2011

A Perceptually Based Comparison of Image Similarity Metrics

Richard Russell
Gettysburg College

Pawan Sinha
Massachusetts Institute of Technology

Sinha, P., and Russell, R. (2011). A perceptually based comparison of image similarity metrics. Perception, 40(11), 1269-1281. http://dx.doi.org/10.1068/p7063

A Perceptually Based Comparison of Image Similarity Metrics

Abstract

The assessment of how well one image matches another forms a critical component both of models of human visual processing and of many image analysis systems. Two of the most commonly used norms for quantifying image similarity are L1 and L2, which are specific instances of the Minkowski metric. However, there is often not a principled reason for selecting one norm over the other. One way to address this problem is by examining whether one metric, better than the other, captures the perceptual notion of image similarity. This can be used to derive inferences regarding similarity criteria the human visual system uses, as well as to evaluate and design metrics for use in image-analysis applications. With this goal, we examined perceptual preferences for images retrieved on the basis of the L1 versus the L2 norm. These images were either small fragments without recognizable content, or larger patterns with recognizable content created by vector quantization. In both conditions the participants showed a small but consistent preference for images matched with the L1 metric. These results suggest that, in the domain of natural images of the kind we have used, the L1 metric may better capture human notions of image similarity.

Keywords

image similarity metrics, human visual processing

Disciplines

Cognition and Perception | Psychology

Perception, 2011, volume 40, pages 1269 – 1281

doi:10.1068/p7063

A perceptually based comparison of image similarity metrics

Pawan Sinha
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 46-4077,
77 Massachusetts Avenue, Cambridge, MA 02139, USA; e-mail: psinha@mit.edu

Richard Russell
Department of Psychology, Gettysburg College, Gettysburg, PA 17325, USA

Received 2 July 2011, in revised form 19 November 2011

Abstract. The assessment of how well one image matches another forms a critical component both of models of human visual processing and of many image analysis systems. Two of the most commonly used norms for quantifying image similarity are L1 and L2, which are specific instances of the Minkowski metric. However, there is often not a principled reason for selecting one norm over the other. One way to address this problem is by examining whether one metric, better than the other, captures the perceptual notion of image similarity. This can be used to derive inferences regarding similarity criteria the human visual system uses, as well as to evaluate and design metrics for use in image-analysis applications. With this goal, we examined perceptual preferences for images retrieved on the basis of the L1 versus the L2 norm. These images were either small fragments without recognizable content, or larger patterns with recognizable content created by vector quantization. In both conditions the participants showed a small but consistent preference for images matched with the L1 metric. These results suggest that, in the domain of natural images of the kind we have used, the L1 metric may better capture human notions of image similarity.

1 Introduction

The operation of comparing images is an integral component of many visual routines. Key visual tasks such as stereo-depth estimation and motion flow extraction depend on being able to establish correspondence between different image regions across space or time (Dhond and Aggarwal 1989; Grimson 1982; Hildreth 1987; Marr and Poggio 1979; Mayhew and Frisby 1981). Correspondence, in turn, depends critically on comparing image regions and evaluating their mutual similarity. Image comparisons play an even more obvious role for tasks like object recognition. In order to be able to build accurate models of these visual abilities, we need to use formally specifiable image similarity metrics that mimick their human counterparts.

The need for choosing appropriate image similarity metrics can be motivated from a more pragmatic perspective as well. The rapidly growing preponderance of digital images in diverse aspects of everyday life necessitates automatic methods for their manipulation, storage, and use. Central to many operations on digital images are image similarity metrics ('distance functions' or, more generally in information theory, 'distortion measures') that quantify how well one image matches another. Three broad classes of applications that rely on appropriately chosen image similarity metrics are image search, image compression, and image quality assessment.

Thus, similarity metrics are important both for understanding human vision and for improving image processing in applied settings. Indeed, the two goals are complementary. Computational metrics can be improved by approximating human similarity judgments, while human similarity judgments can be better understood by comparing them to formally characterized computational similarity metrics.

It is worth pointing out that a similar motivation underlies a growing body of work in the computer graphics domain. Analogous to the problem of choosing a similarity metric on perceptual grounds is the task of selecting (or devising) rendering schemes that are computationally efficient and maximize perceptual realism (Bartz et al 2008; Ramanarayanan et al 2007; Stich et al 2011). The perceptual criterion has proven to be a very fruitful one for narrowing down the set of algorithmic possibilities since it perfectly captures the eventual usage scenario of computer graphics renderings, and also enables significant complexity reduction by exploiting limitations of the human perceptual apparatus (for an example, see Ostrovsky et al 2005).

In both human and computational vision research, a very widely used class of image similarity metrics involves performing some operation on the differences between corresponding pixels in two images, then summing over these modified differences. These are referred to collectively as the $\mathrm{L_P}$ family of similarity metrics, or the Minkowski metric. The general form of the p norm is:

$$\|\mathrm{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

The most commonly used members of this family are the L1 and L2 norms. Formally, the L1 and L2 norms are defined as follows: For two n-component vectors $\bar{a}$ and $\bar{b}$, L1 distance between the vectors is:

$$\sum_{i=0}^{n-1} |a_i - b_i| .$$

The L2 distance between $\bar{a}$ and $\bar{b}$ is:

$$\sum_{i=0}^{n-1} \sqrt{(a_i - b_i)^2} .$$

The L1 metric is also called the Manhattan distance or the Sum of Absolute Error (SAE), and the L2 metric is also called the Euclidean distance or Sum of Square Error (SSE). Examples of studies that have employed these norms, or slight variants in models of visual processing, include: (Balas and Sinha 2006; Bülthoff et al 1989; van Doorn and Koenderink 1983; Förstner 1993; Jenkin et al 1991; Liu et al 1995; Ludwig et al 1994). Examples of their use in more applied settings include: in image compression (Baker and Gray 1982; Delac et al 2008; Gersho and Ramamurthi 1982; Goldberg et al 1986; Mathews 1992; Mathews and Khorchidian 1989); in image retrieval (Kwasnicka et al 2011; Rubner et al 1997; Tao and Dickinson 1996); in image quality assessment (Ahumada 1993). Though many implementations utilize metrics that are more complicated, incorporating specific task-dependent features (Eckert and Bradley 1998; Frese et al 1997; Jayant et al 1993; Watson 1993), they are often modifications made to simpler metrics such as L1 and L2.

