# Perceptual Image Similarity Experiments*

Bernice E. Rogowitz[†], Thomas Frese[‡], John R. Smith[†], Charles A. Bouman[‡] and Edward Kalin[†]

[†]IBM T.J. Watson Research Center,
P.O. Box 218,
Yorktown Heights, NY 10598
{rogowtz, jrsmith, ekalin}@watson.ibm.com

[‡]School of Electrical Engineering,
Purdue University,
West Lafayette, IN 47907-1285
{frese, bouman}@ecn.purdue.edu

## ABSTRACT

In this paper, we study how human observers judge image similarity. To do so, we have conducted two psychophysical scaling experiments and have compared the results to two algorithmic image similarity metrics. For these experiments, we selected a set of 97 digitized photographic images which represent a range of semantic categories, viewing distances, and colors. We then used the two perceptual and the two algorithmic methods to measure the similarity of each image to every other image in the data set, producing four similarity matrices. These matrices were analyzed using multidimensional scaling techniques to gain insight into the dimensions human observers use for judging image similarity, and how these dimensions differ from the results of algorithmic methods. This paper also describes and validates a new technique for collecting similarity judgments which can provide meaningful results with a factor of four fewer judgments, as compared with the paired comparisons method.

## 1 Introduction

Advances in digital cameras, large accessible data storage, internet repositories, and image applications have fueled the recent development of methods for searching, retrieving, and navigating through a set of images.[1–6] In a typical search task, the user selects an image and asks the computer to retrieve images which are similar. The computer compares the features of the selected image with the characteristics of the other images in the set and returns the most similar images for inspection. Typically, this is done by computing, for each image, a vector containing the values of a number of attributes and computing the distance between image feature vectors. The best image matches are typically displayed to the user as an array of "postage stamp" sized images, in descending order of this computed distance. These methods produce several well-documented artifacts. The methods, for one, do not include any knowledge about the intrinsic organization of the images. For example, if a query image is equidistant between two clusters, the Euclidean metric will return the images in metric order, interdigitating images from the two clusters. The success of these metrics, furthermore, depends largely on the power of the identified features. Often, these features are thought to contribute to human judgments of image similarity (e.g. color, texture, and shape), but with few exceptions,[7,6] the characteristics of human similarity perception have not been included in the selection of these features. The third artifact of these methods is that the images are displayed in metric order, but not in an order which is conducive for navigation. Similar images are in order by row, but even if the ordering is perfect, similar images are not near each other in the array. Recently, the idea of representing images in a multidimensional, browsable space has been proposed.[8,9] These navigation spaces, however, are not perceptual, and therefore do not

* Published in Human Vision and Electronic Imaging III, Bernice E. Rogowitz and Thrasyvoulos N. Pappas, Editors, Proceedings of the SPIE, 3299, San Jose, CA, January 26-29, 1998.

give the user intuitive cues for finding images which are "redder", "darker", "out-of-doors", "more natural", or "containing more people".

The purpose of this perceptual image similarity project is to examine the issue of image similarity from the perspective of the human observer. In this paper, we present two psychophysical scaling experiments in which we measure the similarity of 97 carefully selected digital photographic images. In these experiments, each image is compared with every other image, and the observer judges their perceived similarity. The users are free to use whatever criteria they choose for making this decision. We compare these perceptual judgments to similarity ratings produced by two algorithmic methods, one based on similarities in the color histograms of the images, the other based on more sophisticated perceptually-relevant features.[6] We evaluate these perceptual and algorithmic similarity ratings using multidimensional scaling techniques and explore the results of two- and three-dimensional MDS solutions.

The next step in this research is to model perceptual image similarity judgments in terms of calculable image features. This would allow us to extend our findings to the development of better image search and retrieval methods, and to the development of more intuitive navigation spaces.

# 2 Experimental Design

In these experiments, human observers judged the similarity of 97 carefully-selected JPEG images. Two psychophysical scaling methods were used to measure the perceived similarity of each image with every other image in the set. In the "Table Scaling" experiment, observers organized printed thumbnail images on a tabletop so that similar images would be near each other and dissimilar images would be far apart. This design allowed the observer to see all the images at once. In the "Computer Scaling" Experiment, the images were presented on a computer display. On each trial, a reference image was compared with eight randomly-chosen images, and the observer selected the image which was most similar. Both experimental procedures produced a similarity matrix which served as the input to multidimensional scaling algorithms. We also calculated image similarity matrices for these 97 images using the algorithmic similarity metrics. The multidimensional scaling results for these algorithmic matrices were compared with the psychophysical results.

## 2.1 Selecting the Stimuli

Ninety-seven JPEG images were selected from the PhotoDisc collection of 5000 photographic images. This set was chosen because they were reputed to be of good photographic quality, and had been used by other researchers in this field.[3,6] These images were selected according to three explicit criteria. First, we wanted to make sure that we included a wide range of topics. To achieve this goal, we consulted books designed to teach photography and matched our selections to these focus areas. These were: animals, people, indoor scenes, nature, buildings, textures, and man-made. The photographic textbooks also focused on the distance of the object from the viewer, and so, for each category, we explicitly selected images which were wide-angle, normal (e.g., 50 mm for 35 mm film), or close-up. In order to insure that we had selected a set of images which covered a broad range of colors and light levels, we measured each image in calibrated CIELab space, and iterated on the image samples until we achieved a balanced distribution in all three dimensions of this color space. The images were in both landscape and portrait modes.

## 2.2 Observers

Fifteen volunteer observers were recruited from the T.J. Watson Research Center, Hawthorne Laboratory. They were 12 men and 3 women, ranging in age from approximately 23 to 45. Observers received a lunch voucher for each session.

## 2.3 Experiment 1: Table Scaling

For this experiment, we printed the 97 JPEG images on a color printer at 300 dots per inch. The size of the prints was 3 cm by 2 cm, on a 3.5 cm by 3.5 cm white background.

The images were placed randomly on a large round table, and the observers' task was to arrange them so that the physical distances between them were inversely proportional to their perceived similarity. That is, the more similar the images appeared, the closer they were to be placed next to each other. Since the table surface was, by definition, a 2-D surface, this experiment forced the observers to project the multidimensional relationships between the images down to a two-dimensional space.

Nine subjects served in this experiment. Each session took 30-45 minutes, and the observers found this to be a "fun, puzzle-like" task. For each observer's solution, we measured the physical distance between each pair of images and created a similarity matrix. We created a pooled matrix by accumulating these distances across subjects.

## 2.4 Second Experiment: Computer Scaling

In the second experiment, fifteen observers compared each of the 97 thumbnail images with every other image. In the traditional multidimensional scaling paradigm, these measurements would be made using a "paired comparisons" design. In paired comparison, all stimulus pairs are compared, and for each pair, the observer assigns a number proportional (or inversely proportional) to perceived similarity. We modified this procedure for two reasons. First, with $n = 97$ stimuli, the paired comparisons procedure would require $n(n-1)/2$ comparisons, which, in this case, would be 4656 trials. Second, we were seeking a procedure which would not depend on subjective magnitude judgments, since these can be prone to bias. We therefore developed a variation on the traditional paradigm. In this variation, each image was compared with each other, taken eight at a time, thereby reducing the number of trials by a factor of four. Thus, on each trial, the observer viewed a randomly-selected reference stimulus and eight test stimuli, selected randomly from the set of 97, and judged which of the eight appeared most similar.

