Chapter 8

Conclusion

Image matching for content-based image retrieval (CBIR) applications is fundamentally different from image matching in the more traditional computer vision areas of stereopsis and tracking. In stereo, most of the information in one image has matching information in the other image (missing information results when part of the scene is visible from one viewpoint but not the other). The search for a correspondence to a feature in one image can be limited to the associated epipolar line in the other image. Also, the lighting for the two images is usually the same. In tracking, consecutive frames are nearly identical. Estimates of feature velocity can be used to predict where in the next frame a feature will appear, thus limiting the search space in finding correspondences. As in stereo, the lighting is usually constant during tracking. In contrast, the images we desire to match in CBIR can be visually quite different because most of the information in one image may not have any matching information in the other; images are not usually of the same scene. Also, a region in one image might match any region in another. The matching process may be further complicated by differences in illumination that cause the same object to appear differently in two images.

It is now time to take a step back and see what we have done, how it fits into the broader picture of image retrieval and image comparisons, and what remains to be done for the difficult image matching problems in CBIR.

8.1 Thesis Summary and Discussion

The pattern problem is difficult because partial matching and transformations are allowed. The scale component of a transformation plays a critical role here because it determines how much information in the image to compare to the pattern. A system might incorrectly conclude that a pattern is not present at some image location if the system's scale estimate

is very inaccurate. We believe that a good estimate for pattern scale is essential for the efficiency and correctness of a pattern problem solution, and we developed a novel scale estimation algorithm that uses the Earth Mover's Distance (EMD) between two attribute distributions.

The image and pattern signatures used throughout this thesis are distributions of mass in some feature space, where the amount of mass placed at a feature space point is the amount of that feature present in the image. We used a combined color-position feature space for the color pattern problem, and a combined orientation-position feature space for the shape pattern problem. In our scale estimation algorithm, we used a color feature space and an orientation feature space for the color and shape cases, respectively. These choices reflect our general strategy to obtain fast scale estimates by matching attribute-position distributions after marginalizing away position information.

The EMD is a general tool to measure the distance between distributions. Because we believe that mass distributions in a feature space are excellent image signatures for the pattern problem, we devoted a large part of this dissertation to the EMD, including modifications to aid in partial matching, lower bounds to aid in CBIR query efficiency, and computation under transformation sets. We made the EMD more amenable to partial matching with changes that (i) force only part of the mass in both distributions to be matched (the partial EMD), and (ii) measure the amount of mass that cannot be matched if there is a limit on the distance in feature space between allowable matches (the τ-EMD). We extended the centroid lower bound to the partial matching case, and we developed projection-based lower bounds which are also applicable to partial matching. Finally, we made an extensive study of computing the EMD under transformation sets, including theoretical analysis, documentation of the difficulty of the problem, the FT iteration to compute at least a locally optimal transformation (for a large class of transformation sets), cases with special structure that allow us to compute directly a globally optimal transformation, and the previously mentioned scale estimation algorithm that finds the scale of one distribution that minimizes its EMD to another.

The use of an iteration to compute the EMD under a transformation set reflects our decision on a major choice in the design of a pattern problem algorithm. Motivated by the importance of the pattern problem in CBIR, we believe that it is more important to solve many pattern problems quickly with the chance of a small number of errors than to solve every pattern problem correctly but more slowly. Although the FT iteration is not guaranteed to find a globally optimal transformation, its chances are greatly improved if the initial transformation is close to optimal. The combination of SEDL's scale estimation and initial placement methods is an efficient, very effective algorithm for computing a small

set of promising pattern regions within an image.

SEDL uses a simpler distance function than the EMD for efficiency reasons. This distance still allows partial matching under a transformation set, but it gives up the notion of morphing one distribution into a subset of the other in order to handle large image and pattern signatures. Our ideal distance measure involves this morphing notion, but also requires a change in the distribution representation. Currently, each mass in feature space is placed at a single point in that space. A region which is mainly blue, for example, is summarized by a mass at the point (blue, region centroid) in color $\times$ position space. Our ideal representation places masses over continuous regions in the feature space. Instead of summarizing a blue segmentation region as above, we could uniformly distribute mass over the entire extent of the region in position space. This is a more faithful representation of the blue region than mass concentrated at a single (color, position) point.

