amount of human subject effort is to organize the images in the database into clusters of similar textures. Assuming these clusters are visually distinct, we can then obtain pairs of similar (in the same cluster) and dissimilar (in different clusters) textures as ground truth for metric testing. The use of clustering is well established in the literature; for example, see [8–12]. To form the similarity clusters, we propose a new procedure, which we call *Visual Similarity by Progressive Grouping (ViSiProG),* whereby subjects are sequentially presented with textures and asked to form groups (e.g., of nine textures), which are progressively refined and combined with groups formed by other subjects to obtain clusters of visually similar textures. The grouping criterion is that textures blend visually, as if they came from the same tapestry. (Thus, textures with different spatial scale or orientation are dissimilar.) Semantic grouping is strongly discouraged. In this way, our approach is fundamentally different from clustering algorithms based on semantic or emotional categories [9–12], and similar to Balas's attentive texture similarity [8]. Moreover, unlike these studies, our goal is not to determine the underlying texture space, requiring the use of the whole set of textures, but just to find a sufficient number of similar and dissimilar pairs of textures.

ViSiProG makes it possible to obtain a large number of similar and dissimilar texture pairs starting from a large and essentially unconstrained database of textures. This is in contrast to traditional approaches, where a limited number of stimuli are selected for the experiment, based on a certain hypothesis. However, the drawback of such approaches is that the stimuli are designed to fit the experiment. The proposed approach allows us to use a wide range of stimuli without imposing a certain perceptual scale onto the users.

For the experimental results, we constructed a database of 505 photographic texture images that meet basic assumptions about texture signals (repetitiveness, spatial homogeneity), and include a wide variety of textures and a wide range of similarities between texture pairs. Our experimental results demonstrate that the proposed procedure collects subjective data in an efficient and effective manner, so that a relatively large database of textures can be accommodated, and a large number of similar/dissimilar pairs can be generated. We then used the results of ViSiProG to obtain clusters of similar textures, based on which we obtained a labeled set of texture pairs, which we used as ground truth for testing a number of texture similarity metrics in the similar/dissimilar domain, including peak signal-to-noise ratio (PSNR), structural similarity metric (SSIM) [13], complex wavelet SSIM (CW-SSIM) [14], STSIM [1], and the metrics by Do and Vetterli [15] and Ojala *et al.* [16]. As in [1], the metric evaluation was based on a number of statistical tests, which demonstrate that STSIM outperforms the other metrics.

The remainder of this paper is organized as follows. Section 2 reviews texture similarity metrics. Section 3 discusses an exploratory study and the operating domains for similarity metrics. In Section 4 we discuss the necessity of decoupling grayscale and color composition. Section 5 discusses the design of the subjective experiments in general, while Section 6 describes the new subjective testing procedure. The metric testing results are presented in Section 7 and the conclusions in Section 8.

Portions of this work were presented as a conference paper [17]. Some of the conclusions of this paper are also summarized in a review article [18].

2. Brief Review of Texture Similarity Metrics

In this section we review similarity metrics for grayscale textures. As we will argue in Section 4, the color composition and the spatial pattern of a texture are quite separate attributes that should be considered separately. Moreover, Zujovic and co-workers [6, 19] have argued that separating color composition and (grayscale) structure leads to more effective metrics; the color composition metrics they developed are based on the dominant colors of the textures and their percentages.

As we saw in the introduction, the main idea behind the development of texture similarity metrics is to give high similarity scores to pairs of textures that have relatively large point-by-point deviations yet according to human judgment are visually very similar or essentially identical. This can be accomplished by replacing point-by-point comparisons with comparisons of region statistics. Based on this philosophy, Wang *et al.* proposed structural similarity metrics, both in the space domain (SSIM) [13] and in the complex wavelet domain (CW-SSIM) [14]. However, due to cross-image correlations (in the "structure" term), these metrics are not completely free of point-by-point comparisons, which results in low similarity values for perceptually similar textures. To overcome such problems, STSIMs [1, 19, 20] rely only

on comparisons of statistics computed within each image. The basic elements of an STSIM are:

- A real or complex subband decomposition, typically a steerable filter decomposition.
- A set of statistics computed for each image, each subband or pair of subbands, and each window in that subband. Either a local sliding window or a global window (the entire subband) can be used. The statistics typically include the mean, variance, horizontal and vertical autocorrelations, and cross-band correlations, and can be computed on the complex subband coefficients or their magnitudes.
- Formulas for computing similarity scores for each pair of corresponding statistics, one from each image. The form that each formula takes depends on the range of values of the particular statistic and may also include a normalization factor.
- A pooling strategy for combining the similarity scores, over statistics, subbands, and window positions, to produce an overall STSIM score.

The two main variations, STSIM-2 and STSIM-M, are presented in detail in [1]. STSIM-2 computes the statistics (on a local or global window) and compares images in a similar fashion as CW-SSIM, while in STSIM-M each image is represented with a vector of its statistics and the metric computes the dissimilarity between images as the distance between their respective feature vectors. These metrics offer significant improvement over existing methods.

Here we should note that, due to the "structure" term, which is not an image statistic (as it is computed over two images) and thus is not computed as a similarity score between statistics, SSIM and CW-SSIM are not special cases of STSIMs. However, CW-SSIM can be computed using both a local sliding window and a global window. SSIM, on the other hand, becomes trivial with a global window.

We should also point out that the metrics in [1] are not scale or rotation invariant. This is consistent with our grouping criterion that textures must blend visually. However, for applications for which scale or rotation invariance are required, the metrics can be modified to account for such invariance.

3. Operating Domains for Texture Similarity Metrics

In this section we identify operating domains for testing texture similarity metrics. These are based on the capabilities of human perception and the requirements of the intended applications. As we pointed out in the introduction, the establishment of such domains is essential for obtaining statistically meaningful results as well as for vast reductions in the amount of subjective testing. As we will see, in the similar/dissimilar domain, the reductions also depend on the subjective testing procedure we propose in Sections 5 and 6. We first discuss an exploratory study, which motivated the development of the proposed approach.

(a) 5th (b) 7th (c) 23th (d) 26th (e) 50th

Fig. 3. Examples of texture pairs used in the exploratory study, ranked according to decreasing subjective similarity: (a) similar structure and color; (b) similar structure, different color; (c) similar color, different structure; (d) similar structure, different color; (e) dissimilar color and structure

3.A. An Exploratory Study

In an exploratory study of subjective texture similarity, we used 30 color texture images, organized into 50 pairs. Some examples of pairs can be seen in Fig. 3. Ten subjects were asked to rate the similarity of each pair on a scale from 1 to 10, with 10 being the highest score. The results were pooled together, and each pair was assigned the mean value of the subjective scores as its overall subjective similarity score.

For this initial study, we used STSIM-2 as an objective metric (see Section 2 and [1, 19]), computed with a sliding window of size 7×7. We also used the color similarity metric described in [19], also computed with a 7×7 sliding window. We then used a linear combination of the two metrics (appropriately normalized to yield scores in the $[0, 1]$ interval, with 1 indicating highest similarity) to calculate a single similarity score for a pair of color textures. Such linear combination is in accordance with some of the existing literature [21–23]. For comparison, we used a second objective metric, CW-SSIM, also described in Section 2 (combined with the color similarity metric of [19] in the same way as STSIM-2).