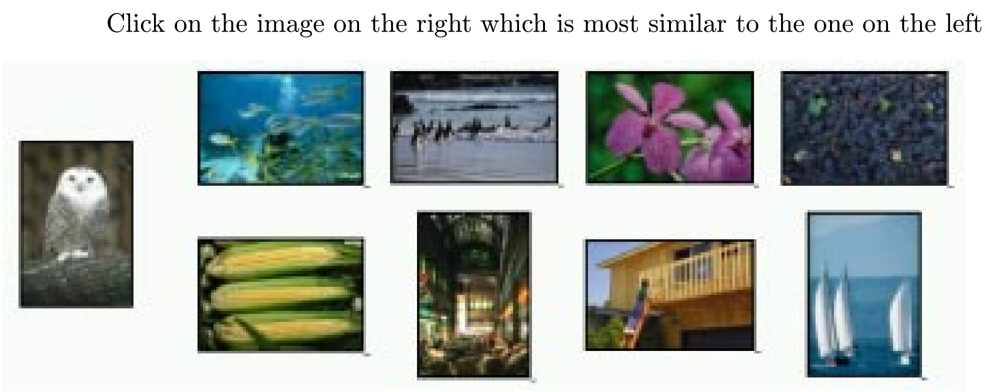
The experiment was conducted on a color-calibrated display monitor, using the Netscape browser to present the stimulus images. The display measured 32 cm by 24 cm, and subtended roughly 40 by 30 degrees of visual angle when viewed at a distance of approximately 47 cm. The image bitmaps used in this experiment measured 123 by 83 pixels. Viewed on the display monitor, the size of each image was approximately 4.7 cm by 3.1 cm, and subtended approximately 6 by 4 degrees of visual angle. Figure 1 shows one trial of this experiment, as displayed to the observer on the computer display. The reference stimulus was presented along the left edge of the display, accompanied by two rows of four test stimuli running horizontally along the display. On each trial, the observer evaluated the eight test stimuli, judged their similarity to the reference stimulus, then used a mouse to "click" on the most similar test image. This response launched the next trial of the experiment. The experiment was self-paced, and observers could take breaks whenever they chose. The 1164 trials and 20 practice trials typically took three 1-hour sessions to complete.

In addition to these objective judgments, we also collected verbal protocols from the observers, asking them to free-associate about the reasons they had for making their selections. The purpose of this technique was to develop a better intuition for what observers thought they were doing, and to gain insight into candidate dimensions of image similarity.

### 2.4.1 Validation of the Experimental Design

In our Computer Scaling paradigm, each image appears as a reference stimulus twelve times, and on each of these trials, is paired with a randomly-selected set of eight test stimuli. Each observer selects just one "most-similar" image on each of these trials, and so, over the course of the experiment, data are only entered into 12 cells in each 97-cell row of the similarity matrix. To create the overall similarity matrix, results of the individual observers were accumulated. In particular, if a subject selected test image $k$ as a match to reference image $r$, then the $rk$-th entry of the similarity matrix was incremented by one. With fifteen observers, there are only 15 x 12 votes in each row of the similarity matrix, or 180 distributed over the 97 images. This produces a rather sparse similarity matrix.

In order to verify that these sparse measurements can lead to a meaningful dataset, we performed a Monte Carlo simulation. The simulation assumes a "true" similarity matrix $S$ whose entries $s_{ij}$ are between zero (least similar) and one (most similar). Preliminary experiments showed that most images are perceived as being very dissimilar to each other. Therefore, similarity matrices tend to be sparse and contain mostly small entries. For the simulation, the similarities were assumed to be approximately exponentially distributed[10] with mean $\lambda^{-1} = 0.5$. Furthermore, the matrix was assumed to be 30% sparse which we consider a conservative estimate. This leads to the probability



Click on the image on the right which is most similar to the one on the left.

Figure 1: One trial of the computer scaling experiment. The reference stimulus on the left is compared to the eight test stimuli on the right. The task was to select the one test stimulus that is most similar to the reference.

distribution

$$p_{s_{ij}}(s) = 0.3\delta(s) + 0.7\frac{\lambda e^{-\lambda s}}{1-e^{-\lambda}}(\mathrm{u}(s) - \mathrm{u}(s-1)) \qquad \text{for} \quad i \neq j \tag{1}$$

where $\delta$ is the Dirac-delta and u denotes the unit-step function. The similarity matrices used for the simulation were obtained by sampling from this distribution and symmetrizing the result. Notice that the symmetry assumption is a consequence of our intent to embed the estimated similarity matrix into a metric space using multidimensional scaling.

In order to model the subject's choice behavior, we assume a choice process based on Gaussian confusion probabilities. For illustration, assume that the subject is presented with one reference image $r$ and two test images $t_1$ and $t_2$. Let $s_{rt_1}$ and $s_{rt_2}$ denote the similarities of the reference image to test images one and two respectively. Then, assuming without loss of generality that $s_{rt_1} < s_{rt_2}$, the probability of selecting the test images is assumed to be

$$P(t_1) = \frac{1}{2}e^{-\frac{(s_{rt_1}-s_{rt_2})^2}{2\sigma^2}} \tag{2}$$

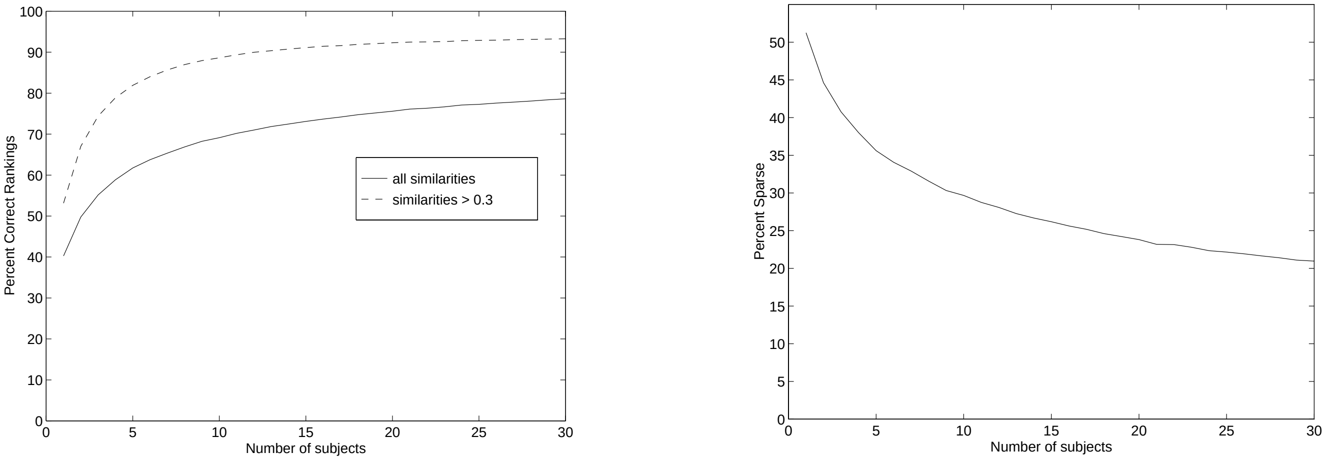
$$P(t_2) = 1 - \frac{1}{2}e^{-\frac{(s_{rt_1}-s_{rt_2})^2}{2\sigma^2}}. \tag{3}$$