Currently both these metrics are used commonly and interchangeably. The L1 metric has the advantage of being slightly less computationally expensive, involving no product terms. This can be relevant in computationally demanding applications such as large database search, where even small improvements in computational efficiency can lead to significant time savings. The L2 metric has the advantage of being continuously differentiable. Yet there is no reason to believe that either of these reasons is of any concern to the human visual system. Since the end result of many image-analysis operations is intended to be viewed by humans, it is the human visual system that in many applications is the ultimate arbiter of the similarity of images. Thus, the most relevant criterion for deciding between the two metrics may be perceptual rather than computational. Surprisingly, very few studies have investigated perceptually based differences between the L1 and L2 norms, and none, to the best of our knowledge, have done so systematically. Mathews and Hahn (1997) have commented on the perceptual interchangeability of the metrics. DeVore and colleagues (1992) have advocated the perceptual superiority of the L1 metric in the domain of wavelet transform coding, based anecdotally on their own subjective judgment of a handful of images. To date, however, a rigorous, perceptually based, comparison of the two metrics has not been performed. The purpose of the present study is to perform such a comparison, in order to determine whether one of the two similarity metrics is closer to human notions of what it means for two images to look similar. The results could provide a well-motivated way to choose between the L1 and L2 metrics for image-analysis tasks. Furthermore, determining which computational metric better captures human notions of similarity can be used to derive inferences regarding similarity criteria the visual system uses. For instance, an important way in which the L1 and L2 norms differ is in terms of how much they penalize outliers. By comparing these two metrics, we can indirectly assess the significance of outliers for human observers. With these motivations in mind, we experimentally investigated whether humans prefer the image matches chosen by the L1 or the L2 metric.

Admittedly, the capabilities of simple metrics like L1 and L2 in capturing high-level image meaning are rather limited. For tasks like semantic content-based image classification, or object recognition under significant transformations, these metrics need to be augmented with sophisticated representations and, often, statistical classifiers. The goal of our study, however, is to examine the effectiveness of the similarity metrics for comparing image structure, without the confounding influence of high-level semantics. Though modest, this goal matches the task requirements involved in several important image-analysis settings. These include local image matching (for instance, to establish correspondence between points in a stereo pair or frames of a motion sequence) and situations where images are not semantically meaningful (say synthetic aperture radar images). It is also a necessary first step toward developing more sophisticated metrics that attempt to handle image semantics, as these metrics typically rely on one or another simple similarity metric for providing a basic set of distances that can then serve as inputs for further analysis/classification.

Pursuant to the discussion above, an important issue that arises in the design of experiments for comparing low-level image similarity metrics is how to deal with the high-level semantic content in images. Such content (for instance, people, flowers, objects, etc) may be more salient to viewers than the abstract patterns of light and dark that constitute the image structure on which similarity metrics operate. This may cause judgments of perceptual similarity to be influenced by high-level semantic considerations. For example, a flower vase and a garden in bloom may be declared to be similar on the basis of high-level information, even though they are quite different at the level of image structure. Figure 1 shows an example of how semantic considerations can lead to classifications that have little to do with similarity at the level of the image structure. If metrics such as the L1 and L2 norm are used to determine the similarity of images with semantic information, they often produce results that may seem poor to humans, for whom the semantic information is apparent and critical. A study by Rogowitz and colleagues (1998) also highlighted the role that semantic information plays in human judgments of image similarity.

Because semantic content is not the focus of our study, we needed to find a way to control for it. Toward this end, we conducted our comparisons of the two metrics in two different experiments. One experiment preserved semantics and the other dispensed with it. In the first experiment, participants viewed images with recognizable semantic content that were composed of many small fragments. Each fragment individually was too small to have any high-level meaning. The computational similarity metrics were used to select these image fragments. In the second experiment, participants viewed these single fragments in isolation, such that there was nothing recognizable in the images. The trials of both experiments involved asking participants to decide which of two images better matched a target image. In these trials one of

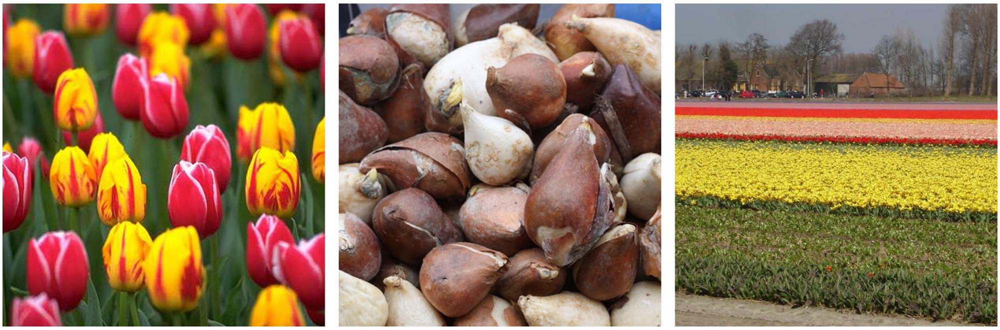

Figure 1. [In color online, see http://dx.doi.org/10.1068/p7063] Semantic criteria can lead to image groupings that are unrelated to similarity at the level of image structure. All of the images shown above belong to the semantic category of tulips ('tulip flower', 'tulip bulbs', 'tulip field'). While the ability to classify images based on semantic criteria is of use in some settings, for the purposes of this paper, we seek to focus on image level similarity that can be captured by metrics such as the L1 and L2 norms.

the two images was chosen/composited from a library of images using the L1 metric while the other image was based on the L2 metric. Thus, on each trial the participants had to decide whether they agreed more with the similarity judgment of the L1 or L2 metric.