To evaluate metric performance, we used Pearson's correlation coefficient, which is typically used to measure the association between metric values and subjective scores. The results are depicted in Fig. 4, which presents a scatter plot of metric values versus subjective scores, each "standardized," i.e., converted into Z-values, which are zero-mean, unit-variance variables. Pearson's r is the slope of the minimum-mean-square error (MMSE) linear fitting of the data, for each metric. It is clear that the performance of both metrics is far away from the ideal linear relationship that Pearson's correlation presumes, i.e., that the slopes of MMSE linear fits are not close to the ideal slope of 1. More importantly, it is also evident that no monotonic curve could describe the relationship between the subjective scores and the values of either metric well. While one cannot claim that the performance shown in Fig. 4 is the best an objective metric can do, the results are indicative of the difficulties of evaluating a metric on a (by necessity) small set of subjective data, which does not necessarily capture

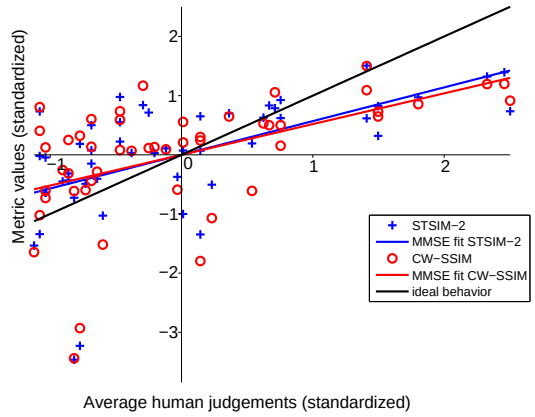

Fig. 4. Scatter plot of metric values versus average subjective scores

the complexity of the problem.

A closer look at the subjective data reveals relatively large disagreements among human judgments. One approach for measuring the consistency of the subjective scores is via the intra-class correlation coefficient (ICC) [24], which is equal to 0.66 for single measures when subjective evaluations for the entire set of texture pairs is used. This number is low, indicating low consistency among human subjects.

A careful examination of the texture pairs in the database elucidates the difficulties in obtaining consistent subjective data. If we look at the texture pairs at the high end of the subjective similarity scale, for example, the pair shown in Fig. 3(a), ranked 5^{th} most similar based on the subjective similarity scores, we see that the textures are similar in both structure and color composition (even though there is a bit of a spatial scale difference). For such pairs, our exploratory data indicate that subjects give consistently high similarity scores. As we move down the scale, we find images that are similar in some respect but different in another. For example, in the 23^{rd} pair, shown in Fig. 3 (c), they have similar color composition but different structure, while the 7^{th} and 26^{th} pairs, shown in Figs. 3 (b) and 3(d), respectively, have similar structure but different colors. In such cases, the subject-to-subject consistency is poor because the relative importance of each attribute (color, structure) differs from subject to subject in determining overall texture similarity. Similar inter-subject inconsistencies should be expected for other perceptual dimensions, such as regularity, scale, and orientation. An interesting observation is that the 7^{th} pair, shown in Fig. 3(b), was given a very high average similarity score. This is apparently because many of the subjects based their decision on the structure and essentially ignored the color difference. Finally, at the bottom of the subjective similarity sale, e.g., the 50^{th} pair, shown in Fig. 3(e), the textures are dissimilar in almost every respect. For such images, the subjective scores are generally consistently low, but the subjects do not necessarily agree in the ranking of the dissimilarities.

The main conclusion of this exploratory study was that, when the data set contains dissimilar textures or textures that are similar in some respect (e.g., structure) and dissimilar in another (e.g., color composition), subjects cannot provide consistent similarity judgments. It thus makes no sense to expect the metric to do better. An effective treatment of texture similarity requires a more careful look at the problem. We first look at human perception.

3.B. Human Perception

The capabilities and limitations of human perception can and should be used to set the expectations on metric performance. Thus, we should expect texture similarity metrics to be consistent with human performance only in the situations in which humans can provide consistent similarity judgments. We should not expect metrics to accomplish what humans cannot.

The results of our exploratory study make it clear that it is inappropriate to use standard statistical approaches, such as Pearson's correlation, to judge metric performance over the entire subjective similarity scale. It is only at the very high end of the scale, where the textures are essentially modifications of the same texture, as in the examples of Fig. 2, that we can expect a monotonic relationship between subjective ratings and metric values. In this range, we can measure the performance of objective metrics using Spearman's rank correlation coefficient, which describes how well the relationship between two variables can be described with a monotonic function, while Pearson's correlation coefficient measures how well this (arbitrary) monotonic function could be represented with a straight line. Indeed, our work with texture distortions for image compression [7] demonstrates that such a monotonic relationship exists at this end of the scale.

For the remainder of the subjective similarity scale, a much simpler task should be considered, namely, differentiating between similar and dissimilar textures. The question is whether subjects can make consistent judgments in this simpler task. According to our exploratory results, this is not possible when the textures are similar in one respect and dissimilar in another. Henceforth, such textures will be considered dissimilar. In further exploratory study, we gave subjects a relatively small number of textures (150) printed on paper cutouts and asked them to form clusters of similar textures. We found that different subjects used different criteria for forming the clusters, some relying on directionality, others on scale, color, regularity, etc. It is only when the textures are *similar in every perceptual dimension* that subjects agree that they are similar. However, given a small random selection of textures, such similar pairs are unlikely. As we will discuss in Sections 5 and 6, to form consistent clusters of similar textures, it is necessary to conduct tests with a large number of textures, and with

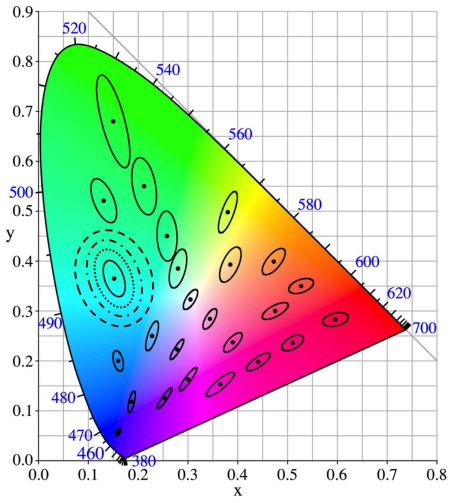

Fig. 5. MacAdam ellipses, CIE 1931 xy chromaticity diagram

explicit instructions to select textures that are similar in every respect.

In summary, the requirement for a monotonic relationship between metric predictions and human judgments should be limited to the high end of the subjective similarity scale. Over the rest of the scale, a metric should only be required to differentiate between similar and dissimilar textures. As we will see in Section 3.C, these conclusions are in line with the needs of applications such as image compression and content-based retrieval. These requirements simplify the subjective testing procedures considerably, and drastically increase the chances of obtaining consistent subjective results.

3.B.1. Analogies With Color

The above conclusion should not be surprising as even in the case of color, which is considerably simpler than texture, consistently quantifying similarity is only possible for colors that are close to each other. It does not make much sense to ask whether blue and orange are more different than green and red, as we would not get consistent answers. Thus, perceptually uniform color spaces, such as CIELAB and CIELUV, provide consistent results only over small distances.

In fact, we can extend the color analogy to the texture similarity problem. Color can be described in a three-dimensional space. The MacAdam ellipses, shown in Fig. 5 (solid lines), encompass the colors that are perceptually indistinguishable from the color at the center of that ellipse. Assuming that textures can be represented in an N-dimensional space (where each dimension represents an independent perceptual dimension), then we would have four kinds of ellipses. The first (solid lines in Fig. 5), would include the textures that are *perceptually indistinguishable* from the texture at the center of the ellipse, and thus correspond to the traditional JND threshold [3, 4]. These ellipses are the exact analog of the MacAdam ellipses. The second (dotted lines in Fig. 5), would be the ellipses that encompass all the textures that are visually *identical* to the texture at the center [1]. The third (dot-dashed lines in Fig. 5), would encompass all the textures that are small modifications/distortions of the texture at the center. In this ellipse, a *monotonic* relationship between metric values and subjective ratings is expected to exist; this corresponds to the range of colors for which the CIELAB space is approximately perceptually uniform. Finally, a fourth ellipse (dashed lines in Fig. 5) encompasses all the textures that are *similar* in every perceptual dimension to the texture at the center. In Section 5, we will propose a procedure for forming clusters of similar images, which are included in the similar texture (dashed) ellipse.