For our subject experiment where eight test images are displayed at each trial, this generalizes as follows: Assume we wish to calculate the probability for choosing test image $t_j$, $1 \leq j \leq 8$. We first divide the 7 remaining test images $t_i$ into two sets $T_j^{(1)}$ and $T_j^{(2)}$. The first set $T_j^{(1)}$ contains those test images that are more similar to the reference image than $t_j$, i.e. $T_j^{(1)} = \{t_i : s_{rt_i} > s_{rt_j}\}$. The second set $T_j^{(2)}$ contains the test images that are less or equally similar to the reference image than $t_j$, i.e. $T_j^{(2)} = \{t_i : s_{rt_i} \leq s_{rt_j}, i \neq j\}$. Generalizing the two-test case from above, we obtain

$$P(t_j) = \frac{\prod\limits_{t_i \in T_j^{(1)}} \frac{1}{2}e^{\frac{-(s_{rt_i}-s_{rt_j})^2}{2\sigma^2}} \prod\limits_{t_i \in T_j^{(2)}} (1 - \frac{1}{2}e^{\frac{-(s_{rt_i}-s_{rt_j})^2}{2\sigma^2}})}{\sum\limits_{k=1}^{8}\left(\prod\limits_{t_i \in T_k^{(1)}} \frac{1}{2}e^{\frac{-(s_{rt_i}-s_{rt_k})^2}{2\sigma^2}} \prod\limits_{t_i \in T_k^{(2)}} (1 - \frac{1}{2}e^{\frac{-(s_{rt_i}-s_{rt_k})^2}{2\sigma^2}})\right)} \tag{4}$$

where the denominator normalizes the sum of the probabilities $P(t_j)$ to one. The standard deviation used was $\sigma = 0.1$.

Using the assumptions above, we performed the Monte Carlo simulation by generating experimental trials and selecting image matches according to (4). The estimate $\hat{S}$ was then formed as follows: Starting with a matrix of zeros, the entry $\hat{s}_{ij}$ was incremented by one if the test image $j$ was selected as a match to reference $i$. The matrices were added across the hypothetical subjects, symmetrized and divided by the number of subjects.



(a)

(b)

Figure 2: Results of the Monte Carlo simulation. Shown in (a) is the percentage of correct similarity rankings as a function of the number of subjects. The solid line evaluates correct rankings for all matrix entries, whereas the dashed line shows percentage correct for entries greater 0.3 only. Shown in (b) is the sparsity of the estimate.

Notice, that the estimates $\hat{s}_{ij}$ are biased, i.e. they will not converge to the true $s_{ij}$ as the number of subjects increases. However, the magnitude *rankings* of the $\hat{s}_{ij}$ converge to those of $s_{ij}$. Since we evaluate the data with respect to its value for the multidimensional scaling procedure we are mostly interested in correct rankings as compared to the correct numerical values.

In order to evaluate the simulation results, we calculated the sparsity of the estimate as well as the percentage of correct ranking relationships of matrix entries. The percentage of correct ranking relationships was computed by examining all pairs {(i,j),(k,l)} of matrix entries and comparing whether their ranking relationship was the same in $S$ and $\hat{S}$.

The results are shown in Fig. 2. Figure 2(a) shows the percentage of correct similarity rankings as a function of the number of subjects. The solid curve shows correct rankings evaluated over all matrix entries. For 15 subjects, the percentage of correct rankings is 73 %. The dashed curve evaluates correct rankings only for entries $s_{ij}$ greater than 0.3. In particular, these entries were paired with all other entries and ranking relationships were compared as in the general case. The percentage correct in this case is higher than in the general case. For 15 subjects, we obtain 91.1% correct rankings. The interpretation is that we measure the larger similarities more exactly than the small similarities. This is consistent with our design using sparse measurements and selecting only the highest similarity match at each trial. Figure 2(b) shows the sparsity of the estimate. Notice that sparsity here is defined as the percentage of zero estimates for non-zero entries in the true $S$-matrix. Ideally, the sparsity should be zero. However, considering, that most of the true matrix entries are assumed to be small, we consider a sparsity of 26.2% for 15 subjects satisfactory. Notice that these results are fairly robust with respect to small changes in the simulation assumptions and parameters. Concluding, the results suggest that using our experimental design, data from 15 subjects should be sufficient to obtain a meaningful estimate of the similarity matrix for multidimensional scaling evaluation.

## 2.5 Computing Similarity Matrices for Image Processing Algorithms

In the above experiments, we use perceptual judgments to measure the distance between the 97 images. For comparison, we also used algorithmic image similarity metrics to characterize these distances. The first metric is based only on global color-histograms of the images. In particular, this metric computes image dissimilarity as the $L_1$-norm of color-histograms in CIELab color-space. The number of bins used for the histograms was 8/16/16 for the $L$, $a^*$ and $b^*$ channels respectively. While this is a very simple algorithm, algorithms like these are often an important component of more sophisticated measures.[3,5]

The second metric[6] is a multichannel model derived from models for image quality assessment. This metric is based on a multiresolution framework of color, contrast and orientation-selective attributes. Image dissimilarity is calculated as a weighted combination of attribute differences. The relative weights of the attributes are based on a training set of perceptual judgments.

# 3 Results

## 3.1 Qualitative Findings

Figure 3 shows the accumulated results for one image for all fifteen observers in the computer scaling experiment. The number of "votes" is indicated below each of the ten most frequently selected images. Although individual observers used different verbal descriptors of their behavior when making these judgments, their selections were very similar. In this example, an image of three snowdrop flowers is judged to be most similar to other images showing a countable number of colorful flowers, followed by other foliage and flower scenes, outdoor scenes with lots of green foliage and color, and later by an image of foliage with animal figures.

Figure 4 shows a similar result for a very different image. Here, the picture of the young girl with a baseball glove is matched most frequently to other images including children, then to a solitary portrait, followed by group images which include either a child or outdoor sports. Qualitatively, thus, there seemed to be a high agreement across subjects in their rating of image similarity.

Although these matches were very consistent across observers, they were not always symmetric. For example, a nature scene might commonly be matched to nature scenes including children. However, when these nature scenes containing children are shown as reference, they will most frequently be matched to images of other children and not to nature scenes without children. This is an important point; the MDS analysis does not capture such non-symmetric behavior, since the images are embedded in a metric space. In future work we will explore methods to analyze these aspects of the data.

