Proceedings of the 1998 IEEE Asian Conference
on Computer Vision, Hong Kong

Adaptive Color-Image Embeddings for Database Navigation *

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas

Computer Science Department, Stanford University, Stanford, CA 94305, USA

Abstract. We present a novel approach to the problem of navigating through a database of color images for the purpose of image retrieval. We endow the database with a metric for the color distributions of the images. We then use multi-dimensional scaling techniques to embed a group of images as points in a two-dimensional Euclidean space so that their distances reflect image dissimilarities as well as possible. Such geometric embeddings allow the user to perceive the dominant axes of variation in the displayed image group, and form a mental picture of the database contents. Furthermore, since these embeddings group similar images together, away from dissimilar ones, the user can refine the query in a perceptually intuitive way. By iterating this process, the user can quickly navigate to the portion of the image space of interest.

1 Introduction

The user of an image retrieval system would typically like to specify queries in semantic terms (e.g. "children playing in a park"). Unfortunately, the state-of-art in computer vision does not yet allow for such queries. Instead, systems use simpler syntactic image features such as color, texture and shape [2, 1, 3, 6, 7], in the hope that these correlate well with semantic features. This discrepancy between syntactic and semantic queries causes a basic problem with the traditional query/response style of interaction. An overly generic query yields a large jumble of images, which are hard to examine, while an excessively specific query may cause many good images to be overlooked by the system. This is the traditional trade-off between good precision (few false positives) and good recall (few false negatives). Striving for both good precision and good recall may pose an excessive burden on the definition of a "correct" measure of image similarity. While most image retrieval systems, including the ones above, recognize this and allow for an iterative refinement of queries, the number of images returned for each query is usually kept low so that the user can examine them one at a time.

In contrast, we propose that with an appropriate display technique, which is the main point of this paper, many more images can be returned without overloading the user's attention. Specifically, if images can be arranged on the screen so as to reflect similarities and differences between their color distributions, the initial queries can be very generic, and return a large number of images. The

* Research supported by grants DARPA DAAH04-94-G-0284 and NSF IRI-9712833.

consequent low initial precision is an advantage rather than a weakness. In fact, the user can see large portions of the database at a glance, and form a global mental model of what is in it. Rather than following a thin path of images from query to query, as in the traditional approach, the user now *zooms in* to the images of interest. Precision is added incrementally in subsequent query refinements, and fewer and fewer images are displayed as the desired images are approached.

In our system, we use the distributions of colors in images as our retrieval features. These have been shown [12, 2, 11, 1, 3, 6, 7] to be useful retrieval cues. When a (usually vague) query is specified or drawn by the user, we locate and display a large number of neighboring images in the database. Since queries in our system are image-like, neighborhood can be defined in terms of the distance between images. The resulting images are then used for more focused queries that return fewer and fewer images. At every step, query results are embedded in two-dimensional space by using *multi-dimensional scaling (MDS)* [10, 4], by which we place picture thumbnails on the screen so that screen distances reflect as closely as possible the distances between the images. While more traditional displays list images in order of similarity to the query, thereby representing n distances if n images are returned, our display conveys information about all the $\binom{n}{2}$ distances between images. As shown by the examples in this paper, this display makes it easy for the user to grasp the entire set of returned images at a glance, understand how the query actually performed, and decide where to go next. In fact, such geometric embeddings allow the user to perceive the dominant axes of variation in the displayed image group. When the user selects a region of interest on the display, a new, more specific query is automatically generated, and returns a smaller set of images. These are again displayed by a new MDS, which now reflects the new dominant axes of variation. Thus, the embeddings are *adaptive*, in the sense that they use the screen's real estate to emphasize whatever happen to be the main differences and similarities among the particular images at hand. By iterating this process, the user is able to quickly navigate to the portion of the image space of interest, typically in very few mouse clicks.

In the next section, we introduce the data structures we use to summarize the color content of images, and a distance measure between them. Section 3 then describes the visualization technique, and Section 4 shows its use for database navigation. Section 5 argues that MDS image embeddings can be usefully applied to other modalities besides color and discusses topics for future work.