Note that the threshold for the perceptually indistinguishable ellipses is well-defined [3, 4]. The threshold for the identical ellipses is discussed in [1], while the threshold for similar versus dissimilar textures is considered in Section 7 [as its choice is guided by the receiver operating characteristic (ROC) curves]. Finally, the threshold for the monotonic region depends on the application and user preferences, and is expected to be more difficult to establish.

3.C. Metric Performance in Different Applications

We now turn from human visual system abilities to the requirements of key applications that make use of texture similarity metrics. As we saw in the introduction, a variety of applications can make use of texture similarity metrics, and each imposes different requirements on metric performance.

For applications such as image compression it is important to have the correct ordering of images according to perceived similarity, that is, to have a monotonic relationship between measured and perceived similarity (or distortion). For example, it is important to quantify the distortions in Figs. 2 (b)-(e) relative to the original in Fig. 2 (a). This holds true for both applying the metric to image quality assessment and for using it as a tool within a compression algorithm. However, this is only needed up to the point where the distorted images are no longer of acceptable quality; beyond that point it should be sufficient that the metric gives a low value. In addition, at the high end of the scale, it is important to have a threshold for identical textures and an absolute scale for image similarity, so that consistent quality can be achieved across different types of content, both within an image and across different images.

In CBR applications, the foremost task is to distinguish between similar and dissimilar images, while the precise ordering of the retrieved images may also be useful but of lesser importance. Thus, the texture pairs in Fig. 1 (b) and (c) should be labeled as similar, and the pairs in Fig. 1 (d) and (e) as dissimilar. The ability to retrieve identical textures (Fig. 2 (a)) is an important special case. The task is to retrieve a certain number of images from a database that are most similar – preferably, but not necessarily, in order of similarity –

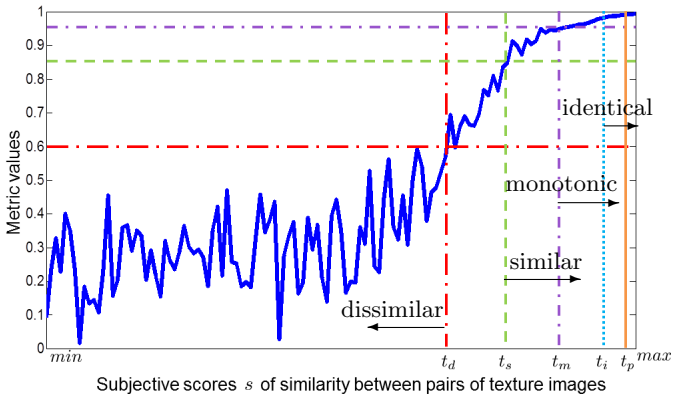

Fig. 6. Conceptual plot of desired metric performance: metric values vs. subjective similarity scores

to the query. However, the ordering may be useful at the high end of the scale. An absolute threshold may also be needed for determining whether any of the retrieved textures are sufficiently similar or identical to the query in the database. Similarly, for computer vision and image understanding, qualitative distinctions (similar/nonsimilar) may be more important than quantitative ones.

3.D. Operating Domains

We thus see a convergence between human perception abilities and application requirements. Figure 6 schematically illustrates the desired performance of a texture similarity metric, normalized to yield scores in the $[0, 1]$ interval, with 1 indicating highest similarity. It plots objective metric values of texture pairs versus subjective scores s, ordered from lowest to highest. For the sake of argument, we assume the subjective scores represent the average or ideal observer. We also assume that the horizontal axis includes all possible pairs of natural textures. In practice, it would include all pairs of textures in a database. Thus, for a database of N images, the x-axis would contain a total of $\binom{N}{2}$ points. We have identified the following important regions of interest in the plot.

At the top of the subjective similarity scale, subjects are able to assign consistent similarity values. This is the region where the textures are "essentially the same," even though substantial differences (including distortions and identical textures) may still be visible. A *monotonic* relationship between subjective similarity scores and metric values is desirable in this region, to the right of the dot-dashed violet line ($s > t_m$). This is the primary focus for compression applications, but a monotonic relationship may also be useful for CBR applications.

In the broader region of *similar* pairs, to the right of the dashed green line ($s > t_s$), the subjects agree that the textures are similar but may not assign consistent similarity values. The metric requirement is generally to assign high values in this area. At the other end of the scale, we have the region of *dissimilar* pairs, to the left of the dot-dashed red line ($s < t_d$). In this region, the subjects agree that the textures are "substantially different," even though they may not assign consistent similarity values. The metric requirement is simply to assign low values in this area; any other constraints on metric behavior in this region are relaxed. A clear gap between the metric values assigned to similar and dissimilar pairs is desired. This is the primary need of CBR and image understanding applications, but as we saw, this is also important for compression applications.

The region between the dot-dashed red line and the dashed green line ($t_d < s < t_s$) is the "non-consensus area," where the subjects cannot agree whether the texture pairs are similar or dissimilar. This should be a transition area for the metric, without any other strict requirements.

Finally, we should add the regions of *identical* (to the right of the dotted blue line, $s > t_i$) and *perceptually indistinguishable* (to the right of the solid orange line, $s > t_p$) textures. As we discussed, the former is important for identical texture retrieval applications; the metrics should assign very high similarity scores in this region. The latter is important for perceptually lossless compression; the metrics should give the highest value in this region.

The plot emphasizes the fact that a monotonic relationship between metric predictions and human judgments in the range of dissimilar textures is not important, and is perhaps unachievable. Even in the range of similar textures, for many applications, it may not be important for the metric to provide a monotonic relationship between subjective and objective values. Thus, the relationship between subjective and objective similarity needs to be monotonic only at the high end of the subjective similarity scale. Our analysis removes unnecessary constraints on metric performance and eliminates irrelevant experiments, thus turning a problem that requires an enormous human effort into a tractable one.

We now summarize the three operating domains where a similarity metric can be tested:

1. The metric ability to retrieve identical textures.
2. The top of the subjective similarity scale, where a monotonic relationship between metric values and subjective scores is desired.
3. The metric ability to differentiate perceptually similar and dissimilar textures.

Each of these domains imposes different performance goals for similarity metrics and requires different metric testing procedures. Sections 5 to 7 are going to focus on the third domain.

4. Decoupling Grayscale and Color Composition

In the discussions of the previous section, we saw that different subjects weigh color composition and structure differently when assessing overall texture similar-

ity. Moreover, their weights may depend on the application and the experimental settings [21, 23]. This suggests that we separate metric development and testing for these two attributes. In addition, one can argue that the same should be done for each perceptual dimension. Indeed, the development of metrics for individual dimensions (directionality, roughness, glossiness, scale, etc.) may be the only way to make quantitative assessments for dissimilar textures. However, deriving and testing similarity metrics for different perceptual dimensions is beyond the scope of this paper. On the other hand, there are several reasons for separating color and (grayscale) structure for metric development and testing:

- It is easy to eliminate color. (The converse is not true [6].)
- Color (in combination with structure) provides strong semantic clues, which we are trying to avoid.
- Eliminating color, substantially increases the domain of textures that are similar without being identical textures.
- We can develop more effective metrics when we separate color composition and grayscale structure [6, 19].

Here we should clarify, that by *color composition* we mean the dominant colors of an image – as perceived by a human, not a histogram of the colors of the individual pixels – without regard to their spatial arrangement in the texture; and by *structure* we mean the form of the spatial arrangement of these colors. One way of isolating the structure of a texture is by looking at the grayscale component of the image. Of course, this ignores any structure in the chrominance components. On the other hand, in most natural textures, the grayscale component is fairly representative of the overall structure.

For the subjective procedure and experimental results we present in the following sections, we will focus on grayscale textures and examine the performance of grayscale similarity metrics like the ones reviewed in Section 2.