## 3.2 Multidimensional Scaling

Multidimensional scaling is a psychometric procedure originally designed to estimate the perceived distances between stimuli which vary along a large number of dimensions, where the goal is to uncover the dimensions along which these judgments are made.[11] This concept has been used, for example, to discover the perceptual dimensions of perceptual relationships between notes in a musical scale[12] and the dimensions of color sensation.[13] It has been used in the image analysis community to identify the perceptual dimensions of textures,[14] and in the image retrieval community to measure the similarity of images based on their color histograms[8,9] as well as to explore the behavior of a texture perception algorithm.[15]

The goal of the multidimensional scaling procedure is to place objects that are specified only by their distances into a lower dimensional space. In particular, the input to the MDS algorithm is a distance matrix of the pairwise object distances. The procedure then places the objects into a metric space while preserving the distances as much as possible. Notice that there is a large variety of MDS algorithms which differ in their assumptions and optimization strategies. For an excellent overview, the reader is referred to.[16]

The multidimensional scaling procedure employed here is metric least-squares scaling. We first convert the similarity matrix $S$ from the subject experiment to a dissimilarity matrix $D$ as

$$d_{ij} = 1 - s_{ij}. \tag{5}$$

Now assume that the images $i$ and $j$ are placed at position $x_i$ and $x_j$ in the metric space. We then define $\tilde{d}_{ij}$ as the Euclidean distance $\tilde{d}_{ij} = \| x_i - x_j \|_2$. The optimization or stress function $\theta$ used for the MDS algorithm is then

$$\theta = \frac{\sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} (\tilde{d}_{ij} - d_{ij})^2}{\sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} d_{ij}^2} \tag{6}$$

where $N$ is the number of images in the set. The minimization of $\theta$ is performed using a conjugate gradient optimization.

We used this multidimensional scaling algorithm to reduce the dimensionality of the four similarity matrices we described above, derived from, 1) the color histogram metric, 2) the perceptually- based image similarity metric, 3) the Perceptual Table Scaling data, and 4) the Perceptual Computer Scaling experiment. In order to compare the MDS results for the different methods, we allowed for translations and rotations to minimize the differences in a least square sense.

## 3.3 The multidimensional scaling results

We performed a multidimensional scaling analysis in two and three dimensions for each of the similarity matrices derived from the different methods.

### 3.3.1 MDS in 2 dimensions

Figure 5 shows the two-dimensional result obtained when using the color histogram algorithm to compute distances between images. Not surprisingly, the two-dimensional solution reflects the influence of color. We have added some color words to the figure to simplify interpretation. Although the images do seem to segment into color regions, these regions are rather diffuse. For example, images with a lot of green span from the landscape view of foliage at the bottom left through the landscape view of agricultural terraces in the middle of the image, and beyond. Looking at this projection, we clearly see that equal distances in the space do not correspond to equal differences in perceived color. For example, the foliage image in the bottom left and the terraces image are very similar in overall color but are quite far apart in this space. Conversely, the terraces image is right next to a sailboat image whose overall color is quite different. This may be due to the high stress (0.27) of the MDS solution. It may also be due to the fact that overall perceived color is not well-described by this simple algorithm. We believe, for example, that the color of the main object in the image contributes more to the overall impression of the color than would be expected given its pixel contribution. That is, we would expect close-up images of people to be close together perceptually because the color of the skin is more important to the overall impression of image color than the color of the background or the color of the clothing. To test results of this type, we are planning to compare them with semantic color descriptors for these 97 images, such as the "overall color", the "color of the main object", "the background color", etc.

One very interesting observation from these data is that overall color alone does capture information about the semantic meaning of these images. For example, a large number of the outdoor, natural scenes appear near to each other in this projection, suggesting that color alone carries semantic information. Since it is very difficult to develop algorithms which capture the semantic information in images, understanding the correlations between color and semantics could be a useful tool in developing semantic image descriptors which rely on calculated image attributes.

Figure 6 shows the results of the 2D MDS for the perceptually-based image processing metric. This algorithm contains terms which capture color attributes, and the influence of color is apparent in this projection. With this algorithm, the achromatic and brown images are at the center of the projection, and the spectral colors rotate around this hub. This algorithm is designed to capture additional features in the image. Low spatial-frequency color images are grouped together which can be seen especially within color regions in this projection. For example, the images with large blue expanses with brown subjects in the foreground are together in the top right region, including buildings, windmills, and horses. The portraits of humans, with large expanses of skin color against a darker background, are also organized together. Images with high-spatial-frequency luminance modulations are also near each other. Looking toward the center of the projection, for example, there is a cluster of traffic, crowd, city street and texture images. The perceptually weighted attributes, low-spatial-frequency color and high-spatial-frequency luminance, do seem to produce a more perceptually-plausible organization of these images than did the color histogram alone, capturing some compositional attributes in the images.

Figure 7 shows the two-dimensional multidimensional scaling result of the perceptual table scaling experiment. The popularity of color-histogram-based image similarity metrics is validated by these perceptual results. Overall color does seem to play a significant role in the perceptual organization of these images. The color organization, however, is much less pronounced, suggesting that other factors are playing significant roles in these judgments. There appear to be groupings of blue, green, and brown/orange images, and perhaps achromatic images. The attributes described in the result for the perceptually-weighted algorithm also seem to be operating. Within color areas, there

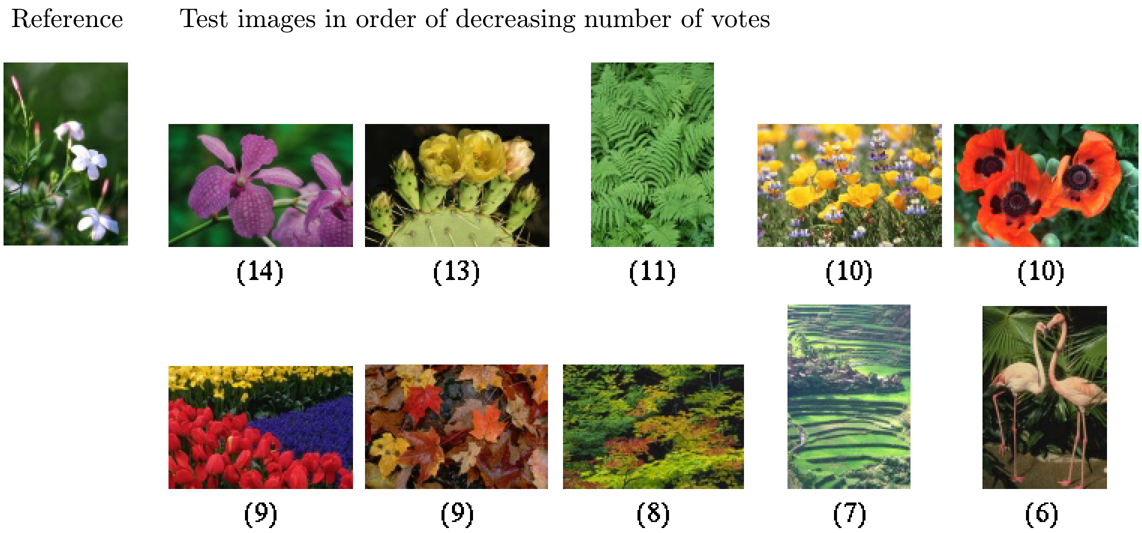


Figure 3: Computer Scaling matching results for snowdrop-flowers image. The two rows on the right show the test images that were most frequently matched to the reference on the left. Indicated below each image is the number of votes this match received across the 15 subjects. In this case, the reference was most frequently matched to other images showing a small number of flowers, followed by other foliage and flower scenes.