2 General methods

Both experiments utilized the same two-alternative forced-choice design, in which participants were instructed to choose which of two (probe) images looked most like a reference (target) image. Participants indicated their preference by pressing a keyboard button, and were not timed, although they were allotted a maximum of 10 s to complete each trial. The same display configuration was used in all trials. This consisted of the three images in the center of the screen with the target image above and the two probe images below. In each trial of the first experiment, one of the two probe images was derived via vector quantization (detailed below) using the L1 metric as the distortion function and the other probe image was derived using the L2 metric. The left – right ordering of the two probe images was counterbalanced across trials and conditions. Participants were not required to determine which of the two probe images was derived using the L1 metric and which using the L2 metric. Each experiment consisted of 384 trials. Participants sat approximately 75 cm from the display monitor in a room with low ambient illumination. Each image subtended 2° of visual angle. Twenty-three individuals participated as participants in the two experiments. One subject, RPR, participated in both experiments and is an author of this report. The remaining participants were paid volunteers, eleven participating in the first experiment and eleven participating in the second. Each experiment thus employed twelve participants.

3 Experiment 1

3.1 *Methods*

The images in the first experiment were generated by vector quantization (VQ) (Nasrabadi and King 1988). Our codebook was a library of randomly selected natural images. Each fragment of the target image was compared with every other fragment of the same size in each image in the library using an image similarity metric, and replaced by the fragment judged to be the most similar by the metric. After this was performed on each fragment in the target image, a new image was created that was composed entirely of fragments from the library. Figure 2a shows an example of image reconstruction by VQ.

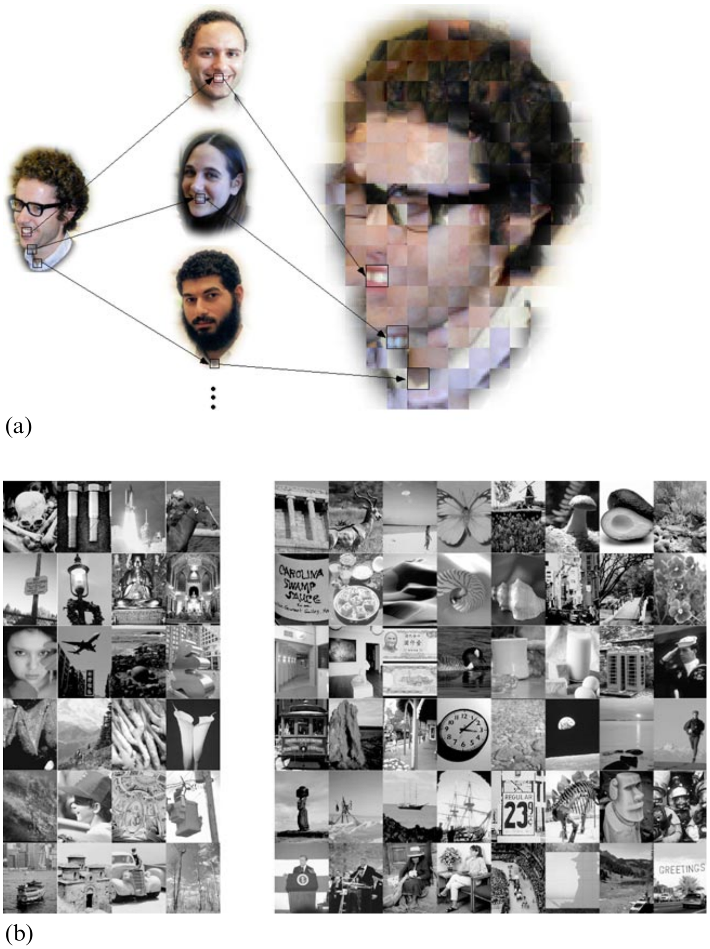

Figure 2. [In color online.] (a) In vector quantization, fragments of the target image are matched to fragments in the library images using an image similarity metric. These fragments from the library images then replace those in the target image, creating a reconstructed image that is composed entirely of fragments from the library images. In this figure, the image to the left is the target image, the middle images are the codebook, or library of images, and image to the right is an enlarged version of the reconstructed image created by placing fragments from the library images together. Exactly which fragments from the library are chosen to replace the target fragments is affected by the choice of image similarity metric. The poor quality of the reconstruction here is due to the very small size of the codebook (5 images). (b) Thumbnails of all the images used in the target and library images. The 24 images on the left are the target images, and the 48 images on the right are the codebook images.

Our decision to use the VQ scheme for generating experimental stimuli was motivated primarily by the need to control for semantic content across images being compared. The fact that this decision makes our results be directly applicable for the design of better VQ systems is a useful side effect.

Two important parameters for a VQ scheme are the number of images in the library and the size of the fragments. Larger libraries and smaller fragment sizes create images that are more similar to the target images. On the other hand, increasing the fragment size and/or decreasing the library size will typically result in reconstructed images that look less like their target images, but have the benefit of yielding greater compression factors.

