massachusetts institute of technology — artificial intelligence laboratory

Perceptually-based Comparison of Image Similarity Metrics

Richard Russell and Pawan Sinha

AI Memo 2001-014 July 2001
CBCL Memo 201

ABSTRACT

The image comparison operation – assessing how well one image matches another – forms a critical component of many image analysis systems and models of human visual processing. Two norms used commonly for this purpose 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 captures the perceptual notion of image similarity than the other. 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 via vector quantization. In both conditions the subjects showed a consistent preference for images matched using 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

Digital images play an increasingly important role in our lives. Their sheer numbers speak to their prevalence. According to popular estimates, 7/8 of data on the internet is graphical in nature. This preponderance of images necessitates automatic methods for their manipulation, storage, and use. Central to many operations on digital images are *image similarity metrics* (also called '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.

A very widely used class of image similarity metrics involve 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 L_P family of similarity metrics, or the Minkowski metric, after the Lithuanian physicist, Hermann Minkowski. The most commonly used members of this family are the L1 and L2 norms. The L1 metric is also called the Manhattan distance or the Mean Absolute Error (MAE), and the L2 metric is also called the Euclidean distance or Mean Square Error (MSE). The L1 and L2 metrics are described graphically in Figure 1. Examples of studies that employ these norms include: in image compression (Baker, 1982; Gersho, 1982; Goldberg, 1986; Mathews, 1989; Mathews, 1992), in image retrival (Rubner, 1997; Tao, 1996), in image quality assessment (Ahumada, 1993). Though many implementations utilize metrics that are more complicated, incorporating specific task-dependent features (Eckert, 1998; Frese, 1997; Jayant, 1993; Watson, 1993) they are often modifications made to simpler metrics such as L1 and L2.

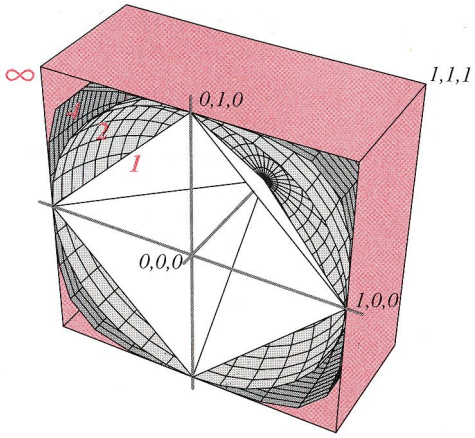

Figure 1. The surfaces of the hemi-octahedron and hemisphere represent points at a unit distance from the origin, as measured using the L1 metric and L2 metric, respectively. Both the L1 and L2 metrics involve taking the difference between the luminance values of corresponding pixels, i, in two images, x and y. In the L1 metric, the absolute value of the differences of each set of corresponding pixels is summed: L1 Distance (image x, image y)=$[\Sigma_i|x_i-y_i|]$ In the L2 metric, the differences are squared before being summed: L2 Distance (image x, image y)=$[\Sigma_i (x_i-y_i)^2]$ Smaller values correspond to greater similarity. (Image from Duda et al., 2001)

Currently both these metrics are used commonly and interchangeably. The L1 metric has the advantage of being slightly less computationally expensive, as no multiplication is needed for its calculation. 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 are 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 have commented on the perceptual interchangeability of the metrics (Mathews, 1997). Devore and colleagues 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 (DeVore, 1992). 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 potentially provide a well motivated way to decide between the L1 and L2 metrics. With this goal in mind, we experimentally investigated whether humans prefer the image matches chosen by the L1 or the L2 metric.

An important issue that arises in the design of such experiments is how to deal with the high-level semantic content in images. Such content (for instance, people, flowers, animals 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. This factor has complicated the interpretation of results in past studies of image similarity assessment. For instance, in a study by Rogowitz and colleagues (Rogowitz, 1998), subjects were asked to compare a reference image with eight test images, and decide which of the test images was most similar to the reference image. Figure 2 shows a trial from the experiment, with the reference image on the left and the eight test images on the right. The target images are structurally very different from the reference image, forcing a similarity judgment based on image semantics rather than structure.