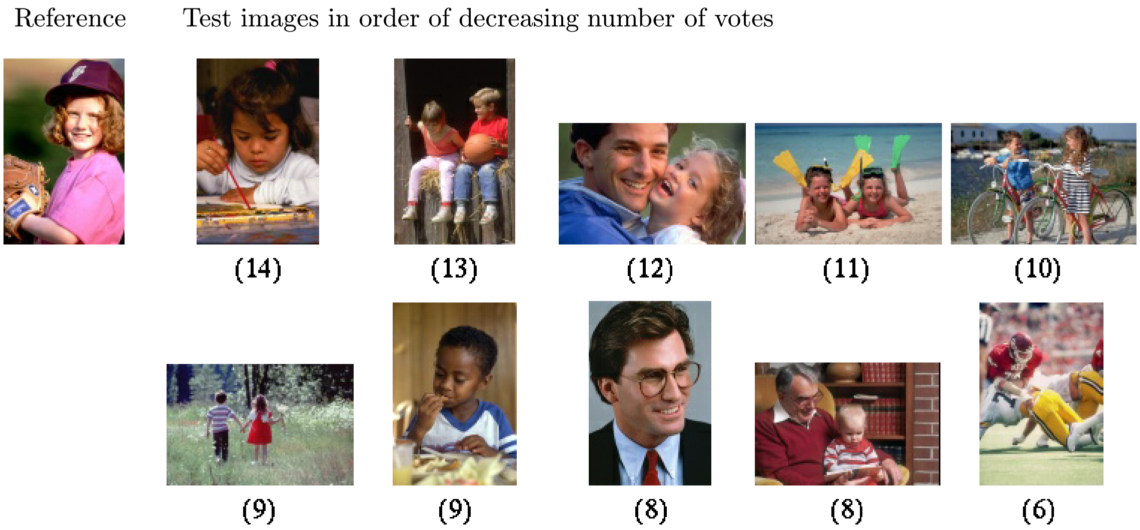


Figure 4: Computer Scaling matching results for “girl-baseball-glove” image. This image is most frequently matched to images of other children, followed by a portrait and group images containing either children or sports scenes.

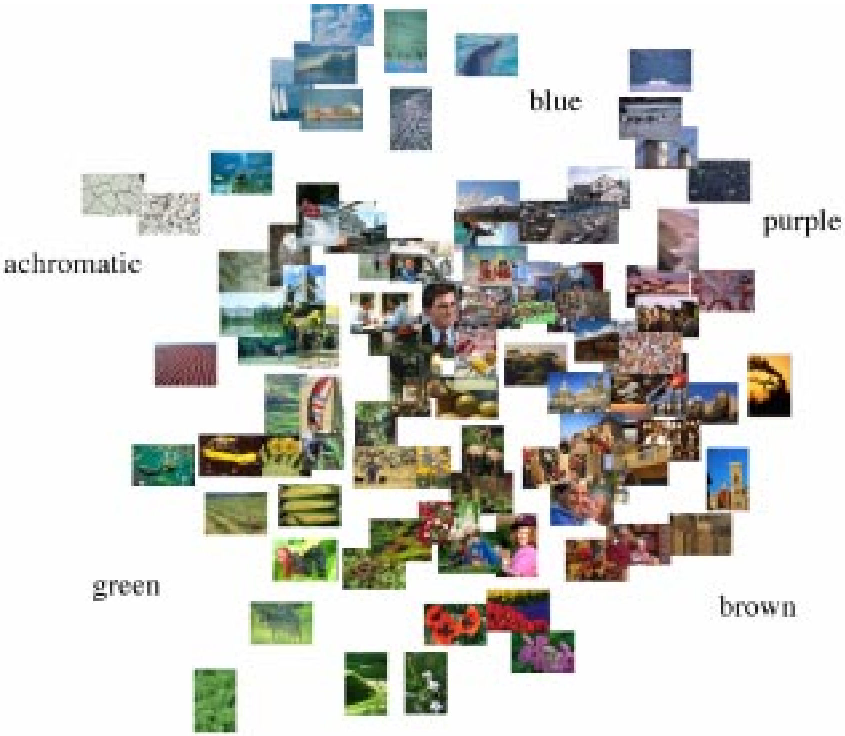


Figure 5: Multidimensional Scaling result using the color histogram distance metric

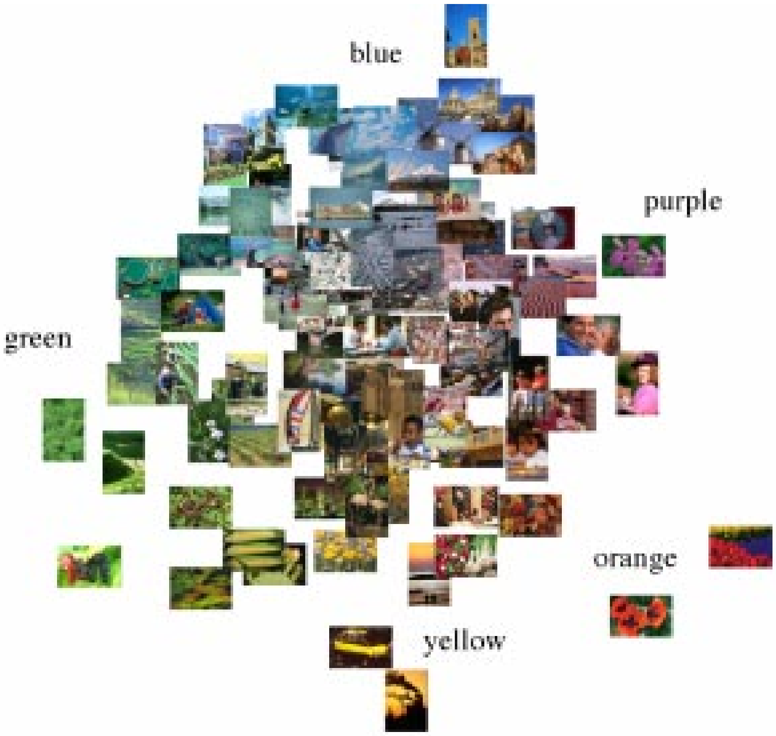


Figure 6: Multidimensional Scaling result using the perceptually optimized distance metric.