5. Design of Subjective Experiments for Similar versus Dissimilar Texture Pair Labeling

We now turn our attention to the problem of identifying pairs of perceptually similar and dissimilar textures. A large number of texture pairs that are labeled as similar or dissimilar can serve as ground truth for metric testing. For this, it is necessary to conduct subjective experiments with a large database of textures. The selection of the database and the subset of texture pairs from the database should ensure that we have a reasonable sampling of similar and dissimilar pairs.

Once the database of textures has been selected, the process of separating similar texture pairs from dissimilar ones can be done in a number of different ways. The most straightforward approach is to perform a single-stimulus binary forced-choice test, where the subjects are asked to rate each pair in the database as "similar" or "dissimilar." Alternatively, we can ask subjects to assign a numerical value (e.g., in the range 1 to 10) to the similarity of each texture pair, but this offers little advantage if our goal is to identify similar and dissimilar pairs. However, in both cases, the number of comparisons grows quadratically with the number of textures in the database: $N(N-1)/2$ pairs for a database of N images. For typical values of N (in the hundreds) this would require far more subject time than is feasible. Accordingly, a far more efficient approach is needed.

An even more laborious approach is to ask subjects to compare two texture pairs at a time and choose the most similar (two-alternative forced choice). This can be used to obtain numerical similarity values [25]. However, in this case, the number of comparisons grows quadratically with the number of pairs in the database, i.e., is proportional to N^4.

Another approach is to conduct the test as multiple rank order judgments [26], where the subjects are asked to rank a small subset of images based on the similarity to a given image from the database. The database is partitioned in such a way that each subject makes a judgment on the similarity between all possible pairs of images. Using this method the subject gets to compare all the pairs in the database with a reduced number of trials, since they are simultaneously comparing a few images, as opposed to comparing them one pair at a time. This method has been applied in image similarity experiments and can reduce the needed number of comparisons by four [27]. However, it still requires lengthy tests and many subjects when the database under consideration is large.

A more efficient procedure for discriminating between similar and dissimilar texture pairs is a *texture clustering experiment,* whereby the subjects are asked to form well-separated *similarity clusters*, i.e., sets of images that are similar to each other and dissimilar from images in other sets. Since most texture pairs in the database are dissimilar, clustering avoids a large number of unnecessary comparisons, concentrating on the most meaningful ones [8]. Indeed, clustering has been used in a number of prior studies [8–12]. However, while such a test is well-defined and easy to carry out for a relatively small set of images, when the number of images is in the order of several hundred or more, practical problems arise, as it is difficult for subjects to see images in multiple clusters simultaneously. This is particularly difficult when the experiment is conducted electronically and the database is too large to be presented in its entirety on a single computer screen.

To alleviate this problem, in the next section, we propose a progressive testing scheme, named ViSiProG. The key idea is that when the database is relatively large, i.e., too large to be presented on a single computer screen, it will be easier for the subjects to form small similarity *groups* one at a time, in a step-by-step fashion, picking similar images out of a small set of images, and repeating

the process with a new set that contains the group and a new set of images, progressively refining the similarity group. Since the subject cannot see the entire database, it is very difficult to decide how large each cluster should be. By necessity, then, the subjects are asked to form similarity groups of a predetermined size. The final clusters are obtained by combining and analyzing multiple groups formed by several subjects.

An alternative approach for handling a large database is to use hierarchical clustering [9–11]. However, hierarchical clustering is suboptimal because as we subdivide the clusters, the clustering criteria change, which may necessitate reassignment of the entire database. Another alternative is to base the initial cluster formation on a subset of the database [8] or on a number of anchor stimuli [9], but this requires previous knowledge of the database.

As we stated in the introduction, our goal is to label a sufficient number of similar and dissimilar pairs of textures in the database for metric testing, not to obtain all possible similarity clusters for a complete texture space characterization. Texture space characterization has received considerable attention in the literature [8–10, 28, 29]. Using ViSiProG as the basis for texture space characterization would be a natural extension of the proposed techniques, but is beyond the scope of this paper.

6. Visual Similarity by Progressive Grouping

The goal of the ViSiProG procedure is to form clusters of similar textures. A number of subjects perform ViSiProg a number of times, each called a *trial* and each producing one group of N_g (e.g., 9) similar images. As will be described, the groups formed from a number of trials by a number of subjects are then pooled and analyzed (e.g., merged and pruned) to form the clusters.

The main idea of ViSiProG is that a subject working at a computer terminal is sequentially presented with subsets of the database and builds a similarity group in a step-by-step fashion. The procedure consists of a series of *rounds*, where the goal in each round is to form a group of N_g most similar images. In the first round, the subject is presented with a *batch* of N_b images, as shown in Fig. 7. Once the similarity group is formed, as illustrated in Fig. 8, a new round begins with a new batch of images. Typical values are $N_g = 9$ and $N_b = 36$ images. A key to the procedure is that the similarity group is visually separated from the rest of the batch, as illustrated in Fig. 8. The idea is that when the group is displayed in this way, it is easy for the subject to visualize how similar the images are. As the procedure moves on and the subject is presented with different images in the database, the subject converges to a stable group. The trial does not end until the subject has seen all the images in the database and is satisfied with the current similarity group.

The choice of the initial batch of images shown to the subject greatly influences the final group formation. Therefore, for each trial, the initial batch presented to the subject is randomly chosen from the database, so that the chance of obtaining a different group in each trial is increased.

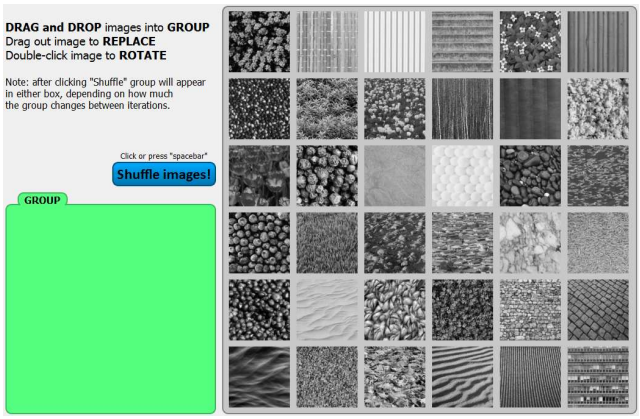

Fig. 7. Snapshot of ViSiProG interface: This is the first batch, before the subject has selected any images.

Although the focus of our experimental results is on grayscale textures, ViSiProG can be applied to similarity grouping according to other criteria, such as color composition.

6.A. Detailed ViSiProG Testing Procedure

Recall that N is the total number of images in the database, N_b the number of images shown to subjects in a batch, and N_g the number of images that are chosen to be in the group. In each round the subject has to pick N_g images out of N_b presented. Typical values are $N = 500$, $N_b = 36$, and $N_g = 9$.