The stimuli for the experiment were created using either the L1 or the L2 metric to choose replacement fragments from the library. The two metrics often choose different fragments, leading to different overall reconstructions. Thus we can ask which of the two metrics produces better reconstructions from a perceptual point of view.

72 images were selected at random from the IMSI MasterClips image catalog. The images were a mixture of indoor and outdoor natural scenes with a variety of objects and people at different spatial scales. Each was cropped to 150 × 200 pixels and converted to grayscale. 24 of these images were used as target images to be compressed (henceforth referred to as 'reconstructed'), and 48 were used as library (codebook) images. Figure 2b is a montage of thumbnails of all 72 images. Reconstructed images were always created in pairs with the L1 metric in one case and L2 in the other.

To investigate whether the fragment size or library size played a role in determining perceptual preferences, we created reconstructed images with four different fragment sizes (5 × 5 pixels, 10 × 10 pixels, 15 × 15 pixels, and 20 × 20 pixels) and four different library sizes (6, 12, 24, and 48 images). Smaller libraries were subsets of larger libraries (eg all of the images used in the 12-image library were used in the 24-image library). Thus, there were 24 target images × 4 fragment sizes × 4 library sizes = 384 pairs of reconstructed images. Each of the 24 target images appeared 16 times—once for each of the 16 different conditions—with a different pair of reconstructed images each time. Figure 3 shows an example trial from the condition with block size of 20 × 20 pixels and a library size of 48 images.

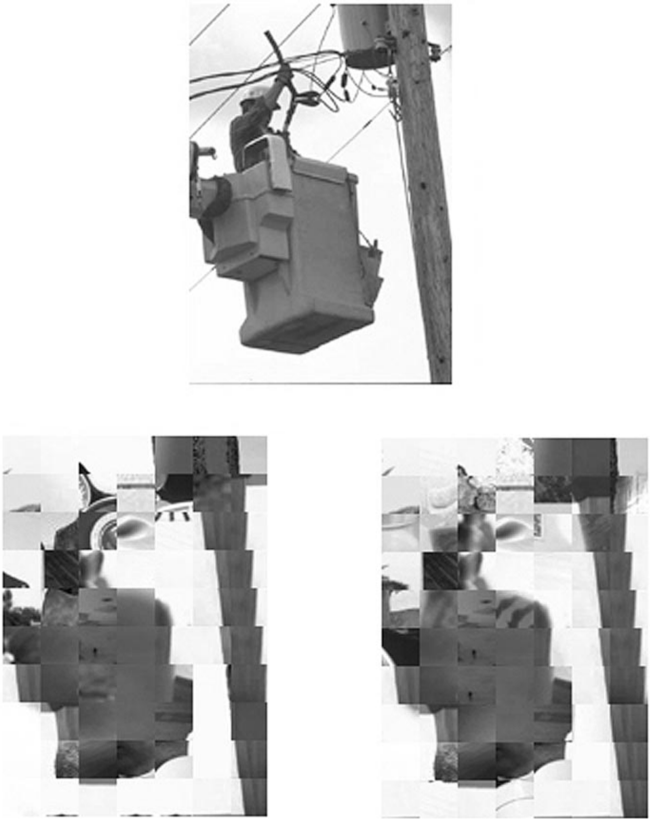

Figure 3. Sample display from experiment 1. The top image is the original. The bottom images are reconstructions, one using the L1 norm as the distortion function and the other using the L2 norm. Participants were instructed to indicate which of the two bottom images looked more like the top image.

3.2 *Results*

Participants displayed a small, but highly consistent preference for reconstructions based on the L1 metric. 54% of all responses (averaged across fragment sizes and library sizes) were made for the reconstructions with the L1 metric. Of the twelve participants, eleven chose the L1 metric on more than 50% of the trials and the remaining subject chose the L1 metric on 49% of the trials. Participants chose images reconstructed with the L1 metric significantly more often than those created with the L2 metric (Student's $t_{11} = 3.98$, $p < 0.01$). The results are shown in figure 4a. There were no significant differences in preferences across the different fragment sizes (single factor ANOVA, $p = 0.69$) or library sizes (single factor ANOVA, $p = 0.61$). Figures 4b and 4c show the results by fragment size and library size.