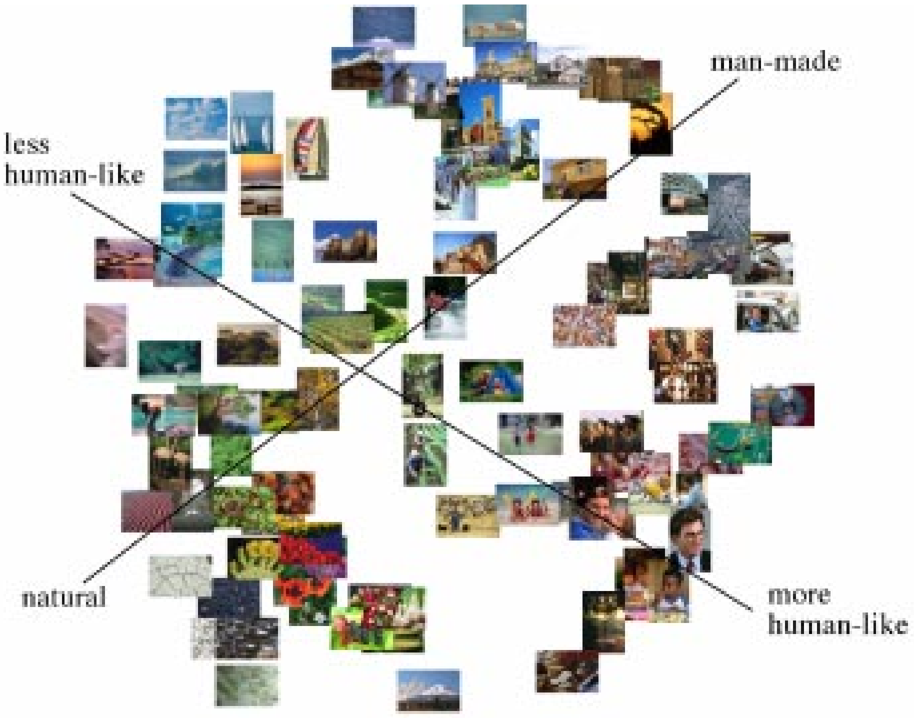


Figure 7: Multidimensional Scaling result using the image distances from the table scaling experiment

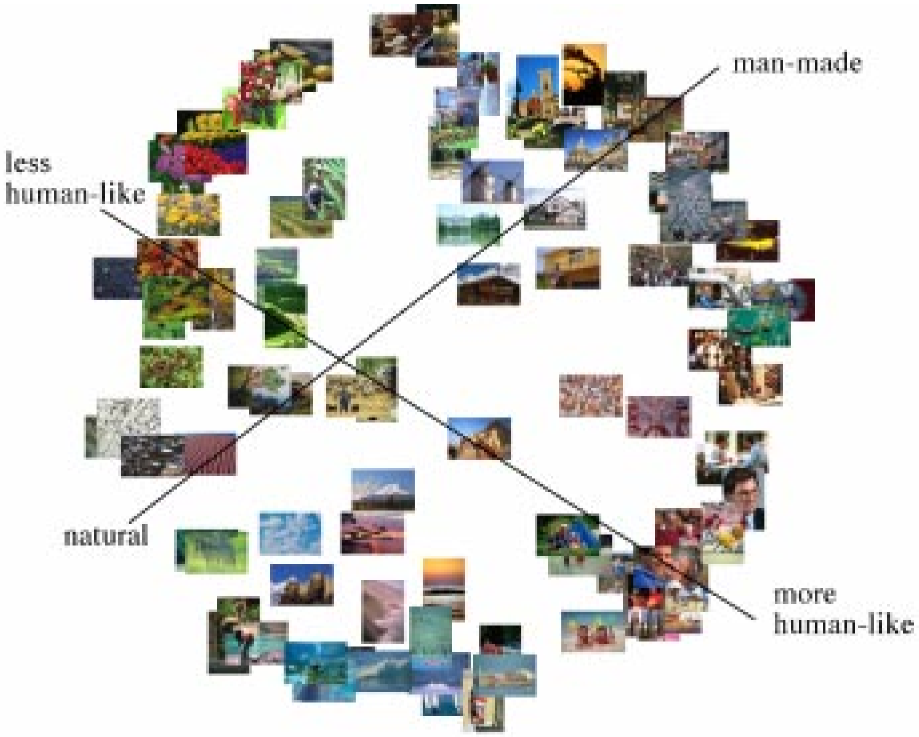


Figure 8: Multidimensional Scaling result using the image distances from the computer scaling experiment

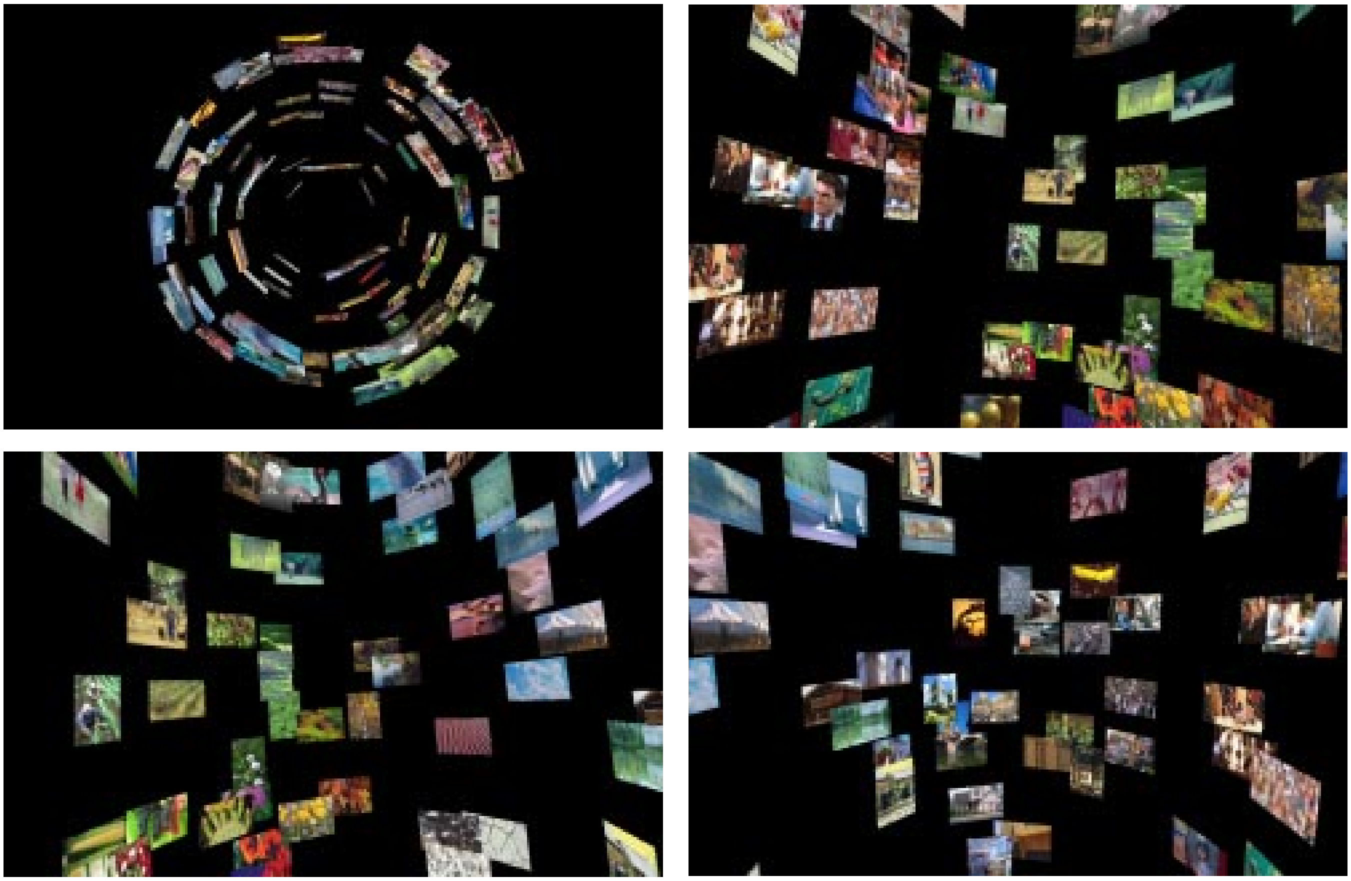


Figure 9: Multidimensional scaling results for 3 dimensions as seen in the VRML browser. The top left shows a top view of the 3-D space. The three other panels (top right; bottom left, bottom right) show three sequential views as the viewer looks out from the center of the space.

appear to be clusters of low-spatial-frequency color images and clusters of high-spatial-frequency luminance images.