A trial begins by randomly selecting a batch of N_b images from the N images in the database. Initially, all the images have the same probability of being selected. The subject then selects N_g images from the batch to form a group (of images that are as similar as possible) in a separate box, colored green and labeled "GROUP," using the mouse to drag the images into the box. Figs. 7 and 8 show the snapshots before and after the user forms the group. As we discussed, visually separating the group within the green box from the remainder of the batch enables the subject to better visualize the similarity of the images. Once an initial group is formed, the subject has the option to replace individual images in the group with other images in the batch, until she/he is satisfied that the group represents the N_g most similar images in the batch. The subject can then press a (blue) button to start a new round, i.e., to request a new set of N_b images, out of which N_g are the ones previously selected to form the group, and $N_b - N_g$ images are randomly selected from the remaining $N - N_g$ images in the database. This new set of images is shuffled before it is presented to the subject. This process is repeated several times. However, after the first round, the selection process is not uniform; the probability that an image

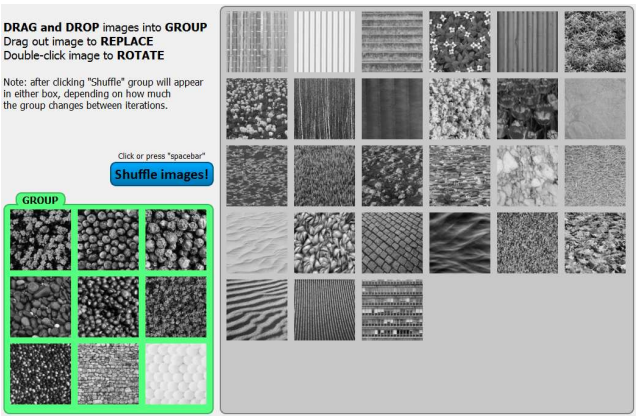

Fig. 8. Snapshot of ViSiProG interface: This is after a group has been formed and the subject is ready to ask for a new batch of textures.

will show up in the new batch is inversely proportional to the number of times it was selected to show up in a batch, i.e., the number of times the subject has seen and rejected it. This holds for all the subsequent rounds, ensuring that the subject will be able to see all N images without going through too many steps in the procedure.

From the second round on, every step is essentially the same. The subject is presented with a batch of N_b images, N_g from the previous group and $N_b - N_g$ newly selected ones, and she/he can use any of the N_b to form the current group. There is no restriction on the number of images from the previous group the subject has to keep, i.e., she/he can completely change the groups between two rounds if she/he can find a more similar set of images. However, if the groups in two consecutive rounds overlap by more than 50%, in the following round the chosen group will stay together in the green group box. Otherwise, if the overlap is less than 50%, in the following round all of the N_b images will be shuffled before they are presented to the subject, and the green group box will be emptied, so that the grouping will start from scratch. This feature exists to ensure that the subject does not feel forced to refine the group she/he selected in the first round but is allowed to drift until converging to a stable group.

Figure 9 shows a sequence of (partial) snapshots of the ViSiProG interface. Note how the group is progressively refined with textures that blend visually, and how the cohesiveness of the group increases from left to right. In fact, the green border between images facilitates the visual blending by masking discontinuities that would be apparent if the images were adjacent. Here, we should point out that for illustrative purposes the textures in Fig. 9 have been kept in the same place in the green group box; this is not guaranteed by the interface.

The rounds continue until the subject has been exposed to all the images in the database. When every image in the database has been displayed at least once, the subject is given the option to end the procedure. This is indicated by the appearance of a special (magenta) button, which can be seen in the rightmost picture in Fig. 9, and which the subject may click to finish the procedure. However, the subject is allowed to continue the rounds indefinitely, until she/he is satisfied with the similarity group or loses hope of improving it. In the latter case, the subject can always replace the images in the group with a new set and start over. Thus, in effect, the test has only a lower bound on duration – it may not end until the subject has seen all the images. Note that toward the later stages of the test, before or after the red button appears, when the similarity group has stabilized, the subject is more or less "hunting" for textures to refine the group.

6.B. Modified ViSiProG to Obtain "MacAdam" Ellipses

The ViSiProG procedure can be easily modified to obtain the similar texture ellipses of the MacAdam analogy we discussed in Section 3. These ellipses encompass all the textures that are similar in every perceptual dimension to the texture at the center of the ellipse. Given such a seed texture, we can build the cluster of similar textures by asking the subjects to form groups of textures similar to the given one. That is, the given texture stays in (the middle of) the green "GROUP" box, while the subject selects textures to form a visually cohesive group. This process can be run for any given texture, as is assumed in the classical MacAdam ellipses.

The advantage of this process is that if all the subjects start with the same seed texture as they form groups, the chances that the seed texture will end up at the center of the cluster are very high. Thus, when comparing the seed texture with the textures in the cluster, the subjective similarities will be (most certainly) higher than the similarity of the seed to any other texture in the database. In contrast, in the original ViSiProG described in Section 6.A, two textures at the opposite ends of the ellipse may be less similar than two textures near the border of the ellipse, one inside and one outside. On the other hand, we must limit the texture queries in the metric evaluation tests to the seed of each cluster, whereas any texture of the cluster can be used as query in the original ViSiProG. In addition, if the goal is to form well separated clusters, this process would require that the seeds are well separated, while the original ViSiProG automatically results in well-separated clusters.

6.C. Cluster Formation

To construct the desired similarity clusters from the groups produced by the subjects, we first form a graph whose vertices are the texture images, and whose edges (connecting vertices) have weights proportional to the number of times the adjacent images were placed by a user in the same group. These weights can be stored in a *weighted graph adjacency matrix,* which is essentially

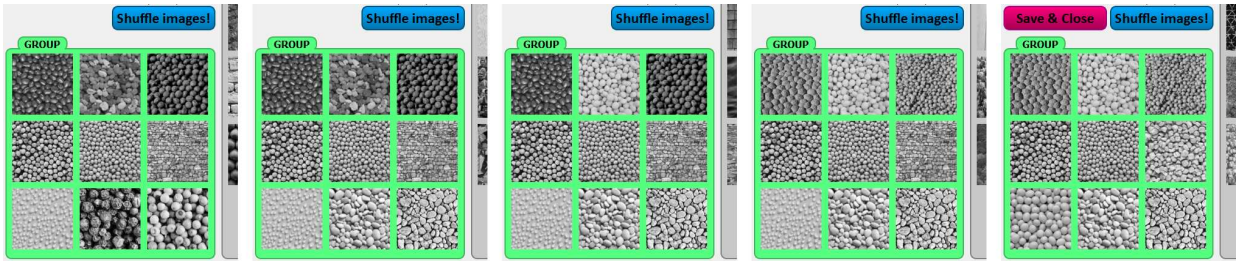

Fig. 9. Sequence of snapshots of ViSiProG interface showing how the formation of a group evolves until the user is given permission to end the test.

the same as what is often referred to as the *similarity matrix*.

In order to eliminate noise and to obtain clearly separated clusters, we set to zero the edge weights that are below a certain threshold. The thresholded weighted graph adjacency matrix can then be analyzed using the spectral clustering algorithm, whose detailed description is given in [30]. This algorithm identifies the disconnected components of the graph, corresponding to non-overlapping clusters of images, or, if all the images are connected, it can partition the data into a desired number of clusters.

Alternatively, the similarity matrix can be analyzed by multidimensional scaling [31, 32], which places the texture images in a multidimensional space, in which distances are inversely proportional to image similarity.

7. Experimental Results

In this section, we present experimental results with a database of textures and the use of ViSiProG to form clusters of subjectively similar textures. We then present the use of these clusters as the ground truth for testing a number of texture similarity metrics. For that we used a number of standard statistical tests, information retrieval statistics and ROC curves, to evaluate metric performance. We first discuss the construction of the database.

7.A. Database Construction

To construct the database, we collected 505 color texture images, from the *Corbis* website [33]. The resolution varied from 170×128 to 640×640 pixels. Out of each of those images we extracted one 128×128 pixel patch to form a database of equally-sized texture images. As was explained in [1], the images were carefully selected to meet some fundamental assumptions about texture signals. For that we used the following commonly used definition of texture (e.g., by Portilla and Simoncelli [2]): *an image that is spatially homogeneous and that typically contains repeated structures, often with some random variation (e.g., random positions, size, orientations or colors).* The textures we collected met the requirement of repetitiveness (at least five repetitions, horizontally or vertically, of a basic structuring element) and spatial homogeneity. In some cases, in order to meet the repetitiveness requirement, we had to downsample the image, typically by a factor of 2. All of the textures are photographic, mostly of natural or man-made objects and scenes. No synthetic textures were included. The database includes a wide variety of textures and a wide range of similarities between texture pairs. In contrast to the experiments described in [1], our database did not include any identical texture matches. Finally, for the texture metric experiments, the 505 test images were converted to grayscale.