Given image signatures which are continuous mass distributions in feature space, we need a continuous version of the EMD to compute morphing distances. In section 2.4, we mentioned work in computing the EMD between two normal distributions and between two uniform distributions over elliptical volumes. We are, however, unaware of more general work on the continuous EMD which is capable of matching distributions such as those we have just proposed.

In an effort to understand why the continuous formulation is better than the discrete one, let us consider matching two distributions of color clusters with the EMD. In one image, a clustering algorithm might produce two clusters of somewhat similar reds, while in another image the matching red may be represented as a single cluster which is roughly the average red color of the corresponding two clusters. The EMD is not sensitive to such a non-canonicality of its input. In this example, it only pays a small cost to match the red mass because all three clusters are located close to one another in color space. Of course, the EMD would pay zero cost if both representations had one cluster with identical reds. As long as the representation is accurate, it affects the efficiency of the EMD computation but not the correctness of the result.

Using continuous mass distributions in attribute $\times$ position space with a continuous EMD aims for the same effect as above. Suppose, for example, that there are two matching green areas from two different images. The effect of representing the area as one region of green in the first image and two adjacent regions of similar greens in the other will be negligible under the continuous EMD. This effect may not be negligible if mass is placed only at region centroids because the centroid of the green region in the first image may be far away from the centroids of the green regions in the other image if the regions are large. To help reduce the impact of such problems, SEDL weighs region distances by pattern region

area, the theory being that large pattern regions are more stable in appearance and easier to detect as a single region within the image. The excellent results obtained show that this strategy is effective, but it is still desirable to have a distance function which is provably robust to non-canonicalities of the representation, yet fast enough to maintain reasonable interactivity with the user. In the shape case, SEDL avoids such representation problems to a large extent by using a fine sampling of image curves. This strategy is feasible because the shape data is one-dimensional, but a fine sampling over the 2D image plane, however, would produce color signatures too large to match in acceptable times for retrieval.

An extreme example in the color case is comparing (for the same image) a segmentation in which every pixel is a region and one which aims for the largest possible "uniform" color regions. The EMD between continuous distributions in color $\times$ position space derived from these segmentations should be small. With a continuous formulation, the segmentation used to produce a color $\times$ position signature is an efficiency issue, but not a correctness issue. It may be possible to match two continuous distributions of masses more quickly if there are fewer masses, but the EMD should be roughly the same if one distribution is replaced by another whose masses cover the same areas of the feature space. Computing the EMD between continuous distributions, or a good approximation if the exact answer is too expensive to compute, is a difficult but worthwhile future research problem.

8.2 Future Work

Developing partial matching distance measures that allow for scale-altering transformations is a crucial problem in CBIR because semantic image similarity often follows from only a partial match of unknown size. The road toward practically useful CBIR systems, however, still has a number of interesting and difficult challenges ahead.

Speed and Scalability to Large Databases

Users will demand semantic retrieval ability, but will not wait more than a few seconds for query results. Simple matching schemes on global image statistics will be fast, but are unlikely to return images which are semantically related to a given query. Speed issues must be addressed, but correctness issues are more important since there is no point in computing undesirable answers rapidly. Perhaps a breakthough will come when algorithms find complicated patterns as quickly as they find simple ones ([78]). This is not the case in SEDL, but intuitively a complicated pattern is more distinctive than a simpler one, and this distinctiveness should make it easier to find the pattern or discover that it is not present. It is not straightforward to avoid explicit query-image comparisons via clustering database

images (and comparing a query to cluster representatives) because partial match distance measures do not obey the triangle inequality and may be asymmetric.

Object Recognition in Cluttered Scenes

The experiments discussed in this thesis allow for changes in either camera pose or lighting, but not both at the same time. For example, our object recognition experiments use images of an object under different illuminants but taken with the same camera pose. In the color pattern retrieval experiments, we allow certain changes in object pose but do not account for changes in lighting. SEDL's pattern search is directed by the colors of regions, so big differences in lighting for the database and query images would cause a problem. A completely general pattern problem algorithm would allow for both photometric and geometric factors. In fact, some viewpoint differences may mean that different parts of an object may be visible in the query and database images. Thus, a pattern problem algorithm may also have to allow for partial matching of the query pattern. Allowing all these factors at once makes it difficult to rule out a pattern occurrence at a particular location. Maybe the pattern is present but the system's scale estimate is inaccurate, or all the information in the query should not be matched, or a color correction must be made to account for lighting.