Another organization, however, appears to emerge from this projection. In multidimensional scaling, the two-dimensional algorithm maps the stimuli onto a two-dimensional plane, and analysts typically try to interpret these dimensions in conceptual terms. Following this tradition, we have superimposed two candidate axes onto this image. One axis is what we call the natural vs. man-made axis, running from the rock and flower images in the bottom left through the nature scenes, to nature scenes with man-made objects, to man-made objects with nature, to the buildings and shipping docks in the upper right-hand side of the projection. The other axis is the human vs. non-human a axis, running from the less human-like images of sunsets and clouds in the top left through scenes with animals, to scenes with small images of humans in various settings to images featuring large full-face portraits of humans in the bottom right side of the projection. These candidate dimensions also seem to describe the images as we move around the obviously circular projection. Starting with natural images of seascapes, in the upper left, we move to increasingly man-made, less natural images, which do not contain humans, then move through the man-made objects, which likewise do not include humans, through to images which increasingly include humans, on to the end of the "more human axis". Continuing our path, we have humans in nature, then natural objects without humans.

Figure 8 shows the 2D result for the Perceptual Computer Scaling experiment. Again, this result supports the idea that overall color is important in image similarity judgments, and also provides support for the low-spatial-frequency color and high-spatial-frequency luminance hypothesis. There appear to be clusters within this organization which clearly reflect these features.

Perhaps more striking is the similarity in organization of these results to those of the table-sorting experiment. In this image we have superimposed onto the 2D projection the same qualitative axes as in the previous figure. Although there are differences in the fine structure of how the images are organized, the overall structure produced by the two psychophysical experiments is extremely similar. In the computer scaling experiment, there is a progression across the circular array from natural to man-made, and, in what appears to be an orthogonal dimension, an equally compelling progression from less human-like to more human-like. Around the circle, we go from nature scenes which do not include humans or man-made objects, to nature scenes which include some man-made objects, such as buildings, through to images of man-made objects without nature or humans, to images of man-made objects with humans, but not nature. From the end of the human axis, we progress to images with humans and natural scenes, and back to images with purely natural scenes without humans or man-made objects. Interestingly, in both psychophysical projections, animals are organized more closely with natural scenes than they are with humans, perhaps because in these images, all the animal images were taken in nature.

### 3.3.2 MDS in 3 dimensions

One of the most interesting features of the two dimensional projections of the psychophysical results is the circular structure. Using the Diamond visualization tool,[17] we were able to look more closely at this three-dimensional MDS structure. We found that the 3-D solution is best described as a sphere of images, where the sphere has a thickness of not more than 10 percent of the distance across the diameter. This means that the images in the central region of the 2-D projections are really images which would belong the shell of the 3D solution. To examine this further, we created a three-dimensional geometry and pasted the center point of each image onto its coordinate position in this 3D space. In order to view these images, we positioned them so that each was normal to a vertical axis in this space. For example, all the images on the top of the sphere were effectively rotated 90 degrees.

We created a VRML "world" which allowed us to view the images in this 3-D geometry from a number of different perspectives. Perhaps the most interesting view of these data was to zoom into the center of this rotating set of images, and watch the images move slowly past. Figure 9 shows a matrix of four views generated from the VRML world. The top left panel contains a view from the top, showing concentric rings of images forming a shell around the center of the space. The remaining panels contain three sequential views around the interior of the 3D shape, looking outward. The shapes of the images appear distorted because we have introduced a fisheye transformation in order to see more images. The view on the top right shows a progression from images with human figures, with children organized together, moving through to outdoor scenes with humans as small, subordinate features, to green outdoor nature scenes with animals, to outdoor, nature scenes without animals or people. A grouping of nature scenes with flowers and vegetables is at the bottom right-hand corner of the panel, and includes the snow-drop image discussed in Figure 3. It is interesting to notice that the ordering of similarity across observers in Figure 3 is well-matched to the distances in this projection.

The panel at the lower left shows another snapshot of this 3D space, the next view in the sequence. This view overlaps the view just discussed, beginning with images of humans in natural scenes. The top of the view shows nature scenes with animals on green backgrounds followed by animal images on blue ocean backgrounds (fish and birds), water scenes without people, then sea scenes with man-made objects. The right-hand side of the panel is filled with man-made objects: ships, bridges, trains and train tracks. The panel at the bottom right picks up where the previous one left off and shows the progression from ocean scenes with boats, to outdoor scenes with buildings, to man-made scenes with humans, back to where we began.

The overall impression generated by this 3D view is that the results generated by human observers are very systematic, with sensible progressions, following semantic, color, and structural characteristics.

### 3.3.3 MDS Stress

As expected, very low dimensional spaces cannot represent the full complexity of perceptual similarity judgments. This is reflected in the stress values shown in Table 1. Comparing the two psychophysical techniques, we notice that the stress in the table scaling method is lower than for the computer scaling method. This suggests that, when constrained to a 2-D projection, subjects are more likely to agree on the two most fundamental dimensions, thereby reducing the dimensionality of the data across subjects.

The similarity between the MDS solutions for the Table Scaling and Computer Scaling experiments, and their intuitive appeal, suggests that, despite this high stress, they do capture some of the most important dimensions of

| Dimensions MDS space | Computer scaling | Table Scaling | Color-histogram algorithm | Perceptually optimized algorithm |
|---|---|---|---|---|
| 2 | 0.33 | 0.25 | 0.27 | 0.26 |
| 3 | 0.23 | 0.16 | 0.17 | 0.19 |
| 4 | 0.18 | 0.10 | 0.13 | 0.14 |

Table 1: Stress values for MDS results as a function of dimensionality.

image similarity.

# 4 Discussion

Algorithmic image similarity metrics commonly make use of attributes which are thought to influence human judgments of image similarity. In this paper, we have explicitly studied human visual similarity judgments in an effort to develop similarity metrics better matched to human judgments. Using psychophysical scaling techniques, we have compared the results of several methods for measuring the similarity of images. The first is a simple color histogram algorithm, the second is a metric with parameters weighted by psychophysical judgments, and the remaining two are the results of psychophysical scaling experiments which explicitly measure human judgments of image similarity.

## 4.1 Evaluating the psychophysical results

Although the two scaling results were designed to test the visual similarity of our 97 test images, the experimental conditions were quite dissimilar. In Table Sorting, all images were in view simultaneously whereas in Computer Scaling, the observer selected the one image among a set of eight which looked most similar. In Table Scaling, the images were printed; in Computer Scaling, they were presented on a computer display. In Table Scaling, the experiment constrained the solution to the 2D tabletop, whereas the Computer Scaling experiment did not. In Table Scaling, the similarity matrix was fully populated, since a physical distance was computed between each pair of images once the solution was obtained. In Computer Scaling, the matrix was sparse. Given these differences, it is quite remarkable that the two techniques produced such similar 2D solutions when their respective similarity matrices were scaled using multidimensional scaling. This similarity suggests that the two techniques tapped the same perceptual processes, independent of differences in experimental methodology.