2 A Metric for Color Images

This section describes the color signature as our basic representation for color distributions, and introduces the earth mover's distance as a measure of distance between signatures. More details on these concepts can be found in [8].

2.1 Color Signatures

The color information of each image is reduced to a compact representation that we call the *color signature* of the image. A color signature contains a varying number of points, each representing a cluster of similar colors in the CIE-Lab color space [14]. The number of points in a signature varies with the color complexity of the image. A weight describing the fraction of the image area with that color is attached to each point.

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. This nonlinear transformation deforms the RGB color space so that Euclidean distance between nearby color coordinates approximates how well colors are discriminated by humans. We coalesce the three-dimensional CIE-Lab distribution of colors in the image into clusters by the algorithm described in [8]. In our database[1], a signature contains eight color clusters on average.

2.2 Distance Between Color Signatures

In order to define a similarity measure between two color signatures, we introduce the notion of the *Earth Mover's Distance (EMD).* This is the minimal amount of 'work' needed to transform one signature into the other, in the following sense. 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 two 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. Although we do not claim that the EMD is a perceptual distance, it is an extension of distances of single colors in the CIE-Lab color space, which are perceptual distances, to distances between sets of colors. In practice, as we show in the following sections, the EMD leads to good results. It also allows for partial matches like "give me images with 20% orange and I don't care about the remaining 80%". The ability to use partial matches is important for navigation as we shall see in section 4. See [8] for more details about the EMD properties, and for efficient algorithms for its computation.

3 Database Visualization

While the EMD 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 this metric. Once image retrieval systems find the best matches for a given query, they usually display them in a list, sorted by their similarity to the query. While this might suffice if the image we are after is in that list, this is usually not the case, especially when we have only a vague idea of what we are looking for. At this point it is desirable to display a coherent view of the query results where the returned

[1] Our database contains a rather diverse set of 20,000 color images from the Corel Stock Photo Library.

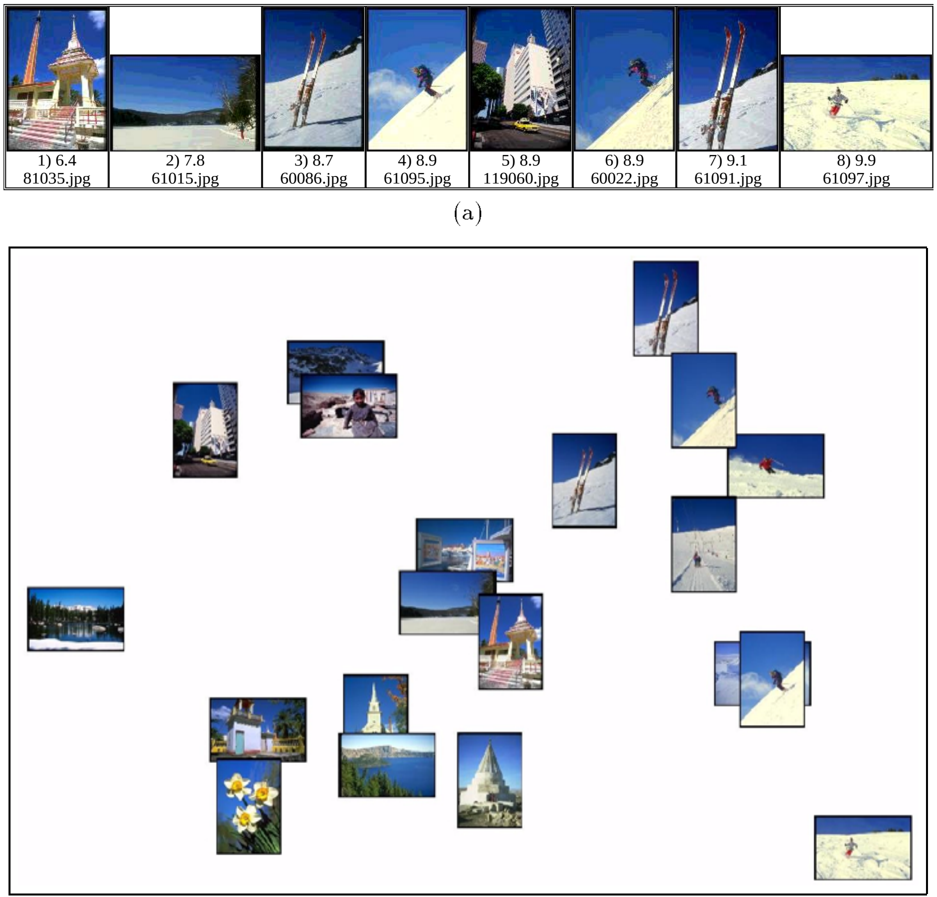
images should be displayed not only in order of their distance from the query, but also arranged according to their mutual distances. With such view, the user can see the relations between the images, better understand how the query performed, and be guided to successive queries. How can such a global picture of part of an image database be created? Our EMD approximates 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 EMDs 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 an exact one, in two dimensions. Such approximate embeddings are an instance of the problem of multi-dimensional scaling (MDS).