Combination of Different Search Modalities

Another important issue in pattern problem algorithms is the use of more than one modality in judging visual similarity, for example region shapes and colors. Although SEDL is a uniform framework for the color and shape pattern problems, it does not use both color and shape information together. A good pattern problem algorithm that uses both color and shape should know when to use which information. Consider, for example, searching for a grayscale Apple logo within a color advertisement. The outline of the apple will match, but the colors will not. The problem of combining different types of information to measure perceptual similarity is a difficult one. For example, how much do the shapes of regions contribute to the perceptual similarity or dissimilarity of two color patterns?

Shape Representations and Distance Measures

We used the word *shape* in this thesis to mean a salient image curve or region boundary, and we considered the problem of measuring distances between individual shapes and sets of shapes. In the former problem, we represented shapes by their arclength versus turning angle curves and measured distance as sums of orientation differences between correponding

points along two curves. Of course, there are many other possible representations and distance measures, including comparison of coefficients in a Fourier decomposition, control points in a B-spline representation, moments in an area-based representation, or the amount of energy to deform one shape into another, just to name a few. We represented a set of curves as a set of (curve point, orientation) pairs obtained by a dense sampling of the curves. As in the single shape case, the distance between two sets of shapes was measured as a sum of orientation differences between corresponding points.

Our choices of representation and distance measures were motivated by the need to handle partial matching and transformations, but there may be better representations and distances for CBIR. The best choice may depend upon the application domain; nature images, product advertisements, Chinese characters, textbook drawings, and CAD models, for example, might all require different representations and distances. Perhaps the continuous EMD can be used as a common distance measure for different representations of both single shapes and sets of shapes, where mass is spread uniformly along curves or over regions. A B-spline representation, for example, describes the distribution of mass as a collection of uniform distributions over Bézier arcs. Is the morphing distance provided by the continuous EMD a perceptual one? For comparisons of one shape to another, maybe using the EMD to compare distributions of energy in frequency space will yield a perceptual measure of shape similarity, as it does in the texture case ([69, 68, 66]).

Image Browsing

We did not touch on the subject of navigating in the space of database images, but this is also an important mode of user interaction with a system. A user may not know exactly what he/she wants, but instead would like to browse the images. Navigation according to global color similarity is accomplished quite effectively by embedding images in a low (two or three) dimensional space such that distances in this space approximate well Earth Mover's Distances between image color signatures ([67, 69, 65]). How can this be done when there is no inherent metric structure imposed by the image distance function as for partial match distance measures, and when there are many dimensions along which images differ. Perhaps a solution to the general image matching problem of finding all pairs of similar regions from two images can help. Imagine a graph structure on database images where there are many links connecting two images, one for each pair of similar regions. These links may help navigate locally, but how can we give a user the big picture of database contents when the underlying image space has high dimension?

The Image Matching Problem

Solutions to the more general image matching problem will take us further toward semantic image retrieval than solutions to the pattern problem. As mentioned in the introduction, an algorithm for the pattern problem may be useful as a subroutine within an algorithm for the image matching problem. But a solution to the image matching problem is far from the end of the story. It is also a major challenge to interpret the results to provide semantic retrieval. This involves looking at positions of similar regions within the database and query images. For example, blue above green in a nature image likely indicates a scene of grass against sky, while blue below green probably means a scene of grass leading to water. Perhaps artificial intelligence knowledge representations and learning algorithms will prove useful here.

8.3 Final Thoughts

In the midst of all the technical details and interesting problems that arise in content-based image retrieval, we should not forget that the ultimate goal is to allow users to find information reliably and efficiently. Image-based and text-based search should complement one another to advance toward this goal. At the present, there is little collaboration between the traditional, text-oriented database community and the image-oriented computer vision and image understanding communities, but this will need to change if we are to produce the best possible information retrieval systems.