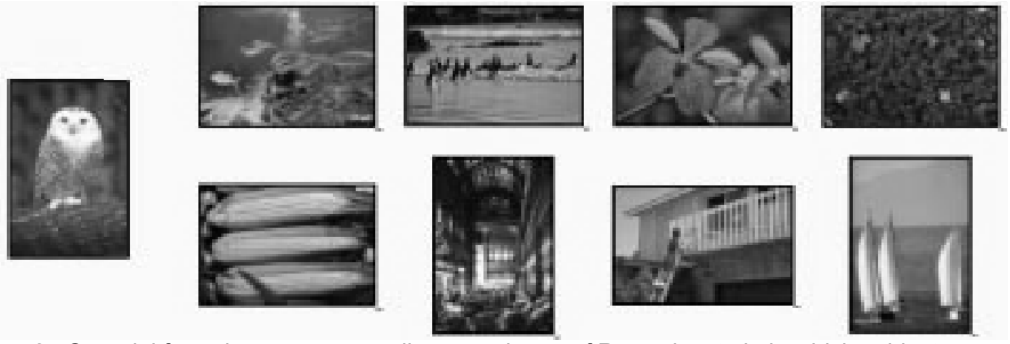

Figure 2. One trial from the computer scaling experiment of Rogowitz et.al., in which subjects were asked to decide which image on the right was most similar to the reference image on the left. The similarity judgment is based on entire images with very different low-level structure.

To avoid the problem imposed by uncontrolled semantic content, we conducted our comparisons of the two metrics in two different experiments. One experiment dispensed with semantics and the other preserved it. In the first experiment, subjects 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. In the second experiment, subjects viewed these single fragments in isolation, such that there was nothing recognizable in the images. The trials of both experiments involved asking subjects to decide which of two images better matched a target image. In these trials one of the two images was chosen/created from a library of images using the L1 metric and the other image was chosen/created from the same library using the L2 metric. Thus on each trial the subjects had to decide whether they agreed more with the similarity judgment of the L1 or the L2 metric.

2. GENERAL METHODS

Both experiments utilized the same two-alternative forced-choice design, in which subjects were instructed to choose which of two (probe) images looked most like a reference (target) image. Subjects indicated their preference by pressing a keyboard button, and were not timed, though they were allotted a maximum of 10 seconds 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 in a triangular formation, 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. Each experiment consisted of 384 trials.

Stimuli were presented using the DMDX experimental software (Forster, 1990) running on a PC under Windows 2000. Subjects sat approximately 75 cm from the display monitor in a room with low ambient illumination. Each image subtended approximately 2^{o} of visual angle. 23 people participated as subjects in the two experiments. One subject, RPR, participated in both experiments and is an author of this report. The remaining subjects were paid volunteers, eleven participating in the first experiment and eleven participating in the second. Each experiment thus employed 12 subjects.

3. EXPERIMENT 1

A. Methods

The images in the first experiment were generated using a technique for image compression called *vector quantization (VQ)*. Vector quantization involves dividing up an image into a grid of small fragments, and replacing each fragment with a similar image fragment from a *codebook*. In our experiment the codebook was a library of other images. Each fragment of the target image is compared with every other fragment of the same size in each image in the library using an image similarity metric. The target fragment is replaced by the fragment from the library that is judged to be the most similar by the metric. After this is performed on each fragment in the target image, a new image is created that is composed entirely of fragments from the library, and can serve as a compressed version of the target, since the image can then be represented by the indices of the fragments rather than the individual pixel values. Figure 3 shows an example of image reconstruction via vector quantization.