7.B. Cluster Formation

The goal of our experiment was to form clusters of similar textures. This experiment was carried out using the ViSiProG procedure, explained in detail in Section 5. As we saw above, the dataset contained a total of $N = 505$ grayscale texture images. The users were presented with $N_b = 36$ different texture images in a batch, and their task was to form a group of $N_g = 9$ similar textures in each run of the test. A graphical user interface (GUI) was developed using Qt application program interfaces (APIs).

The selection of the number of textures in a batch ($N_b = 36$) was determined by the number of textures of a certain quality that can be displayed on the screen. The actual number is not very important because the user sees all the textures during the cluster formation and is free to pick the ones that form the tightest cluster. Moreover, even if a huge screen were available, looking at a large number of textures can be overwhelming, and would not necessarily make going through the entire database faster.

The number of textures in a group ($N_g = 9$) was selected to be large enough to facilitate the building of the similarity matrix in a reasonable number of trials. On the other hand, a smaller number makes the formation and refinement of a similarity group easier. Again, the actual number is not very important. What is important is that the number is fixed. Allowing the user to collect an arbitrary number of textures in the group may result in different users having different cutoff criteria for determining the size of the group. Forcing them to pick a relatively small number of textures ($N_g = 9$) ensures that they pick the most similar textures. In this sense, this is analogous to, or a generalization of, forced alternate choice. Note that there is no guarantee that the

collection of textures in the groups formed in all trials by all users will not miss any textures of a particular similarity cluster. However, the procedure guarantees that the formed clusters will be different from each other.

The subjects were instructed to form the similarity groups based on the overall similarity of the images. The emphasis was on "visual blending" of the images in the similarity group, which implies similarity in all perceptual dimensions, in agreement with our conclusions of Section 3. The subjects were instructed to *ignore any semantic information* they could extract from the images, e.g., images of flowers that look different should be classified as dissimilar, and images of different things that look similar should be classified as similar. Such emphasis on appearance rather than semantics can also be found in Balas's texture similarity study [8].

The total number of subjects was 33. Each subject was asked to perform four trials of the ViSiProG test, each resulting in one subjective similarity group. However, not all the subjects performed four runs of the test. The total number of runs (and resulting groups) was 126. In order to increase the number of candidates for the similarity groups, the subjects were allowed to rotate the images by increments of 90° to align textures' rotations in order to find more matches.

Regarding the influence of the initial batch on the final group formed by a user in a trial, we calculated that the average number of texture images belonging both to the first batch and the final group was 1.46 out of 9, while the average number of texture images belonging both to the first group and the final group was 1.38 out of 9. In addition, in 24% of the trials, no image from the first batch ended up in the final group, while in 30% of the trials, no image from the the first group ended up in the final group. These numbers indicate that the first batch has an influence on, but does not determine, the final group. Note also that even though the random selection of the initial batch of textures presented to the user usually leads to different groups in different trials, there is no requirement or guarantee that a user cannot form groups that are similar to each other.

After collecting all the groups from all the subjects, we formed the weighted graph adjacency matrix. For the construction of the weighted graph adjacency matrix, we discarded the rotational information and formed clusters regardless of the chosen rotations of the images; after the clusters were formed, the images were rotated to match the user data. The matrix was then thresholded, eliminating links between images that had weights less than three, resulting in only 120 "active" images, i.e., images with at least one adjacent edge whose weight is not zero. The thresholded weighted graph adjacency matrix was then analyzed using the spectral clustering algorithm [30]. In this case, the spectral clustering algorithm resulted in 11 non-overlapping clusters. The images in the clusters have no edges connecting them to images that are outside of their own cluster. The 11 extracted clusters are shown in Figs. 10 through 19.

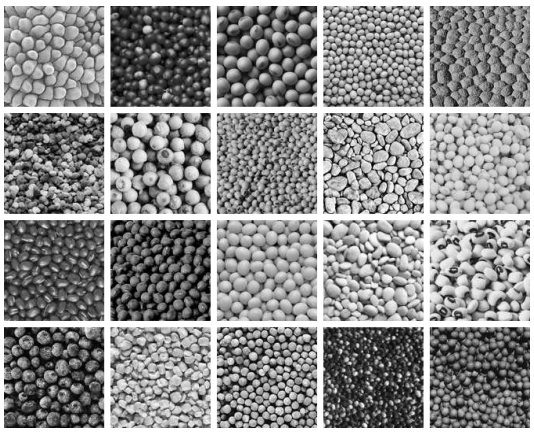

Fig. 10. Grayscale cluster 1

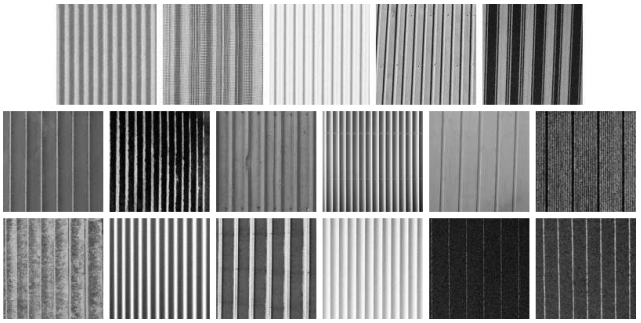

Fig. 11. Grayscale cluster 2

Based on these clusters, we obtained 758 pairs of similar textures (they belong to the same cluster) and 6,382 pairs of dissimilar textures (they belong to different clusters), which form the ground truth for the metric testing experiments we discuss next. For the remaining images that do not belong to any cluster we cannot reliably draw any conclusions.

In summary, out of a total of 505 images, 120 ended up in clusters. Thus, out of 127,260 possible pairs, only 7,140 were rated (about 6%). Yet, we have meaningful results; that is, we obtained a good mix of similar and dissimilar pairs. Achieving such results with ratings of randomly selected texture pairs would require a very large number of trials. Moreover, running ViSiProG is easy and fast and can be fun to run as a game.

7.C. Statistical Analysis of Metric Performance

Given the labeled set of texture pairs, we used them as ground truth for evaluating the ability of objective metrics to distinguish perceptually similar and dissimilar textures. As in the retrieval of identical textures [1], the metric evaluation was based on two types of statistical tests, information retrieval statistics and ROC curves. The former measure the ability of a metric to distinguish similar and dissimilar textures in a relative sense, that is, without reference to an absolute threshold, whereas the latter is based on an absolute threshold [1]. In our experimental results, we compared the fol-

