The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval

Yossi Rubner, Leonidas Guibas, Carlo Tomasi *
Computer Science Department, Stanford University
Stanford, CA 94305
[rubner,guibas,tomasi]@cs.stanford.edu

Abstract

In this paper we present a novel approach to the problem of navigating through a database of color images. We consider the images as points in a metric space in which we wish to move around so as to locate image neighborhoods of interest, based on color information. The data base images are mapped to distributions in color space, these distributions are appropriately compressed, and then the distances between all pairs I, J of images are computed based on the work needed to rearrange the mass in the compressed distribution representing I to that of J. We also propose the use of multi-dimensional scaling (MDS) techniques to embed a group of images as points in a two- or three-dimensional Euclidean space so that their distances are preserved as much as possible. Such geometric embeddings allow the user to perceive the dominant axes of variation in the displayed image group. In particular, displays of 2-d MDS embeddings can be used to organize and refine the results of a nearest-neighbor query in a perceptually intuitive way. By iterating this process, the user is able to quickly navigate to the portion of the image space of interest.

1 Introduction

Rummaging through a large catalog of pictures in search of a particular image is unrewarding and time-consuming. Image database retrieval research [Bach *et al.*, 1996, Guibas and Tomasi, 1996, Niblack *et al.*, 1993, Pentland *et al.*, 1996] attempts to automate parts of this task. The most popular proposals for formulating a query into an image database is to sketch the desired picture or to provide an example of a similar image. Yet often we do not know the precise appearance of the desired image(s). We may want a sunset, but we do not know if sunsets in the database are on beaches or against a city skyline. When looking for unknown images, browsing, not query, is the preferred search mode. And the key requirement for browsing is that similar images are located nearby. Current retrieval systems list output images in order of increasing distance from the query. However, the distances among the returned images also convey useful information during browsing. In this paper, we present a novel framework for computing the distance between images, and a set of tools to visualize an entire image data base or parts of it during browsing.

The question of image similarity is complex and delicate. Semantic similarity (two images with cats are similar to each other) is still out of the question, and we must make do with similarity of appearance. More specifically, in this paper we focus on the overall color content of an image as the main criterion for similarity. The overall distribution of colors within an image contributes to the mood of the image in an important way, and is a useful clue for the image's contents. Sunny mountain landscapes, sunsets, cities, faces, jungles, candy, and fire fighters scenes lead to images that have different but characteristic color distributions. If the pictures in a database can be arranged in a geometric space so that their locations reflect differences and similarities in their color distributions, browsing the database becomes intuitively meaningful. In fact, the database is now endowed with a metric structure, and can be explored with a sense of continuity and comprehensiveness: all we care, as far as the parts of the database that have undesired color distributions are concerned, is that we need not traverse them. On the other hand, interesting regions can be explored with a sense of getting closer or farther away from the desired distribution of colors. In summary, the user can form

*This work was sponsored by the Defense Advanced Research Projects Agency under contract DAAH04-94-G-0284 monitored by the US Army Research Office. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the United States Government, or Stanford University.

a mental, low-detail picture of the entire database, and a more detailed picture of the more interesting parts of it. If a picture is worth a thousand words, a picture of an image database is worth a whole book.

Of course, arrangement criteria other than color distribution are possible. For instance, information about the position of colors in the images, as well as shape and texture, ought to be considered eventually. However, color distribution is at the same time useful in its own right and complex enough to let us illustrate the main issues. Thus, while we experiment with the notion of similarity in the context of color information, we define a framework in which shape and texture descriptors can also be accommodated, leading to a skeletal theory of image database visualization. In particular, we address the following questions:

- How do we summarize the color distribution of an image?
- When do two images have similar color distributions and, more generally, how do we measure the 'distance' between these distributions?
- How can we arrange a collection of images so that similar images are near each other?