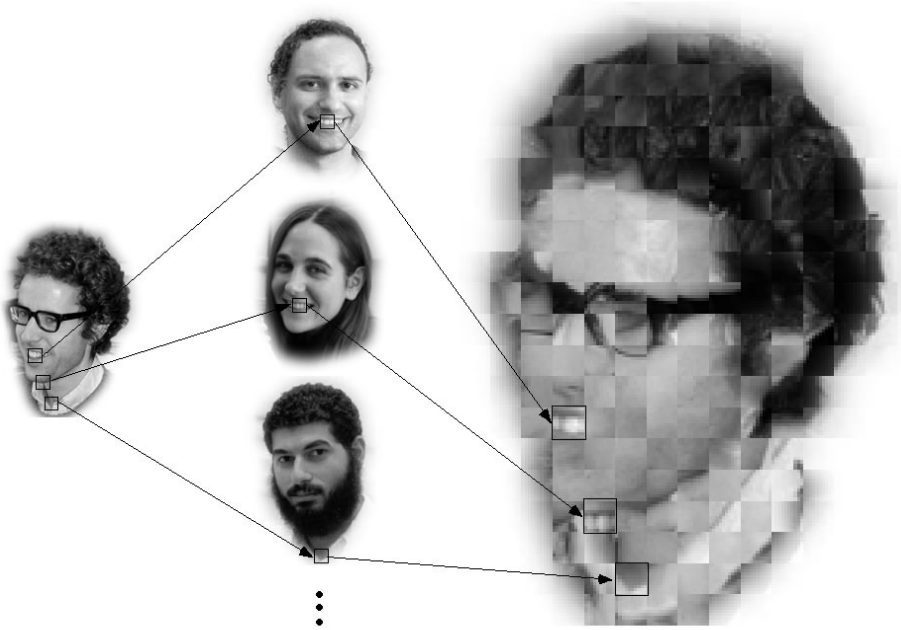

Figure 3. In vector quantization, fragments of the target image are matched to fragments in the library images using an image similarity metric. These fragments from the library images then replace those in the target image, creating a reconstructed image that is composed entirely of fragments from the 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. Compression derives from the ability to represent each block (here of 10x10 pixels) by 3 numbers (image index, x, y) rather than 100 individual pixel values. The poor quality of the reconstruction here is due to the very small size of the codebook (5 images).

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 one 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 rescaled to 150x200 pixels and converted to grayscale using Adobe Photoshop. 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 4 is a montage of thumbnails of all 72 images. Reconstructed images were always created in pairs using the L1 metric in one case and L2 in the other.

Figure 4. 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.

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 and four different library sizes. We used fragments of 5x5 pixels, 10x10 pixels, 15x15 pixels, and 20x20 pixels. For library sizes we used 6, 12, 24, and 48 images. Smaller libraries were subsets of larger libraries (e.g. all of the images used in the 12 image library were used in the 24 image library). Thus there were 24 target images x 4 fragment sizes x 4 library sizes = 384 pairs of reconstructed images. So 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 5 shows an example trial from the condition with block size of 20x20 pixels and a library size of 48 images.

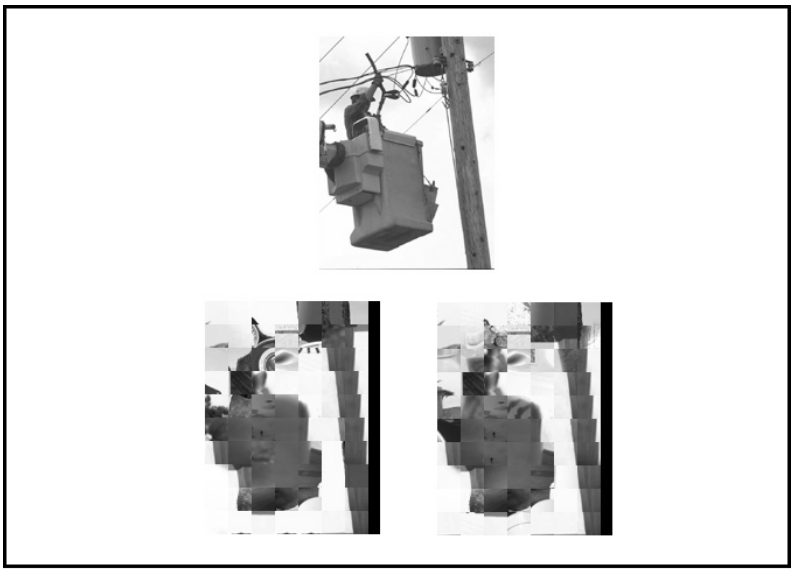

Figure 5. Each trial consisted of a display similar to the figure above. The top image is the original. The bottom pair of images are reconstructions, one based on the L1 norma as the distortion function and the other on the L2 norm. Subjects were instructed to indicate which of the two bottom images looked more like the top image. The left-right ordering of the L1 and L2 reconstructions was counterbalanced across trials and conditions.

B. Results

Subjects 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 using the L1 metric. Of the 12 subjects, 11 chose the L1 metric on more than 50% of the trials and the remaining subject chose the L1 metric on 49% of the trials. Subjects chose images reconstructed using the L1 metric significantly more often than those created using the L2 metric (Student's t-test, $p=0.002$). The results are shown in Figure 6. There were no significant differences across the different fragment sizes (single factor ANOVA, $p=0.69$) or library sizes (single factor ANOVA, $p=0.61$). Tables 7 and 8 show the results by fragment size and library size.