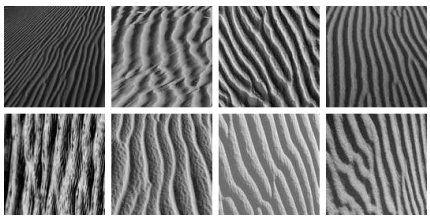

Fig. 12. Grayscale cluster 3

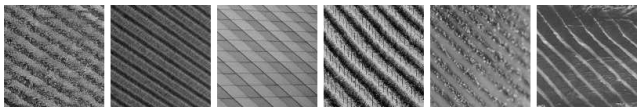

Fig. 13. Grayscale cluster 4

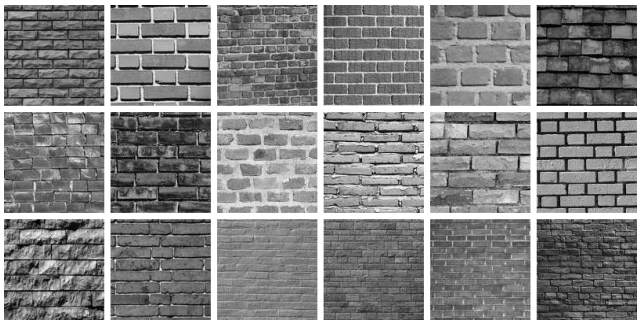

Fig. 14. Grayscale cluster 5

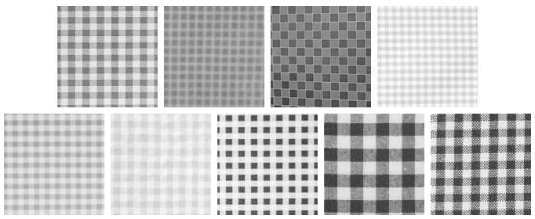

Fig. 15. Grayscale cluster 6

Fig. 16. Grayscale cluster 7

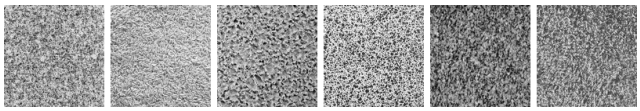

Fig. 17. Grayscale cluster 8

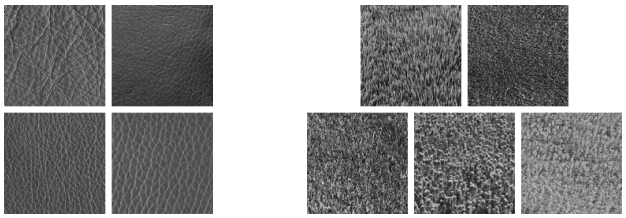

Fig. 18. Grayscale clusters 9 (left) and 10 (right)

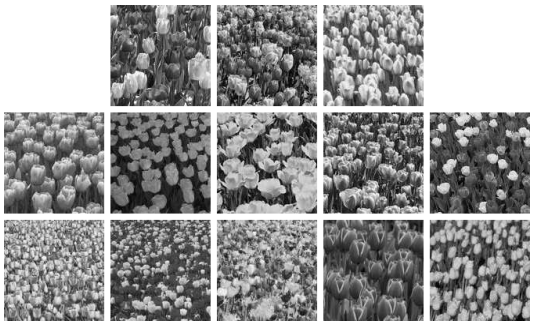

Fig. 19. Grayscale cluster 11

lowing similarity metrics: PSNR; SSIM, with a 7×7 window; CW-SSIM, with a 7×7 window and a global window; STSIM-2, with a 7×7 window and a global window; STSIM-M; Do and Vetterli [15]; and Ojala *et al.* [16].

7.C.1. Information Retrieval Statistics

The metric evaluation experiment can be treated as a retrieval task. We treated as queries only the 120 images that belong to a cluster, as images that do not belong to any cluster do not have any similar images to retrieve. For each of the query images, the remaining 119 images were ordered according to decreasing similarity to the query image.

We used the following performance metrics. Precision at one (P@1) is the number of times the first retrieved image is similar. The mean reciprocal rank (MMR) is the average value of the inverse rank of the first similar retrieved image [34], and provides an estimate of how far down the list the first similar image is. Finally, the mean average precision (MAP) [35] takes into account all the similar images for each query (average precision of all queries of length n where the nth image is similar, averaged over all images).

The results are presented in the first three columns of Table 1. We also calculated the statistical significance of the differences in the performance of the various metrics with significance level set to $\alpha = 0.05$ using the appropriate tests (Cochrane's Q test [36] for P@1, and Friedman test [37, 38] followed by the Tukey-Kramer honestly significant difference test [39] for MRR and MAP). Among the three top performing metrics (STSIM-2, CW-SSIM, and STSIM-2 global), there are no statistically significant differences in performance. However, there are sig-

Algorithm	P@1	MRR	MAP	ROC-AUC
PSNR	0.14	0.23	0.17	0.50
SSIM	0.41	0.49	0.24	0.52
CW-SSIM	0.84	**0.90**	0.64	0.87
CW-SSIM global	0.72	0.82	0.54	0.86
STSIM-2	**0.86**	**0.90**	**0.69**	**0.88**
STSIM-2 global	0.84	0.88	0.65	0.85
STSIM-M	0.84	0.89	0.62	0.80
Do and Vetterli	0.79	0.85	0.56	0.79
Ojala *et al.*	0.57	0.68	0.39	0.54

Table 1. Information retrieval statistics and area under ROC curves for clustering experiment

nificant differences between the bottom two (PSNR and SSIM) and the other metrics. The statistical significance of the remaining comparisons is mixed.

Note that, in contrast to the experiments reported in [1], the local metrics seem to outperform the global ones. This means that the local metrics provide more accurate estimates of texture similarity in general, while in the special case of the retrieval of identical textures, the robustness of the global metrics becomes more important. Thus, different metrics must be used in different applications.

7.C.2. Receiver Operating Characteristic curves

An alternative approach is to treat the evaluation experiment as a binary classification problem. Given two texture images, the metric value is the test variable that will determine whether they are similar (null hypothesis) or not (alternate hypothesis). For each hypothesis, we estimated the probability density functions as the histograms of metric values for every pair of textures in the database for which both textures belong to a cluster. If they belong to the same cluster, then they are similar; if they belong to different clusters, they are dissimilar. Figure 20 (top) shows distributions that correspond to the STSIM-2 metric. Note that the distribution for similar textures is peaky, an indication that the metric provides comparable values for similar textures irrespective of content. On the other hand, the distribution for dissimilar textures is broader, due to the variety of textures in the database. Figure 20 (bottom) shows distributions that correspond to PSNR. Note the almost complete overlap of the two distributions.

The receiver operating characteristic (ROC) is then a plot of the true positive versus the false positive rate for different values of the threshold. The area under the ROC curve is a good indicator of overall performance. ROC curves are stronger indicators of metric performance than the retrieval statistics in the sense that they are based on an absolute threshold for metric values above which textures can be considered to be similar.

The ROC curves are plotted in Fig. 21. The areas under the ROC curves are summarized in the last column of

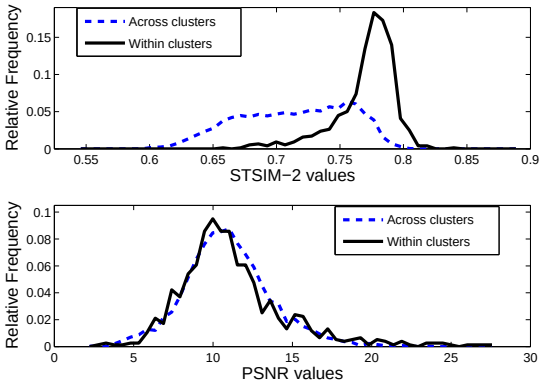

Fig. 20. Histograms for different metrics

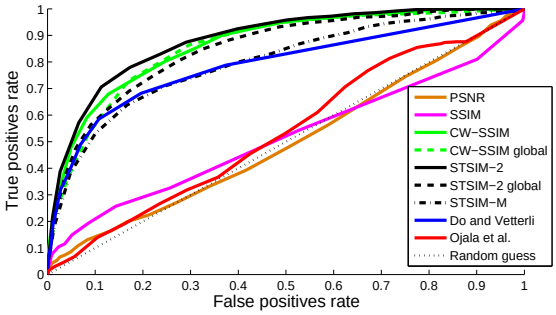

Fig. 21. ROC curves for different metrics

Table 1; the closer the area is to 1.0, the better the performance. By utilizing the method of DeLong *et al.* [40] for comparing the areas under ROC curves, and setting $\alpha = 0.05$, the performance difference between STSIM-2 and all the other metrics is statistically significant. Again, the local metrics outperform the global ones, as was the case with the information retrieval statistics.

8. Conclusions

We proposed a new way of looking at the evaluation of objective texture similarity metrics. Based on the capabilities of human perception and the requirements of different applications, we identified three operating domains for texture similarity metrics: (1) the ability to retrieve identical textures; (2) the top of the similarity scale, where a monotonic relationship is desired; and (3) the ability to distinguish between similar and dissimilar textures. Each of these domains imposes different performance goals for similarity metrics and requires different metric testing procedures. Identifying these domains is essential for obtaining meaningful results, and at the same time results in sizable reductions in the amount of subjective testing that is required for metric evaluation.