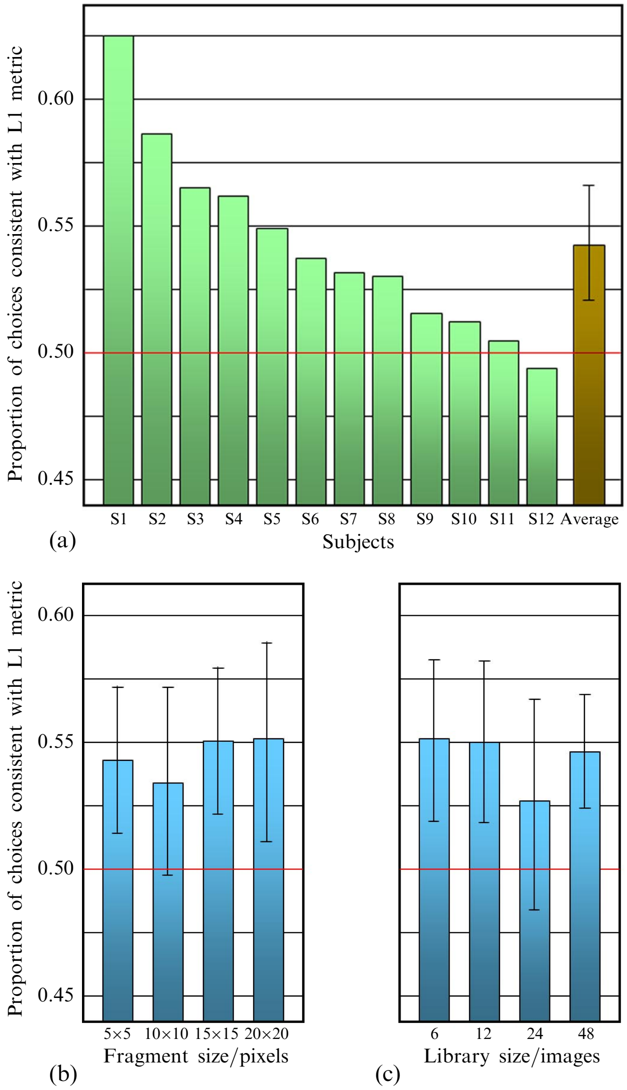

Figure 4. [In color online.] (a) Subject preferences in experiment 1. (b) Preferences across four different fragment sizes. Each bar represents the responses to reconstructions with a given fragment size, averaged across all participants. Preferences for the different metrics did not differ significantly by fragment size. (c) Preferences across four different library sizes. Each bar represents the responses to reconstructions with a given library size, averaged across all participants. Preferences for the different metrics did not differ significantly by library size.

4 Experiment 2

4.1 *Methods*

The second experiment was identical to the first, with the exception of the stimuli. Rather than using the entire original images and reconstructions, individual image fragments and their best L1 and L2 matches were displayed. The fragments used were from the 20 × 20 pixel fragment with 48 images library condition. Of the 1680 possible fragments, 384 pairs of original fragments with their L1 and L2 image matches were used. These pairs of original and match fragments were chosen from across the spectrum of pairs, from fairly similar to quite different from one another. The similarity of the pairs was determined by calculating the L1 distance between the two matches (not between the matches and the original fragments). Because the selection of pairs was made such that there was a range of similarity of pairs, and because their pair distances were not subsequently analyzed, the choice of the L1 norm for this determination in no way biases the subsequent results toward either of the two norms. Because these fragments were quite small, they were scaled to 60 × 60 pixels by bicubic interpolation. Figure 5a shows an example trial.

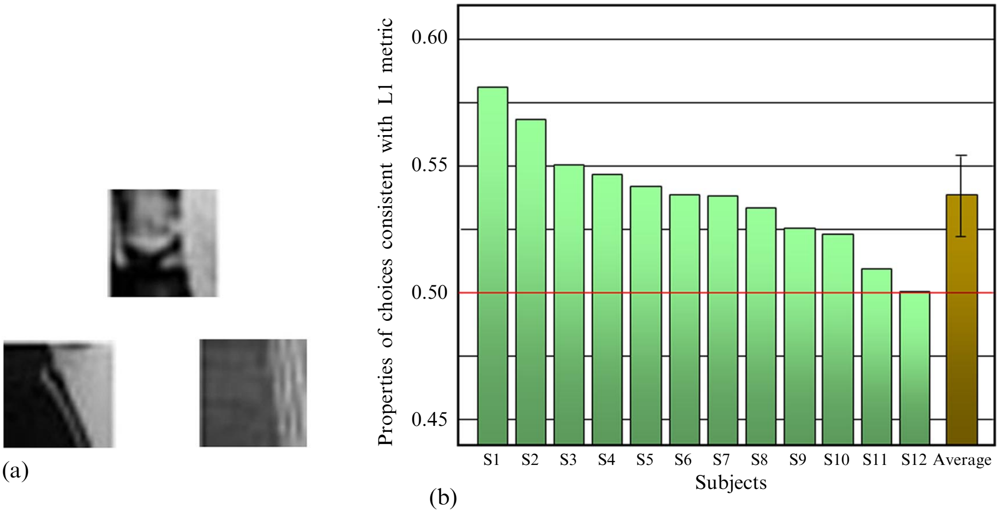

Figure 5. [In color online.] (a) Each trial in experiment 2 consisted of a display similar to the figure above. The top image is the original. The bottom images are the best matching fragments retrieved with the L1 or the L2 metrics. Participants were instructed to choose which of the two bottom images better matched the top image. The left – right ordering of the L1 and L2 reconstructions was counterbalanced across trials and conditions, and was not known to the participants. (b) Subject preferences in experiment 2. All but one of the participants showed a bias towards the L1 metric. One subject's preferences were exactly evenly split across the two metrics.

4.2 *Results*

Interestingly, the results from experiment 2 were qualitatively very similar to those from experiment 1. 54% of all responses were made for the match found with the L1 metric. Of the twelve participants, eleven chose the L1 metric on more than 50% of the trials and one subject chose the L1 metric on exactly 50% of the trials. The participants' choices of the L1 metric were significantly greater than their choices for the L2 metric (Student's $t_{11} = 6.48$, $p < 0.001$). The results are shown in figure 5b.

5 Discussion

We examined human perceptual preferences for image matches produced by two widely used computational metrics. Our results showed a small but highly consistent preference for the L1-based matches. This finding is consistent with and extends the anecdotal assertion of DeVore et al (1992) that the L1 norm is superior in the domain of wavelet transform coding, but is in contrast to the suggestion that the two norms are interchangeable (Mathews and Hahn 1997).