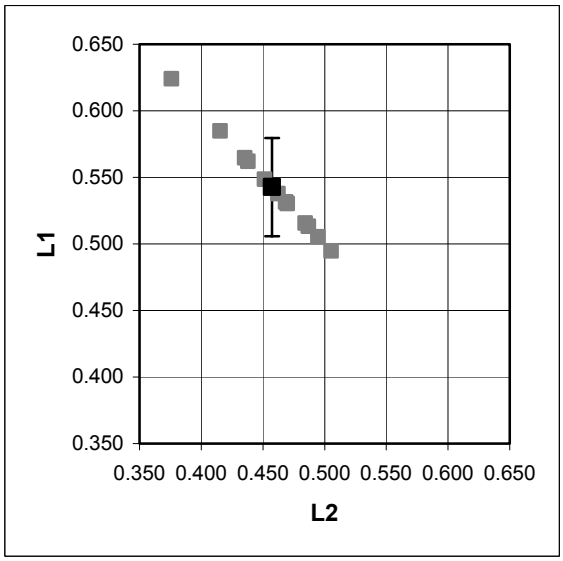

Figure 6. Subject preferences for the L1 metric plotted against preferences for the L2 metric. Each point represents a single subject. The mean of the data set is represented by the black square with error bars (the error bars are shown vertical rather than aligned with the data axis for the sake of clarity). On the x-axis is the proportion of trials on which the subject chose the L2 reconstruction. On the y-axis is the proportion of trials on which the subject chose the L1 reconstruction. Because the proportion of choices for either the L1 or the L2 metric must equal one, all of the points lie on the line y = 1-x. Points above the line y = x indicate more choices in favor of the L1 metric than the L2 metric, while points below the line y = x indicate more choices favoring the L2 metric. All but one of the points lie above the line.

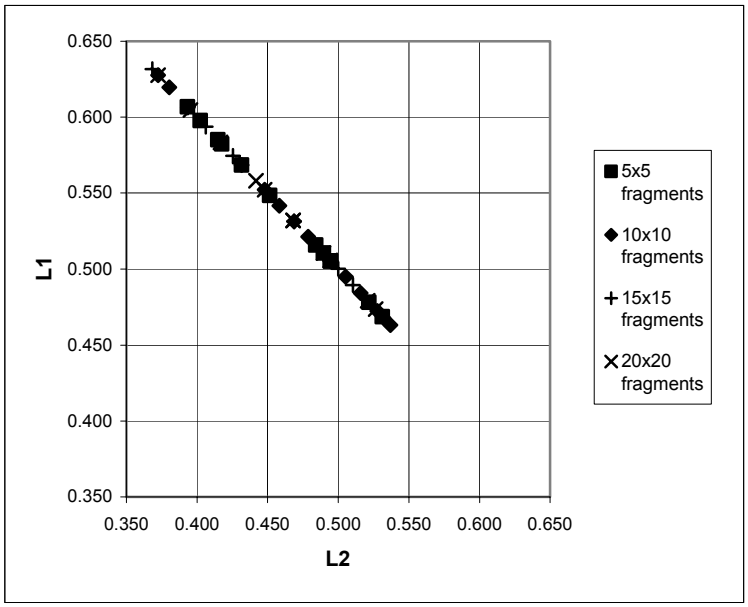

Figure 7. Subject preferences for the L1 metric plotted against preferences for the L2 metric across four different fragment sizes. Each point represents a single subject's responses to reconstructions with a given fragment size. Preferences for the different metrics did not differ significantly by fragment size.