We then focused on the third domain, for which we presented ViSiProG, a new procedure for conducting subjective experiments to organize a texture database into clusters of visually similar images. These clusters can be used to obtain ground truth for testing a texture metric in the similar-dissimilar domain. ViSiProG substantially reduces the length of the subjective tests for

obtaining similar and dissimilar texture pairs compared to traditional approaches. In ViSiProG, the subjects form similarity groups one at a time, in a step-by-step fashion, progressively refining the similarity group by selecting images from sequentially presented small subsets of the database. A key element of ViSiProG is that the grouping is based on visual blending.

Experimental results demonstrate that ViSiProG collects subjective data in an efficient and effective manner, using a relatively large database of textures to obtain a large number of similar/dissimilar pairs. Using these pairs as ground truth, we then evaluated the performance of a number of texture similarity metrics based on standard statistical tests. Our results demonstrate that recently developed structural texture similarity metrics are effective in discriminating between similar and dissimilar textures, outperforming other metrics.

Acknowledgment

The authors would like to thank all subjects for their participation in the experiments. This work was supported in part by the U.S. Department of Energy National Nuclear Security Administration (NNSA) under Grant No. DE-NA0000431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily relfect the views of NNSA.

References

[1] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," IEEE Trans. Image Process. **22**, 2545–2558 (2013).

[2] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statictics of complex wavelet coefficients," Int. J. Computer Vision **40**, 49–71 (2000).

[3] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," Signal Processing **70**, 177–200 (1998).

[4] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in "Handbook of Image and Video Processing," , A. C. Bovik, ed. (Academic Press, 2005), pp. 939–959, 2nd ed.

[5] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Perceptual similarity metrics for retrieval of natural textures," in "Proc. IEEE Wksp. Multimedia Signal Proc.", (Rio de Janeiro, Brazil, 2009).

[6] J. Zujovic, "Perceptual texture similarity metrics," Ph.D. thesis, Northwestern Univ., Evanston, IL (2011).

[7] J. Zujovic, T. N. Pappas, D. L. Neuhoff, R. van Egmond, and H. de Ridder, "Subjective and objective texture similarity for image compression," in "Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)," (Kyoto, Japan, 2012), pp. 1369–1372.

[8] B. J. Balas, "Attentive texture similarity as a categorization task: Comparing texture synthesis models," Pattern Recognition **41**, 972–982 (2008).

[9] Y. Chuang and L.-L. Chen, "How to rate 100 visual stimuli efficiently," Int. Journal of Design **2**, 31–43 (2008).

[10] M. R. Greene and A. Oliva, "Recognition of natural scenes from global properties: Seeing the forest without representing the trees," Cognitive Psychology **58**, 137–176 (2009).

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," Int. J. Computer Vision **42**, 145–175 (2001).

[12] I. K. Teeselink, F. Blommaert, and H. de Ridder, "Image categorization," Journal of Imaging Science and Technology **44**, 552–559 (2000).

[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 Process. **13**, 600–612 (2004).

[14] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in "IEEE Int. Conf. Acoustics, Speech, Signal Processing," , vol. II (Philadelphia, PA, 2005), vol. II, pp. 573–576.

[15] M. N. Do and M. Vetterli, "Texture similarity measurement using Kullback-Leibler distance on wavelet subbands," in "Proc. Int. Conf. Image Proc.", , vol. 3 (Vancouver, BC, Canada, 2000), vol. 3, pp. 730–733.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell. **24**, 971–987 (2002).

[17] J. Zujovic, T. N. Pappas, D. L. Neuhoff, R. van Egmond, and H. de Ridder, "A new subjective procedure for evaluation and development of texture similarity metrics," in "Proc. IEEE 10th IVMSP Wksp.: Perception and Visual Signal Analysis," (Ithaca, New York, 2011), pp. 123–128.

[18] T. N. Pappas, D. L. Neuhoff, H. de Ridder, and J. Zujovic, "Image analysis: Focus on texture similarity," Proc. IEEE **101**, 2044–2057 (2013).

[19] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural similarity metrics for texture analysis and retrieval," in "Proc. Int. Conf. Image Processing," (Cairo, Egypt, 2009), pp. 2225–2228.

[20] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in "Proc. Int. Conf. Image Processing (ICIP)," (San Diego, CA, 2008), pp. 1196–1199.

[21] A. Guerin-Dugue, S. Ayache, and C. Berrut, "Image retrieval: A first step for a human centered approach," Proc. 2003 Joint Conf. of the Fourth Int. Conf. on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conf. Multimedia. **1**, 21–25 (2003).

[22] Y. Lu, Q. Zhao, J. Kong, C. Tang, and Y. Li, "A two-stage region-based image retrieval approach using combined color and texture features," in "AI 2006: Advances in Artificial Intelligence," , vol. 4304/2006 of *Lecture Notes in Computer Science* (Springer Berlin/Heidelberg, 2006), pp. 1010–1014.

[23] I. Markov and N. Vassilieva, "Image retrieval: Color and texture combining based on query-image," in "Image and Signal Processing," , vol. 5099 of *Lecture Notes in Computer Science* (Springer, 2008), vol. 5099 of *Lecture Notes in Computer Science*, pp. 430–438.

[24] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," Psychological Bulletin **86**, 420–428 (1979).

[25] N. A. Macmillan and C. D. Creelman, *Detection Theory: A User's Guide* (Lawrence Elbaum Associates, 2005), 2nd ed.

[26] H. Gulliksen and L. R. Tucker, "A general procedure for obtaining paired comparisons from multiple rank orders," Psychometrika **26**, 173–183 (1961).

[27] B. E. Rogowitz, T. Frese, J. R. Smith, C. A. Bouman, and E. Kalin, "Perceptual image similarity experiments," in "Human Vision and Electronic Imaging III," , vol. Proc. SPIE, Vol. 3299, B. E. Rogowitz and T. N. Pappas, eds. (San Jose, CA, 1998), vol. Proc. SPIE, Vol. 3299, pp. 576–590.

[28] W. Richards and J. J. Koenderink, "Trajectory mapping ('tm'): A new non-metric scaling technique," Tech. rep., DTIC Document (1994).

[29] R. Gurnsey and D. J. Fleet, "Texture space," Vision Research **41**, 745–757 (2001).

[30] U. von Luxburg, "A tutorial on spectral clustering," Statistics and Computing **17**, 395–416 (2007).

[31] W. S. Torgerson, *Theory and methods of scaling* (Wiley, New York, NY, 1958).

[32] J. B. Kruskal and M. Wish, *Multidimensional scaling* (Sage Publications, Beverly Hills, CA, 1977).

[33] Corbis, "Corbis stock photography," (2013). Accessed July 2013.

[34] E. M. Voorhees, "The TREC-8 question answering track report," in "In Proc. 8th Text Retrieval Conf. (TREC-8)," , vol. 8, E. M. Voorhees and D. K. Harman, eds. (1999), vol. 8, pp. 77–82.

[35] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," Information Processing & Management **36**, 697–716 (2000).

[36] W. G. Cochran, "The combination of estimates from different experiments," Biometrics **10**, 101–129 (1954).

[37] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," J. American Statistical Association **32**, 675–701 (1937).

[38] J. D. Gibbons, *Nonparametric statistics: An Introduction*, Quantitative Applications in Social Sciences 90 (Sage Publications, London, 1993).

[39] J. W. Tukey, "Quick and dirty methods in statistics," in "Part II: Simple Analyses for Standard Designs. Quality Control Conference Papers," (1951), pp. 189–197.

[40] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics pp. 837–845 (1988).