Summarization of color distribution has to do with perceptual significance, invariance, and efficiency at the same time. Colors should be represented in a way that reflects a human's appreciation of similarities and differences. At the same time, the distribution of colors in an image should be represented by a collection of data that is small, for efficiency, but rich enough to reproduce the essential information. The issue of relevance to human perception has been resolved by the definition of appropriate color representations, among which we choose the CIE-LAB standard. Section 2 addresses the issue of summarization by presenting a new, efficient clustering scheme based on *k-d* trees. This scheme buys efficiency at the expense of reduced guarantees about the size of the output. While more expensive algorithms may guarantee a minimal number of clusters, this is an unnecessary requirement for our application. The result of this method is a small collection of (weighted) points in color space which represent well the full distribution; we call this set of points a (color) *signature*. Section 3 introduces the *Earth Mover's Distance* (EMD) [Stolfi, 1994] as a useful and flexible measure of distance between signatures, and presents an efficient algorithm for its computation based on linear programming. This distance endows the image database with an appropriate metric, thereby addressing the question of image similarity. Section 4 addresses the third question above, and shows how to use the technique of *Multi-Dimensional Scaling* (MDS) [Kruskal, 1964] in order to visualize either the entire database or just the part of it returned in response to a query in a two- or three-dimensional space. The resulting composite image properly reflects the distribution of color distributions within the database. Finally, section 5 argues that the techniques and issues introduced in this paper generalize to other aspects of image description.

2 Color Signatures

The color information of each image is reduced to a compact representation that we call the *signature* of the image. In general a signature contains a varying number of points in a Euclidean space where a weight is attached to each point. In the case of color images, the points represent clusters of similar colors and the weight of a point is the fraction of the image area with that color.

To compute the signature of a color image, we first slightly smooth each band of the image's RGB representation in order to reduce possible color quantization and dithering artifacts. We then transform the image into the CIE-LAB color space [Wyszecki and Styles, 1982] using D65 as the reference white. This nonlinear transformation deforms the RGB color space so that the resulting Euclidean distance between color coordinates approximates how well colors are discriminated by humans.

Each image implies a distribution of points in the three-dimensional CIE-LAB color space where a point corresponds to a pixel in the image. We coalesce this distribution into clusters of similar colors. We define these as clusters that do not exceed 30 units in any of the L, a, b axes. Because of the large number of images to be processed in a typical database, clustering must be performed efficiently. To this end, we devised a novel two-stage algorithm based on a *k-d* tree [Bentley, 1975]. In the first phase, approximate clusters are found by a balanced partition of color space through a *k-d* tree. Subdivision stops when a cell becomes smaller than the allowed cluster size. This process can result in excessive subdivision. The second phase then tries to merge close clusters computed in the first phase by performing a second *k-d* tree clustering on points which represent the centroids of the clusters that are produced in the first phase, after shifting the space coordinates by one half of the minimal allowed cell size. Each cluster contributes a pair (p, w_p) to the signature representation of the image where p is the average color of the cluster and w_p is its weight which is the fraction of image pixels that are in that cluster. Figure 2 shows examples of color signatures for three images.

The signatures thus obtained are compact: the color distribution of an entire image is summarized by a handful of points, typically eight to twelve. Because of the clustering algorithm used, signatures represent well the image's overall color distribution. Since signatures represent distributions in the CIE-LAB color space, they are perceptually significant, in that Euclidean distances between points are strongly correlated with perceptual differences. Because of clustering, small variations in the colors of an image have little effect on signatures, thereby providing a moderate degree of invariance to changes of viewpoint and lighting. Finally, signatures are simple and flexible abstractions for which we can define meaningful metrics, as shown in the following section.

3 Distance Between Color Signatures

In image retrieval, it is important to define a similarity measure between two color distributions or, in particular, between two color signatures. When considering only the color content of images, and ignoring the actual positions of the pixels within the image, this problem is known as the color indexing problem which was introduced by Swain and Ballard [Swain and Ballard, 1991] and was approached in several ways by others [Hafner *et al.*, 1995, Stricker and Orengo, 1995, Werman *et al.*, 1985]. Our approach is closest to, but more general and at the same time more efficient than that of [Werman *et al.*, 1985]. The other methods are bound to retrieve false positives [Stricker and Orengo, 1995]. We define the distance between two signatures to be the minimum amount of 'work' needed to transform one signature into the other (figure 1). The work needed to move a point, or a fraction of a point, to a new location is the portion of the weight being moved, multiplied by the Euclidean distance between the old and the new locations. When changing one signature to another, the work is the sum of the work done by moving the weights of the individual points of the source signature to those of the destination signature. We allow the weight of a single source signature point to be partitioned among several destination signature points, and vice versa. We call this distance function the *earth mover's distance*. This is a name suggested by Stolfi [Stolfi, 1994], by analogy with some CAD programs for road design which have a function that computes the optimum earth displacement from roadcuts to roadfills. As compared with the match distance of [Werman *et al.*, 1985], our distance is more general because it allows fractional/partial matches. Furthermore, it can be computed much more efficiently, as we now show.