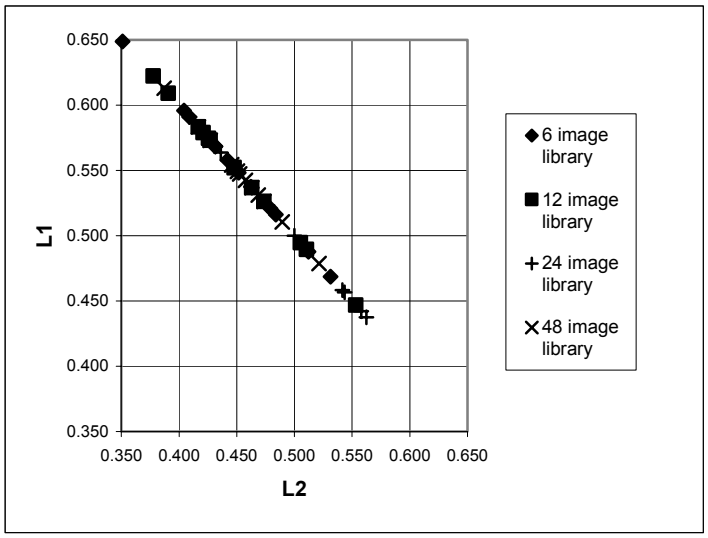

Figure 8. Subject preferences for the L1 metric plotted against preferences for the L2 metric. Each point represents a single subject's responses to reconstructions with a given library size. Preferences for the different metrics did not differ significantly by library size.

4. EXPERIMENT 2

A. 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 20x20 pixel fragment size and 48 image library size condition. 384 sets of original fragments with their L1 and L2 image matches were used. Sets of fragments were chosen such that the L1 and L2 matches of different sets ranged from fairly similar to distinctly different. The similarity of the pairs was determined by calculating the L1 distance between the two matches. It is important to note that the choice of the L1 metric for this determination in no way biases the subsequent results toward either of the two norms. Because these fragments were quite small, they were enlarged to 60x60 pixels using bicubic interpolation. Figure 9 shows an example trial.

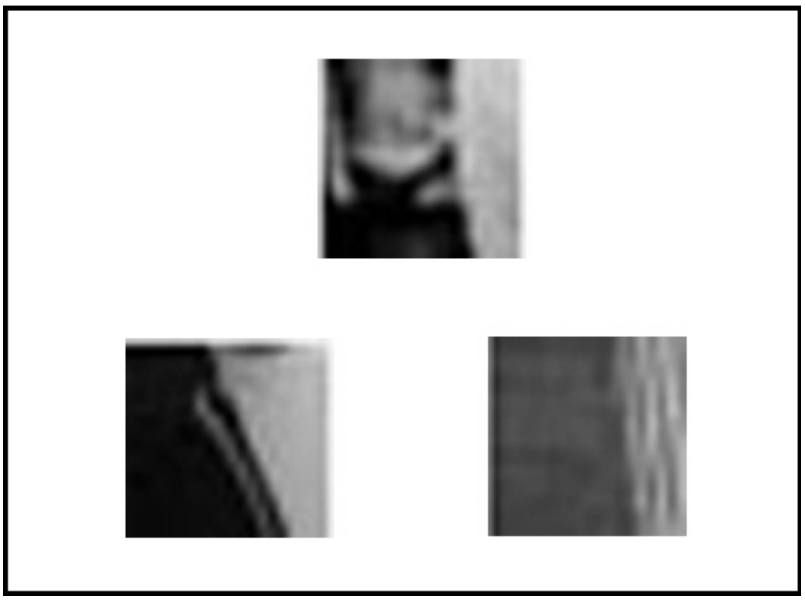

Figure 9. Each trial consisted of a display similar to the figure above. The top image is the original. The bottom image pair comprises the best matching fragments retrieved using the L1 or the L2 metrics. Subjects were instructed to choose which of the two bottom images better matched the top image. The ordering of the L1 and L2 reconstructions was counterbalanced across trials and conditions.

B. Results

The results from experiment 2 were very similar to those from experiment 1. 54% of all responses were made for the match found using the L1 metric. Of the 12 subjects, 11 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 subject's choices of the L1 metric were significantly greater than their choices of the L2 metric (Student's t-test, p=0.00009). The results are shown in Figure 10.