What might underlie the preference for L1 matches? The only divergence between the two norms is that the L2 norm squares the difference between corresponding pixels, while the L1 norm does not. An important result of squaring for the L2 norm is that large pixel differences will have a disproportionately large effect on the overall distance between two image fragments. Thus the L2 metric will be more sensitive to some pixels than to others. This sensitivity to outliers should result in the L2 metric choosing matches that have few pixels that are very different from the corresponding target pixels. However, the L1 matches should be better able to tolerate such pixel outliers.

To investigate whether this was true for our data, we computed difference images for each pair of target image and library match. For each target image, two difference images were computed—one for the L1 match and one for the L2 match. The value of a given pixel in a difference image is the difference between the pixel value in the match image and the pixel value in the target image. In the difference images, a large pixel value (brighter) corresponds to a large difference between the match pixel and the target pixel, and a small pixel value (darker) corresponds to a small difference between the match pixel and the target pixel. Figure 6 shows histograms of all the L1 difference images and all the L2 difference images, as well as the difference between the two histograms. The L1 difference images had more small values and slightly more very large values. The L2 difference image contained more intermediate values. The means of both distributions were very similar—28.0 for the L1, and 28.2 for the L2—indicating that the few very large difference values produced with the L1 metric were balanced out by the larger number of small values. However, the medians of the two distributions diverged—21.6 for the L1 and 24.6 for the L2.

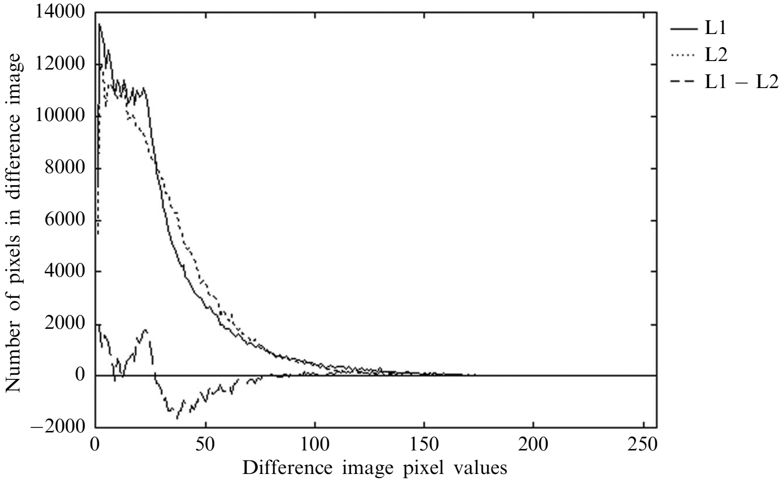

Figure 6. Histograms of difference images. The *X*-axis indicates the difference image pixel values, which are the differences between the pixel values of the original and match fragments. These values run from 0 (for corresponding pixels that were exactly the same) to 255 (for corresponding pixel pairs in which one was completely white and the other completely black). The solid line and dotted line denote the distribution of values created by taking the difference between matches and original images based on L1 and L2 metrics, respectively. The dashed line is the distribution of values created by subtracting the dotted line from the solid line.

Thus in the two sets of differences, the means are the same, while the median of one set is lower than that of the other. It was the set with the lower median difference that the participants considered to be most similar. These data suggest that the participants' notion of image similarity was more similar to that of the L1 norm (placing equal weight on each pixel) than that of the L2 norm (placing greater weight on those pixels with greater differences).

The heightened sensitivity of the L2 norm to outliers makes it more susceptible to signal noise than the L1 norm. A computational experiment demonstrates this convincingly. The task we considered was that of face recognition. Our database comprised 10 grayscale images each of 40 different individuals (this database was compiled by the AT&T Laboratories in Cambridge, UK; more information is available at http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html). Each image was 112 × 92 pixels