The earth mover's distance computation can be formalized as the following linear programming problem: Given two signatures: $\mathbf{p} = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$ and $\mathbf{q} = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$ where p_i and q_j are points in some Euclidean space, the CIE-LAB color space in our case, and w_{p_i}, w_{q_j} are the corresponding weights of the points, find an $m \times n$ cost matrix $\mathbf{C}$ where C_{ij} is the amount of weight op p_i matched to q_j, that will minimize the function:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} \, \|p_i - q_j\|$$

($\|\cdot\|$ is the Euclidean distance) subject to the following constraints:

$$\begin{aligned}
C_{ij} &\geq 0 \qquad 1 \leq i \leq m \,,\; 1 \leq j \leq n & (1)\\
\sum_{i=1}^{m} C_{ij} &\leq w_{q_j} \qquad 1 \leq j \leq n & (2)\\
\sum_{j=1}^{n} C_{ij} &\leq w_{p_i} \qquad 1 \leq i \leq m & (3)\\
\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} &= \min(w_{\mathbf{p}}, w_{\mathbf{q}}) \,. & (4)
\end{aligned}$$

where $w_{\mathbf{p}} = \sum_{i=1}^{m} w_{p_i}$ and $w_{\mathbf{q}} = \sum_{j=1}^{n} w_{q_i}$. The earth mover's distance is defined as the normalized distance between points $\mathbf{p}$ and $\mathbf{q}$:

$$\begin{aligned}
\mathrm{EMD}(\mathbf{p}, \mathbf{q}) &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} \, \|p_i - q_j\|}{\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij}} \\
&= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} \, \|p_i - q_j\|}{\min(w_{\mathbf{p}}, w_{\mathbf{q}})} \,.
\end{aligned}$$

Constraint 1 allows only for positive amounts of 'earth' to be moved. Constraints 2 and 3 limit the capacity of 'earth' a point can contribute to the weight of the point. Constraint 4 forces at least one of the signatures to use all of its capacity, otherwise a trivial solution is not to move any 'earth' at all.

The earth mover's distance has many desirable properties relevant to our application. As long as the total weight of each of our signatures is the same, the earth mover's distance is symmetric and satisfies

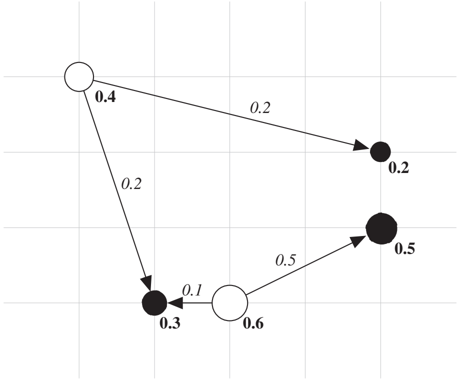


Figure 1: *The earth mover's distance in 2D between a signature with three points (black) and one with two (white). Bold and italic numbers are the weights of the points and the weights moved between points, respectively.*

the triangle inequality — thus we really work with a metric space. The 'optimal assignment' problem which the earth mover's distance computes also gives us a way to 'morph', or continuously transform, two distributions into each other: simply imagine the appropriate weight fractions moving at constant rates along the segments joining the corresponding source and destination points in color space. During the morph the centroid of the morphing distribution will move continuously from the centroid of the source to that of the destination signature. This shows that the distance between the centroids of the two signatures involved, is a lower bound on the earth mover's distance: Assuming that $w_{\mathbf{p}} = w_{\mathbf{q}} = w$ then

$$\mathrm{EMD}(\mathbf{p}, \mathbf{q}) \geq ||\mathbf{p}_{ave} - \mathbf{q}_{ave}||$$

where

$$\mathbf{p}_{ave} = \frac{1}{w} \sum_{i=1}^{m} w_{p_i} p_i \qquad \mathbf{q}_{ave} = \frac{1}{w} \sum_{j=1}^{n} w_{q_j} q_j \ .$$

This is useful for quickly recognizing dissimilar distributions.