Given a set of n objects together with the dissimilarities δ_{ij} between them, the MDS technique [10,4] computes a configuration of points $\{p_i\}$ in a low-dimensional Euclidean space $\mathbf{R}^d$, (we use $d = 2$) so that the Euclidean distances $d_{ij} = \|p_i - p_j\|$ between the points in $\mathbf{R}^d$ match as well as possible the original dissimilarities δ_{ij} between the corresponding objects. Kruskal's [4] formulation of this problem requires minimizing the following quantity

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

with the additional constraint that the d_{ij}s are in the same rank ordering as the corresponding δ_{ij}s. STRESS is a nonnegative number that indicates how well distances are preserved in the embedding. Zero STRESS indicates a perfect fit. Rigid transformations and reflections can be applied to the MDS result without changing the STRESS. Embedding methods such as SVD and PCA are not appropriate here because our signatures do not form a linear space, and we do not have the actual points, only the non-Euclidean distances between them. In our system we used the ALSCAL MDS program [13].

An example is shown in Figure 1. Suppose that we are looking for images of skiers. These images can be characterized by blue skies and white snow, so we use as our query “find images with 20% blue, 20% white and 60% don't care”. Part (a) shows the eight best matches out of 20,000 pictures, sorted by their similarities to the query. Notice that the best match has nothing to do with skiing (although its color signature matches the query well), and that consecutive images in the list can be very different from each other. Part (b) displays the MDS embedding of the best twenty matches where images of skiers are placed on the right: images with more snow at the bottom-right, and images with more sky at the top-right. The bottom-left holds images which contain also green, and the top-left holds images which contain also darker colors. Notice that image thumbnails placed at the coordinates returned by the MDS algorithm might occlude other thumbnails. Up to a point, this is not really a problem since these images are likely to be similar, and are therefore represented well by the topmost thumbnail.



(a)

(b)

Fig. 1. Looking for ski images; (a) Traditional display. Only the eight best matches are shown; (b) MDS display of the best twenty images (STRESS=0.148); A color version of this figure can be found at `http://vision.stanford.edu/~rubner`.

4 Navigation

Using the MDS embedding can assist navigation in the space of images, as we now illustrate. In our system the user starts a query by specifying the color content of the requested image. Using "don't care" is encouraged when the user is not confident about certain colors. A large set of images is returned and embedded in 2D space using MDS. Once the user selects an area of interest, an appropriate new query is generated and submitted. Now a smaller number of images is returned by the system. The new set of images is not necessarily a subset of the previous set, so images which were not returned by the previous query can still be retrieved at later stages, as the query becomes more precise. A new MDS is computed with new axes of variation which are based on the new image set.