## 4.2 Using these Results

The systematic nature of these data make them an excellent basis for future research in image similarity. It is clear that human observers use many dimensions in their evaluations of image similarity, including color, high-frequency luminance information, low-frequency color information, and perhaps most important, semantic information. That the two- and three-dimensional solutions to the scaling judgments produce smooth, orderly transitions from image to image suggests that these cues may be smoothly combined. This suggests that it may be possible to develop descriptors which describe these smooth transitions as a function of measurable features of the images.

In the analysis of the 2D MDS projections for the psychophysical scaling experiments, we suggested the possibility of dimensions running from more- to less-human-like, and from natural to man-made. Returning to these images, it is clear that other organizations are also possible. For example, it is easy to see that the images are grouped according to semantic category: buildings, animals, boats, children, and man-made objects are organized near each other. This suggests that it might also be useful to use cluster analysis techniques.

### 4.3 Modeling

The psychological image similarity spaces we have uncovered using the multidimensional scaling techniques appear to be visually intuitive, and if this were the entire set of existing images in the universe, we might be inclined to offer our 3D VRML world as a navigation tool. Certainly images which appear similar to each other appear near each other, and it is easy to understand how to navigate within this space. However, this is not the universe of images, and in order to make these results useful, it must be possible to compute where new images should go. This means that we need to model image positions in terms of images features which can be computed automatically. To do so, our first task will be to try to model the positions of the images in terms of measurable image characteristics, then test this model with a new set of images.

Another use for these data is to test the performance of various image similarity metrics. We have seen in this study that, at least qualitatively, the color histogram and the perceptually optimized metric capture some aspects of the human judgment data, but do not adequately model the human similarity space. At minimum, an error measure could be devised comparing the results of the proposed metric with the perceptual results. As more is learned about the visual mechanisms underlying these results, however, it is hoped that a more theoretical, quantitative model could be developed which could describe the deficiencies of image similarity metrics in terms of visual processes.

### 4.4 Image Semantics

It is quite clear that these results suggest semantic categories. It is easy to see semantic clusters in the data, of people, outdoor scenes, seashore scenes, etc. These semantic categories, however, appear to correlate with image descriptors. For example, images with indoor scenes tend to be brownish, have low light levels, and many straight edges. One idea for getting a handle on these semantic categories is to explore how much of the information contained in these semantic categories can be accounted for in terms of calculable image processing descriptors and image statistics.

### 4.5 Why are the MDS results so circular/spherical?

The circular structure of the computer-scaling result could reflect the sparsity of the similarity matrix. In particular, the distance matrix contains a considerable number of entries equal to one. The optimal geometric solution to preserve these distances in two dimensions is a circular structure. This explanation, however, does not explain the circular structure of the table-scaling result, since that matrix was fully populated. The circular structure might disappear if a non-metric MDS algorithm is used instead of the metric algorithm discussed above.

## 5 Conclusions

We have conducted two psychophysical scaling experiments aimed at uncovering the dimensions human observers use in rating the similarity of photographic images, and have compared the results with two algorithmic image similarity methods. Although these experiments were conducted with different media, different tasks, and different methodologies, they produced very similar multidimensional scaling results. First, the overall color of the images was an important factor in judging similarity, and the dimensions "Human vs. non-human" and "natural vs. manmade" were very salient. These low-dimensional solutions did not capture all the richness in these multidimensional judgments, as reflected in the overall stress of the multidimensional analysis, but they did provide systematic structures with relatively smooth transitions and intuitive organizations. These features encourage us to use these results as a basis for developing perceptually-based image similarity metrics and intuitive navigation environments.

## ACKNOWLEDGMENT

This work was supported by an IBM Graduate Student Fellowship.

# 6 REFERENCES

[1] P. Aigrain, H. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: A state of the art review," *Multimedia Tools and Applications*, vol. 3, no. 3, pp. 179–202, November 1996.

[2] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11 – 32, November 1991.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23 – 32, September 1995.

[4] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233 – 254, June 1996.

[5] J. Y. Chen, C. A. Bouman, and J. P. Allebach, "Multiscale branch and bound image database search," *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases V*, vol. 3022, February 13 - 14 1997, San Jose, CA, pp. 133 – 144.

[6] T. Frese, C. A. Bouman, and J. P. Allebach, "A methodology for designing image similarity metrics based on human visual system models," *Proc. of SPIE/IS&T Conf. on Human Vision and Electronic Imaging II*, vol. 3016, February 10 - 13 1997, San Jose, CA, pp. 472 – 483.

[7] B. Scassellati, S. Alexopoulos, and M. Flickner, "Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments," *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases*, vol. 2185, February 7-8 1994, San Jose, CA, pp. 2 – 14.

[8] Y. Rubner, L. J. Guibas, and C. Tomasi, "The earth mover's distance, multidimensional scaling, and color-based image retrieval," *Proceedings of the ARPA Image Understanding Workshop*, May 1997, New Orleans, LA, pp. 661–668.

[9] J. MacCuish, A. McPherson, J. E. Barros, and P. M. Kelly, "Interactive layout mechanisms for image database retrieval," *Proc. of SPIE Conf. on Visual Data Exploration and Analysis III*, vol. 2656, Jan. 31 - Feb. 2 1996, San Jose, CA, pp. 104 – 115.

[10] R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 3, pp. 1317 – 1323, September 1987.

[11] J. B. Kruskal, *Multidimensional scaling.* Beverly Hills, CA: Sage Publications, 1978.

[12] C. L. Krumhansl, "The psychological representation of musical pitch in a tonal context," *Cognitive Psychology*, vol. 11, no. 3, pp. 346–374, July 1979.

[13] C. R. Cavonius, M. Mueller, and J. D. Mollon, "Difficulties faced by color-anomalous observers in interpreting color displays," *Proc. of SPIE Conf. on Perceiving, Measuring, and Using Color* (M. H. Brill, ed.), vol. 1250, 1990, Santa Clara, CA, pp. 190 – 195.

[14] A. R. Rao and G. L. Lohse, "Identifying high level features of texture perception," *CVGIP:Graphical Models and Image Proc.*, vol. 55, no. 3, pp. 218 – 233, May 1993.

[15] M. Vanrell, J. Vitrià, and X. Roca, "A multidimensional scaling approach to explore the behavior of a texture perception algorithm," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 262 – 271, 1997.

[16] T. F. Cox and M. A. A. Cox, *Multidimensional scaling.* Monographs on statistics and applied probability, London: Chapman & Hall, 1st ed., 1994.

[17] D. A. Rabenhorst, J. A. Gerth, and C. N. Mills, *BMDP/DIAMOND for Windows: User's Guide.* Los Angeles, CA: BMDP Statistical Software, Inc., 1995. Currently available through SPSS, Inc., Chicago, IL.