Notice that in our formulation we do allow the total weights of the two signatures to be different. This is useful for content-based image retrieval systems for example, when a color query specifies only a part of the wanted color distribution, leaving the rest as "don't care". In this case, of course, the EMD is not a true distance and the lower bound we show does not hold.

The earth mover's distances between the images in figure 2 can be summarized by the following symmetric distance matrix:

$$\begin{bmatrix} 0 & 19.46 & 71.94 \\ 19.46 & 0 & 60.03 \\ 71.94 & 60.03 & 0 \end{bmatrix} .$$

As expected, the first two images are relatively close since they contain similar colors (blues and greens). The third image is relatively far from the first two but somewhat closer to the second image because the colors of the house and the trees in the second image are similar to the colors of the sunset in the third image.

4 Database Visualization

A metric for color signatures is crucial for image retrieval, because it quantifies the intuitive notion of image similarity. If the metric corresponds to perceptual similarity, retrieving images in response to a given query amounts to returning images whose distance from the query is small in the space of color signatures. While the earth mover's distance is indeed at the core of our image retrieval system, and has proven very effective, in this paper we want to emphasize a related but distinct use of the signature metric defined in the previous section. When browsing an image database, we often have only a vague idea of what our target images look like. This is especially true when we have not seen the images in the database beforehand. The standard format of interaction with the database, that is, iterations of a query answered by the presentation of a list of images, is not satisfactory in this case. First, one would like to have a global view of the returned images. As figure 3 (a) shows, images in the returned list can be related to one another and yet appear at separate places in the list. The returned images should be displayed not only in order of their distance from the query, but also arranged according to their mutual distances. In brief, the user of the system would benefit from a more coherent view of the query results.

Second, browsing and navigating in a large database is disorienting unless the user can form a mental picture of the entire database. Only having an idea of the surroundings can offer an indication of where to go next. The wider the horizon, the more secure navigation will be. How can such a global picture of an image database be created? Signatures offer once again a solution. Our earth mover's distance quantifies the perceptual difference that separates two signatures. Consequently, each signature can be represented by a single point in a suitably high-dimensional space, such that distances between these points are equal to the earth mover's distances between the corresponding signatures. The computation of the coordinates of these high-dimensional points is called an *embedding*. However, humans can only visualize low-dimensional spaces, typically in two or three dimensions. We then look for an approximate embedding, rather than for an exact one.

The approximate embedding problem was formalized by Kruskal [Kruskal, 1964] into the so-called Multi-Dimensional Scaling (MDS) problem. Given a set of n objects together with the matrix of distances δ_{ij} between them, and given a (small) dimension d, the problem is to find a set of n points in d-dimensional space whose distances $\{\hat{\delta}_{ij}\}$ are as close as possible to the original distances $\{\delta_{ij}\}$. The choice of closeness that was suggested by Kruskal is to minimize:

$$\text{STRESS} = \left[\frac{\sum_{i,j} (\hat{\delta}_{ij} - \delta_{ij})^2}{\sum_{i,j} \delta_{ij}^2} \right]^{1/2} .$$

Rigid transformations and reflections can be applied to the MDS result without changing the STRESS. Using MDS can assist navigation in the space of images both locally and globally, as we now illustrate.