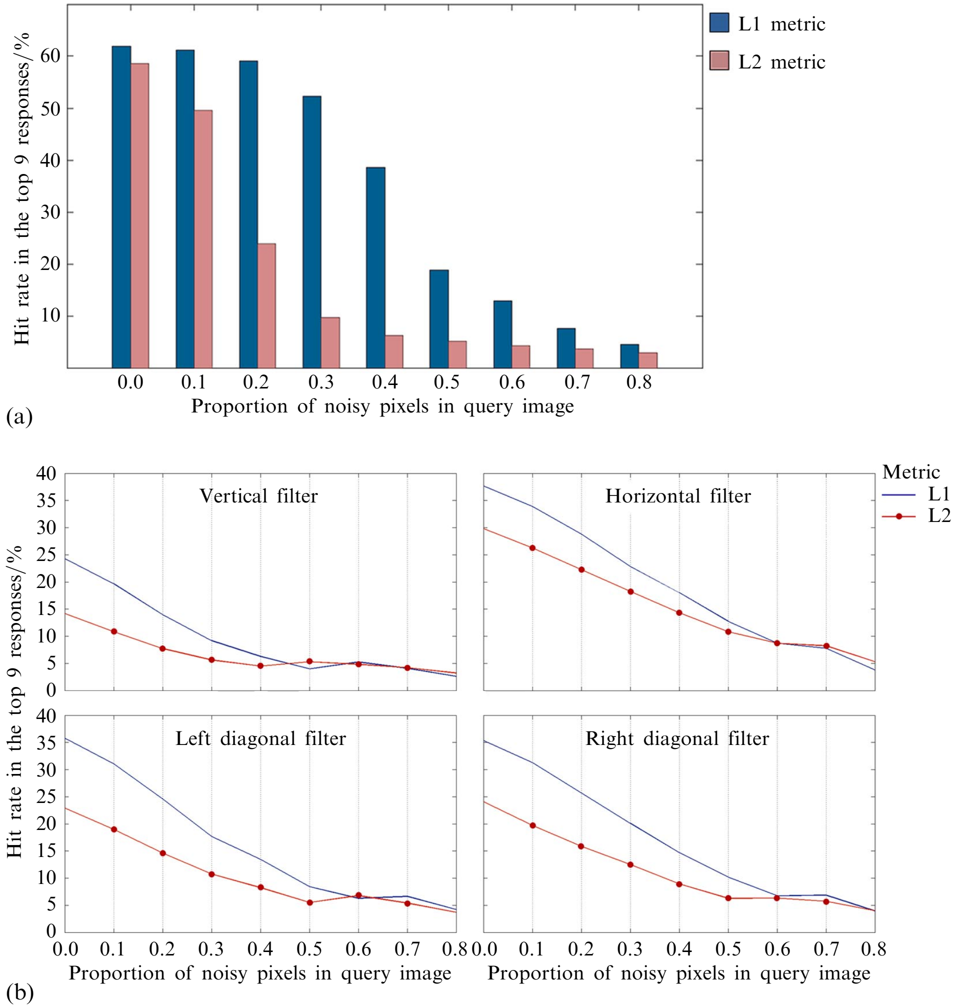

Figure 7. [In color online.] (a) Results of a computational experiment assessing the performance of L1-based and L2-based matching on a simple face-recognition task. Raw pixel values are used to compute the similarity metrics. Although the two metrics perform comparably in the no-noise condition (leftmost bar pair), with the introduction of noise in the query images, the performance of the L2-based retrieval system falls much faster than that of the L1-based system. (b) Same analysis as in (a), except that instead of raw pixel values, the outputs of four different filters are used to compute the similarity metrics. The filters employed here are 3 × 3 Sobel masks (each image is 112 × 92 pixels) at four different orientations, as indicated.

in size. The 10 images of a given person varied slightly in pose, expression, and lighting. We tested the efficacy of two image matching systems, one based on the L1 metric and the other on L2 for retrieving the images of a person given one query image. The similarity criterion used for rank-ordering the database conditioned on a query image was simply the L1 or L2 distance computed across each entire image in the set. For a perfect system, the top 9 returned images would all be of the person depicted in the query image. Figure 7a shows the results. When no noise is present in the images, the two metrics perform comparably, with L1 having a slight edge over L2, consistent with our behavioral results reported above. However, with the introduction of noise, the results from the two metrics diverge markedly. Even small amounts of speckle noise, wherein the values of some randomly selected pixels are clamped at ceiling or floor, greatly reduces the hit-rate of the L2-based retrieval system. The L1-based system, however, proves robust across a much greater range of noise.

It is worth noting that the perceptual biases towards the L1 norm, though statistically significant, are modest in magnitude. This may be attributable to multiple factors. First, VQ reconstructions that we used in the first experiment suffered from block artifacts (see figure 3). These artifacts, due to their perceptual salience, could have interfered with participants' choices of images. Second, the L1 and L2 norms were computed directly over raw pixel values. An interesting extension of this work would involve the investigation of the effects of physiologically plausible pre-processing on the perceptual bias towards L1 and L2 metrics. For instance, instead of computing similarity metrics on the raw pixel values, how would the results change if the images had first been subjected to a Laplacian operator or a Gabor filter of the kind that are known to exist in the early stages of the visual pathway? As a first step towards this goal, we have extended our analyses to compute relative performance of the two metrics on the afore-described face-recognition task when the images are pre-processed with oriented filters. Specifically, we used Sobel filters at four different orientations following the parametric addition of noise to the original images. The results are shown in figure 7b. Comparing them to figure 7a, it is evident that the absolute magnitude of performance for both metrics registers a decrease. This is because post-filtering image information is spatially sparse and even minor misalignments result in large costs in terms for the matching metric. However, for the purposes of the present discussion, it is worth noticing that the relative performance of the two metrics follows the same trend that we have observed so far, ie the L1-based matcher yields better results than the L2-based matcher. Third, to keep the duration of experimental sessions within reasonable lengths, the number of test images was constrained to be fairly small (24). This does make the study susceptible to inadvertent biases in query image selection. Finally, the participants' bias for one over the other norm might have been greater for a different set of tasks. Indeed, results of the computational experiment reported above bear out this expectation. Given that human responses would likely have been identical to the ground truth on the face-recognition task, the bias for the L1 norm they would have exhibited would have been markedly greater than for L2 for even small amounts of image noise.

The results of these experiments give a principled reason for choosing the L1 metric rather than the L2 metric for use in image analysis. The difference, though small, is highly consistent and suggests that in applications related to the retrieval, manipulation, and compression of natural images, use of the L1 metric should result in better performance than that achieved with the L2 metric. In settings with significant sensor noise, the differences may in fact prove to be quite substantial, as our computational experiment shows. Also, the findings reported here may be of use in designing new similarity metrics and more accurate models of human visual processes underlying such tasks as stereo and motion correspondence that rely critically on image comparisons.

Acknowledgments. This research was supported in part by funds from the Alfred P Sloan Fellowship in Neuroscience, the Jim and Marilyn Simons Foundation and the James S McDonnell Foundation to PS, and from NIH grants R01 EY020517 to PS and MH20007 to the Department of Brain and Cognitive Sciences at MIT.