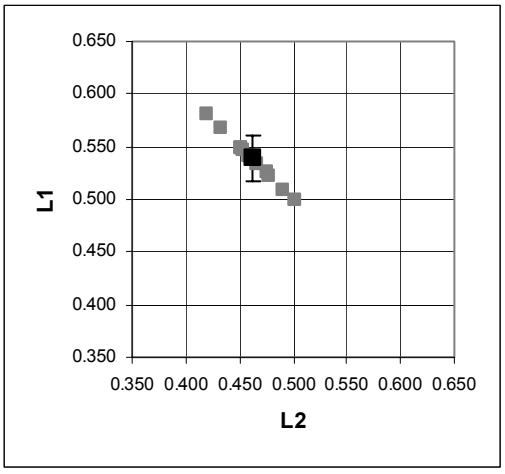

Table 10. Subject preferences for the L1 metric plotted against preferences for the L2 metric. Each point represents a single subject. The mean of the data set is represented by the black square with error bars. On the x-axis is the proportion of trials on which the subject chose the L2 reconstruction. On the y-axis is the proportion of trials on which the subject chose the L1 reconstruction. Points above the line y = x indicate more choices of the L1 metric than the L2 metric, while points below the line y = x indicate more choices of the L2 metric than the L1 metric. All but one of the points lie above the line. One point lies exactly on the line

5. DISCUSSION

We experimentally examined human preferences for the notion of image similarity embodied in two widely used computational metrics. We found a highly consistent preference for the L1 based matches in our results.

These results point to a host of interesting extensions and open questions. The vector quantization reconstructions that we used in the first experiment suffered from block artifacts (see figure 5). A possible extension of the work described here would be to perform a similar experiment, but with a scheme to reduce the block artifacts. A simple method of achieving this would be to use overlapping windows for the fragments, as depicted in Figure 11. This could be accomplished very simply by windowing the fragments to create Gaussian or linear ramp boundaries rather than step edges.

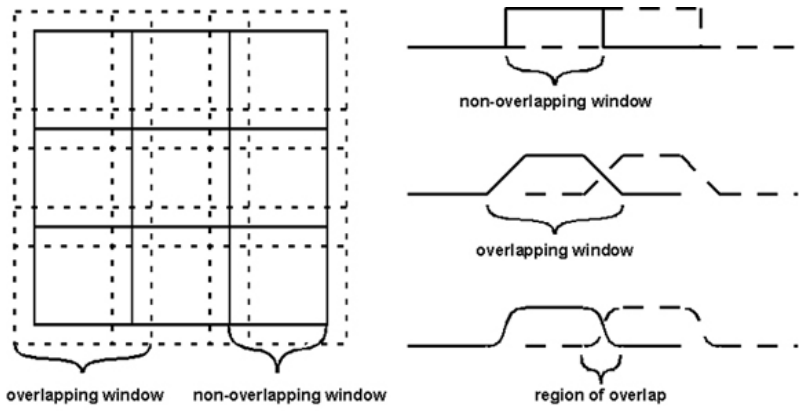

Figure 11. The figure at left shows both overlapping and not-overlapping windows. The solid lines represent non-overlapping windows, which result in blocking artifacts at the window junctions. The dashed lines represent overlapping windows, where the pixel values in the regions of overlap are influenced by more than one fragment. The figures at right are cutaway views of non-overlapping windows (top), overlapping windows with linear convolutions (middle) and overlapping windows with gaussian convolutions (bottom).

Though they are among the most commonly used distance metrics, the L1 and L2 norms are far from being the only available options. It would be very useful to have information about the perceptual dimensions of other similarity metrics beside the L1 and the L2. This would be helpful in the selection of appropriate similarity metrics for the many computational applications that require them. In general, this experimental methodology may be helpful in the design of new similarity metrics based on perceptual considerations.

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 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.

6. 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*, 55-61.

DeVore, R. A., Jawerth, B., & Lucier, B. J. (1992). Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, **38**(2), 719-746.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). Pattern Classification, 2nd edition, John Wiley and Sons Publishers.

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, Arizona: University of Arizona.

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.

Jayant, N., Johnston, J., & Safranek, R. (1993). Signal compression based on models of human perception. *Proceedings of the IEEE*, **81**, 1385-1422.

Mathews, V. J., & Khorchidian, M. (1989). *Multiplication-free vector quantization using L/sub 1/ distortion measure and its variants*. Paper presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89).

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

Mathews, V. J., & Hahn, P.J. (1997). Vector quantization using the L-infinity distortion measure. *IEEE Signal Processing Letters*, **4**(2), 33-35.

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, May 1997). *The earth movers distance, multidimensional scaling, and color-based image retrieval*. Paper presented at the Proceedings of the ARPA Image Understanding Workshop.

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. (Ed.). (1993). *Digital images and human vision* . Cambridge, MA: MIT Press.