4.1 Local MDS

Performing MDS on the images returned from a query gives us a better way to display the query results. Instead of the traditional one-dimensional list of images sorted by their distances from the query, we can display a two or three dimensional map of the images, where each image is positioned according to the MDS result. In this way we are presenting information reflecting $\binom{n}{2}$ distances, instead of only n in the traditional method. In addition to visually representing the relative distances between *all* pairs of images, images with similar color content tend to group together. Figure 3 shows the result from a sample query into our image retrieval system. The query asked for images with 20% blue and 80% don't care and requested only the ten best matching images. Figure 3(a) shows the traditional way of displaying the resulting images as a one-dimensional list sorted by the distances from the query, while figure 3(b) shows the same images arranged according to a two-dimensional MDS. In the MDS display, similar images of desert scenes with yellowish ground group together at the top left, images with green plants group at the bottom, and the two other images – a desert image with a white ground and an image of a statue, are to the right. An all-blue image is comparatively dissimilar from the others, and is accordingly relegated to the far right. In this iterated-query framework, navigation can proceed by choosing a promising area in the MDS display and using a representative image out of that area as the next query.

4.2 Global MDS

Performing MDS on a large set of images can help the user understand the space of color images of the set. In figure 4 we see the MDS map of 500 images. It is easy to see that images group by their average chroma. For example, blue images are at the top-left, green images are at the top-middle, yellow images are at the top-right, and so forth. The images are also ordered from bottom-right to top-left by their average lightness, dark images are at the bottom-right and bright images are at the top-left. Higher dimensional MDS can be done on the image database where different characteristics of the images will be revealed, such as their average chroma (the projection of the images on the appropriate axes gives the chromaticity diagram), average lightness, the colorfulness of the images, and so forth. Now when we look for a sunset we see immediately where to go. At a glance, we can write off most of the database, and home in to the "sunset-looking" part of it. At the same time, we form a mental picture of the entire database. We see everything in coarse detail, and we have the impression of grasping the overall database content, at least in terms of color distributions. Given a joystick that lets us get closer to the area of interest, we have at the same time focus, because nearby images are large on the display, and context, because all or most other images are still visible at a distance. As we move about, we have the comforting impression that the whole database is there all the time, rather than being handed down to us in small fragments.

5 Conclusions

The methods presented in this paper open a novel set of tools and possibilities for image data-base navigation and visualization. The color signatures we have defined and the earth mover's distance between them seem to capture well the perceptual similarity or dissimilarity of images based on their color content. Furthermore, the low-dimensional geometric embeddings we compute using MDS techniques provide an intuitive way for the user to refine his/her query and to continue exploring interesting neighborhoods of the image space — or to see large portions of it all at once.

All image query system are ultimately based on computational approximations to perceptual image distance — approximations whose quality we are often asked to take for granted. Our approach appears to be the first one to allow the user to explore, in an intuitive way, the area of the image space beyond what the system considers the neighborhood of the query. Such an exploration can provide increased confidence that what is wanted will not be missed.

Clearly much remains to be done. It is likely that distances between similar images provide much more information than distances between images which have little in common. Yet currently we compute large distances as accurately as small ones. As indicated in Section 3, we can gain significant speedups by simply using lower bounds for the earth mover's distance when the corresponding images are far apart. A major extension of our work will be to apply the concepts of signature and the earth mover's distance to other modalities which also convey information about the content of the image, such as shape and texture. The principle of our approach will remain that we measure the distance between images by the minimum 'work' needed to make their signatures the same. Thus the data in a signature need not be fully homogeneous, as long as we provide a set of modification operations, with associated costs, for each type of data present. We consider this ability to combine different kinds of feature sets and modalities (both in building the image database index and in computing the appropriate geometric embeddings) to be a unique advantage of our approach.

For the intuitive use of the geometric embeddings computed by MDS methods, it is crucial that the 'axes of variation' be perceptually clear to the user. This worked well for us in the case of color, in part because we started from data in a geometric color space whose axes have a familiar significance. Getting the same effect in the case of shape and texture seems more of a challenge. We intend to explore how to 'advise' MDS algorithms about what are desired coordinate axes to use. We also need to study more the relations between the axes chosen by MDS for related or overlapping image sets. Knowing the correspondence between these 'local charts' (in the sense of topology) of the image space can greatly help in providing a globally stable and consistent sense of navigation.

Acknowledgments: The authors wish to acknowledge helpful discussions with Scott Cohen and Jorge Stolfi.

References

[Bach *et al.*, 1996] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu. Virage image search engine: an open framework for image management. In *SPIE Conference on Storage and Retrieval for Image and Video Databases IV*, volume 2670, pages 76–87, March 1996.