References

Ahumada A J, 1993 "Computational image quality metrics: a review" paper presented at the Society for Information Display International Symposium Digest of Technical Papers, Playa del Rey, CA

Baker R L, Gray R M, 1982 "Image compression using non-adaptive spatial vector quantization" *Conference Record of 16th Asilomar Conference on Circuits, Systems, Computers* pp 55 – 61

Balas B J, Sinha P, 2006 "Receptive field structures for recognition" *Neural Computation* **18** 497 – 520

Bartz D, Cunningham D, Fischer J, Wallraven C, 2008 "The role of perception for computer graphics", in *Proceedings of the 29th Annual Conference Eurographics (EG 2008)* (Blackwell, Oxford) pp 65 – 86

Bülthoff H H, Little J J, Poggio T A, 1989 "A parallel algorithm for real-time computation of optical flow" *Nature* **337** 549 – 553

Delac K, Grgic S, Grgic M, 2008 "Image compression in face recognition—a literature survey", in *Recent Advances in Face Recognition* Eds K Delac, M Grgic, M S Bartlett (Vienna: I-Tech) pp 236 – 250

DeVore R A, Jawerth B, Lucier B, 1992 "Image compression through wavelet transform coding" *IEEE Transactions on Information Theory* **38** 719 – 746

Dhond U R, Aggarwal J K, 1989 "Structure from stereo—a review" *IEEE Transactions on Systems, Man and Cybernetics* **19** 1489 – 1510

Doorn A J van, Koenderink J J, 1983 "The structure of the human motion detection system" *IEEE Transactions on Systems, Man, and Cybernetics* **13** 916 – 922

Eckert M P, Bradley A P, 1998 "Perceptual quality metrics applied to still image compression" *Signal Processing* **70** 177 – 200

Forster K I, Forster J C, 1990 "The DMASTER display system for mental chronometry" (Tuscon, AZ: University of Arizona)

Förstner W, 1993 "Image matching", in *Computer and Robot Vision* volume 2, Eds R M Haralick, L G Shapiro (Boston, MA: Addison Wesley) chapter 16

Frese T, Bouman C A, Allebach J P, 1997 "A methodology for designing image similarity metrics based on human visual system models" paper presented at the *Proceedings of the SPIE/IS&T Conference on Human Vision and Electronic Imaging II, San Jose CA*

Gersho A, Ramamurthi B, 1982 "Image coding using vector quantization" *IEEE Conference on Acoustics, Speech, and Signal Processing* **1** 428 – 431

Goldberg M, Boucher P R, Shlien S, 1986 "Image compression using adaptive vector quantization" *IEEE Transactions on Communication* **COM-34** 180 – 187

Grimson W E L, 1982 "A computational theory of visual surface interpolation" *Philosophical Transactions of the Royal Society of London, Series B* **298** 395 – 427

Hildreth E C, 1987 "The analysis of visual motion: From computational theory to neuronal mechanisms" *Annual Review of Neuroscience* **10** 477 – 533

Jayant N, Johnston J, Safranek R, 1993 "Signal compression based on models of human perception" *Proceedings of the IEEE* **81** 1385 – 1422

Jenkin M R M, Jepson A D, Tsotos J K, 1991 "Techniques for disparity measurement" *CVGIP: Image Understanding* **53** 14 – 30

Kwasnicka H, Paradowski M, Stanek M, Spytkowski M, Sluzek A, 2011 "Image similarities on the basis of visual content—An attempt to bridge the semantic gap", in *Intelligent Information and Database Systems, Lecture Notes in Computer Science* **6591** 14 – 26

Liu Z, Knill D C, Kersten D, 1995 "Object classification for human and ideal observers" *Vision Research* **35** 549 – 568

Ludwig K O, Neumann H, Neumann B, 1994 "Local stereoscopic depth estimation" *Image and Vision Computing* **12** 16 – 35

Marr D, Poggio T, 1979 "A computational theory of human stereo vision" *Proceedings of the Royal Society of London, Series B* **204** 301 – 328

Mathews V J, 1992 "Multiplication free vector quantization using L1 distortion measure and its variants" *IEEE Transactions on Image Processing* **1** 11 – 17

Mathews V J, Hahn P J, 1997 "Vector quantization using the L-infinity distortion measure" *IEEE Signal Processing Letters* **4** 33 – 35

Mathews V J, Khorchidian M, 1989 "Multiplication-free vector quantization using L_1 distortion measure and its variants", paper presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)

Mayhew J E W, Frisby J P, 1981 "Psychophysical and computational studies towards a theory of human stereopsis" *Artificial Intelligence* **17** 349 – 385

Nasrabadi N M, King R A, 1988 "Image coding using vector quantization: A review" *IEEE Transactions on Communication* **COM-36** 957 – 971

Ostrovsky Y, Cavanagh P, Sinha P, 2005 "Perceiving illumination inconsistencies in scenes" *Perception* **34** 1301 – 1314

Ramanarayanan G, Ferwerda J, Walter B, Bala K, 2007 "Visual equivalence: towards a new standard for image fidelity" *ACM Transactions on Graphics, Proceedings of SIGGRAPH* **26**(3)

Rogowitz B E, Frese T, Smith J R, Bouman C A, Kalin E, 1998 "Perceptual image similarity experiments" paper presented at the Conference on Human Vision and Electronic Imaging, San Jose, CA

Rubner Y, Guibas L J, Tomasi C, 1997 "The earth movers distance, multidimensional scaling, and color-based image retrieval" paper presented at the Proceedings of the ARPA Image Understanding Workshop

Stich T, Linz C, Wallraven C, Cunningham D, Magnor M, 2011 "Perception-motivated interpolation of image sequences" *ACM Transactions on Applied Perception* **8**(2) 11:1 – 11:25

Tao B, Dickinson B, 1996 "Template-based image retrieval" paper presented at the Proceedings of the International Conference on Image Processing, Lausanne

Watson A B, 1993 *Digital Images and Human Vision* (Cambridge, MA: MIT Press)

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION

VOLUME 40 2011

www.perceptionweb.com