To this end, in order to generate the new query, our user interface allows the user to draw a circle in the MDS map around images that are to be used to form the next query (In general, it would be better to let the user draw any closed region). The color signatures of these k images are used to generate a color signature that will be used for the next query. We want the new color signature to reflect the common clusters in the k color signatures, and ignore the others by using "don't care". This is done by representing the clusters as points in the CIE-Lab color space, and clustering these points using a similar algorithm to the one used in Section 2.1 to generate the color signatures. We reject clusters that are not represented by points from at least 50% of the k images. Each cluster is assigned a weight which is the median of the weights of the points in that cluster.

An example of navigation is given in Figure 2, where we are looking for images of deserts. The initial generic query was 20% blue for the sky and 80% "don't care". The MDS embedding of 400 returned images is shown in Part (a) of the figure. By glancing at the results, we see immediately the organization of the returned images: images of airplanes are in the bottom-left, diving images are in the top-left, images with green trees are in the mid-right, and so forth. Our desert images are in the bottom-right, so we select them (indicated in the figure by a black circle) and ask for 80 images. In the new MDS map shown in Part (b), most of the images are desert-like, with some buildings on the left. Although we could probably stop here and pick our favorite desert image, one more iteration is shown in Part (c) where we asked for only 20 images. Now all the images are deserts. Notice the cacti at the top-right, and the cougars at the mid-left.

Although MDS embeddings can be computed quickly for small sets of images (about 2 seconds for 80 images, and 0.15 seconds for 20 images on an SGI Indigo 2 with a 250 MHz processor), the computation time grows rapidly as the number of images increases (about a minute for 500 images). This is mostly because of the full distance matrix computation. The MDS technique, however, can tolerate missing data [4], and we are currently investigating ways to decrease computation times by computing only sparse distance matrices.

5 Conclusions

The methods presented in this paper open a novel set of tools and possibilities for image database navigation and visualization. Our color signatures and the EMD between them seem to approximate well the perceptual similarity or dissimilarity of images in terms of their color content. Furthermore, the low-dimensional embeddings we compute by MDS provide an intuitive way for the user to refine a query and to continue exploring interesting neighborhoods of the image space — or to see large portions of it all at once.

The idea of an adaptive image embedding can be applied to other modalities besides color, as long as some notion of similarity, metric or not, continuous or discrete, can be defined. For instance, for texture, shape, eigenimage similarity, or any other image features [1–3, 5–7, 9, 11] . In this context, the key question, which we leave for future work, is then to determine whether the main axes of variation