[Bentley, 1975] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18:509–517, 1975.

[Guibas and Tomasi, 1996] L. J. Guibas and C. Tomasi. Image retrieval and robot vision research at Stanford. In *Proceedings of the ARPA Image Understanding Workshop*, pages 101–108, Palm Springs, CA, 1996.

[Hafner *et al.*, 1995] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–735, July 1995.

[Kruskal, 1964] J. B. Kruskal. Multi-dimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[Niblack *et al.*, 1993] W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, and Y. Heights. Querying images by content, using color, texture, and shape. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume 1908, pages 173–187, April 1993.

[Pentland *et al.*, 1996] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, June 1996.

[Stolfi, 1994] J. Stolfi. Personal communication, 1994.

[Stricker and Orengo, 1995] M. Stricker and M. Orengo. Similarity of color images. In *SPIE Conference on Storage and Retrieval for Image and Video Databases III*, volume 2420, pages 381–392, February 1995.

[Swain and Ballard, 1991] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[Werman *et al.*, 1985] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multi-dimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32:328–336, 1985.

[Wyszecki and Styles, 1982] G. Wyszecki and W. S. Styles. *Color Science: Concepts and Methods, Quantitative Data and Formulae.* John Wiley and Sons, New York, NY, 1982.

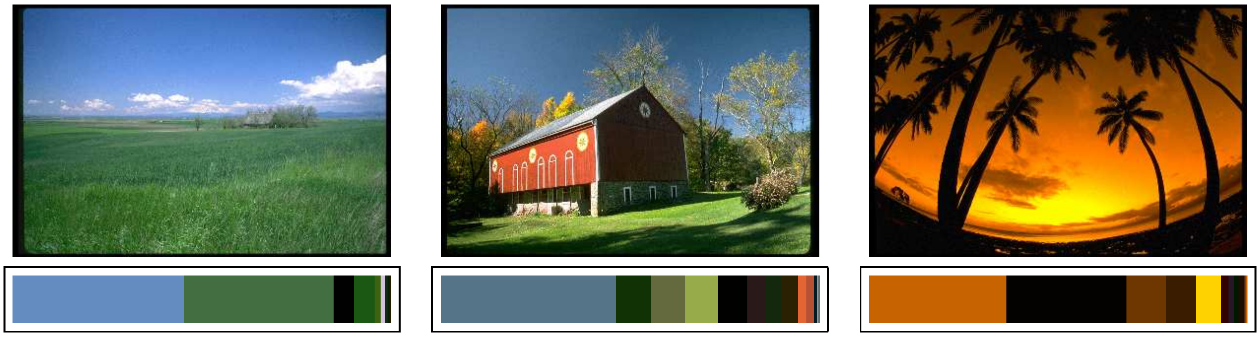
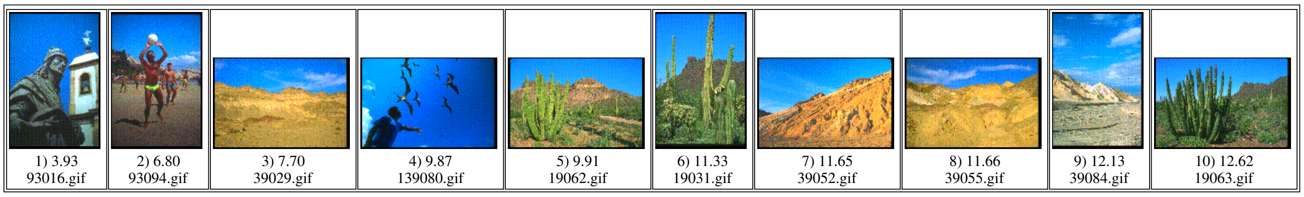


Figure 2: Three (color) images together with their color signatures. The left image contains mostly greens and blues, the middle image contains mostly greens, blues and browns, and the right image contains mostly yellows, browns and blacks.



(a)

(b)

Figure 3: The top ten images for a query that asked for 20% blue and 80% don't care. (a) Traditional display. (b) MDS map.

Figure 4: 2D MDS map of 500 images.