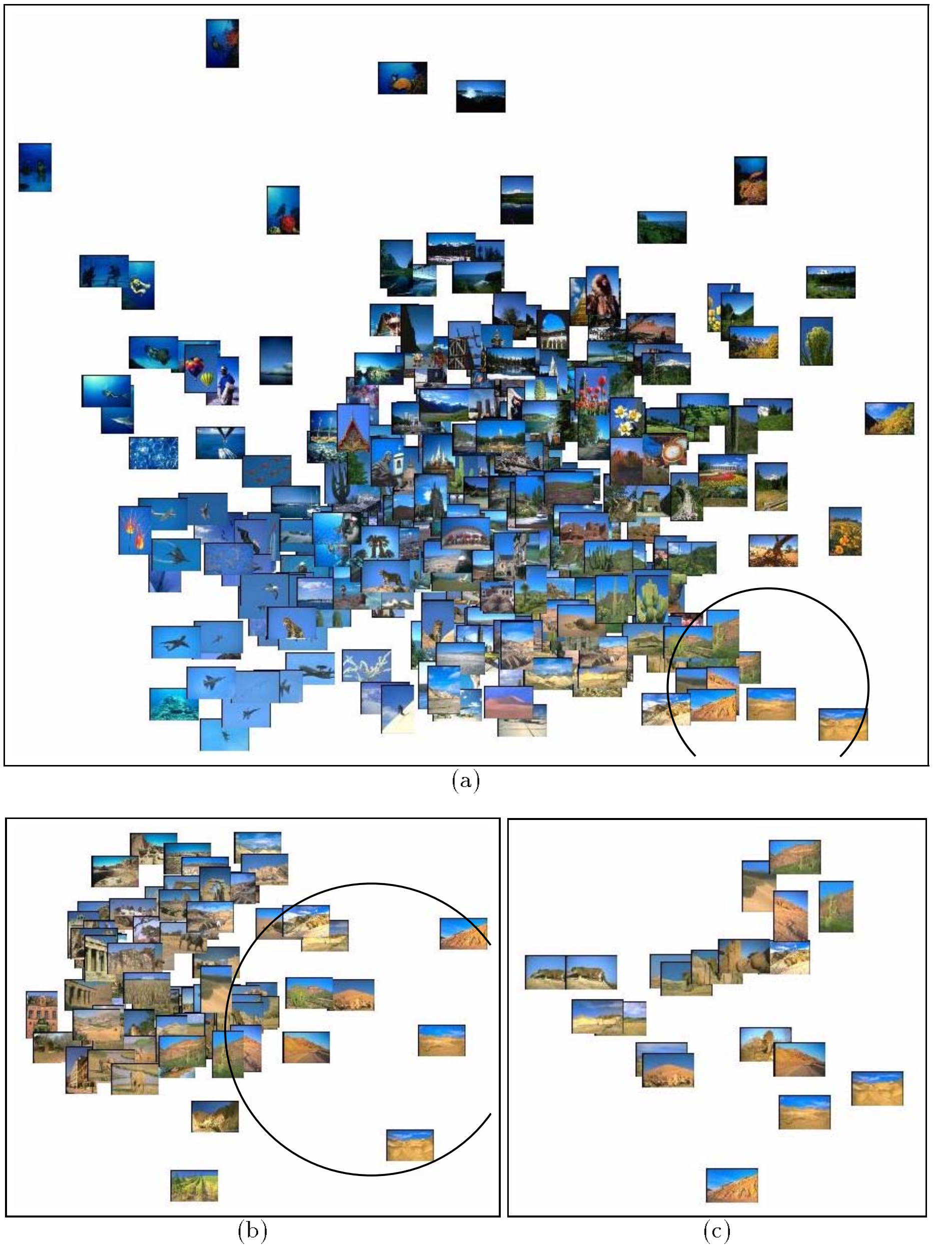


Fig. 2. Looking for a desert landscape; (a) 400 images; (b) 80 images; (c) 20 images. The black circles shows the user selection. A color version of this figure can be found at `http://vision.stanford.edu/~rubner`.

"discovered" by MDS for each of these distances and for various types of image distributions is perceptually meaningful. We believe that since the MDS groups together similar images — away from dissimilar ones — this is often the case. For instance, in [8] we defined a metric for textures, based on their spectral contents. We used this metric to compute a two-dimensional MDS on texture patches. The results were perceptually meaningful and agreed with psychophysics results. We also plan 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 may help in providing a globally consistent sense of navigation.

All image query systems 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.

References

1. 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. *SPIE* , 2670:76–87, 1996.
2. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.
3. D. Forsyth, J. Malik, M. Fleck, H. Greenspan, and T. Leung. Finding pictures of objects in large collections of images. *International Workshop on Object Recognition for Computer Vision*, 1996.
4. J. B. Kruskal. Multi-dimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, 29:1–27, 1964.
5. W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. *CVPR*, 425–430, 1996.
6. G. Pass and R. Zabih. Histogram refinement for content-based image retrieval. *IEEE Workshop on Applications of Computer Vision*, 1996.
7. A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *IJCV*, 18(3):233–254, 1996.
8. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. *IEEE ICCV*, 1998.
9. S. Santini and R. Jain. Similarity queries in image databases. *CVPR*, 646–651, 1996.
10. R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function, i and ii. *Psychometrika*, 27:125–140,219–246, 1962.
11. M. Stricker and M. Orengo. Similarity of color images. *SPIE*, 2420:381–392, 1995.
12. M. J. Swain and D. H. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.
13. Y. Takane, F. W. Young, and J. Leeuw. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, 42:7–67, 1977.
14. G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae.* Wiley, 1982.