Chapter 1

Introduction

The invention of the World Wide Web has brought the need for automated image indexing and content-based image retrieval (CBIR) to the forefront of image processing and computer vision research. Although there has been significant progress in CBIR in the past five years, the general problem is far from solved. Semantic image understanding is beyond the scope of current state-of-the-art CBIR systems. A user who hopes to retrieve all database images of dogs by presenting a CBIR system with a query dog image is likely to be disappointed. Today's systems with automated indexing mechanisms record color, texture, and shape indices in the hope that similarity of these low level features and their locations in the database and query images will imply a high level semantic relationship. Thus, CBIR users today must make do with visual similarity instead of semantic similarity.

Measuring visual similarity with an eye toward semantic similarity between images is still a very difficult problem. In general, database images of interest are unconstrained input to a CBIR system. Images may be taken from any distance, at any time of day, under any weather conditions, under any illuminant, from any angle or viewpoint. Two images of the same object imaged under different illuminants and from different viewpoints will still look similar even though corresponding image pixels may be quite different in color. This visual similarity will persist even when the object is partially occluded from one of the viewpoints.

Comparing images of different scenes, on the other hand, is difficult even without the complication of lighting and viewpoint changes. If we ever want to obtain semantic similarity from measures of visual similarity, then our notion of visual similarity must allow for partial matching of images. A database image with regions that are similar to regions in a query image is likely to be related to the query in some way that the user cares about, and is therefore a good candidate for retrieval. It is common for a semantic relationship to exist even when only part of the information in the database image matches information in the

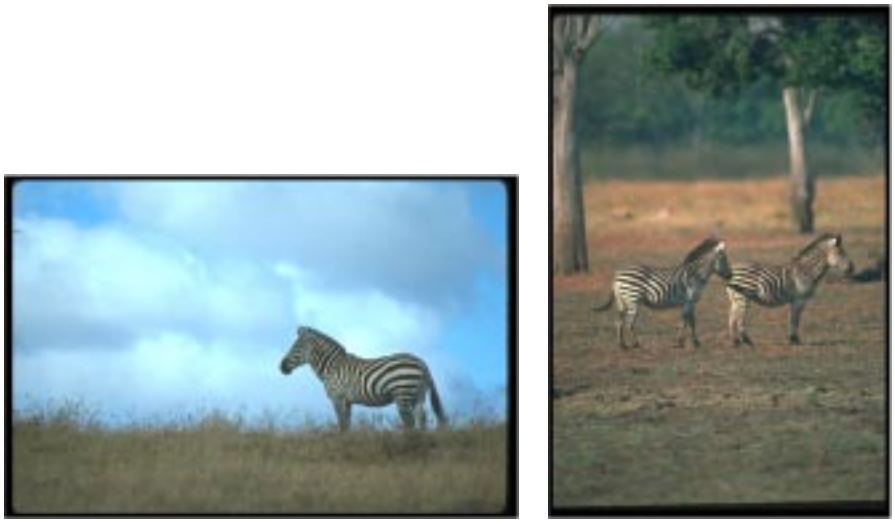

Figure 1.1: The Importance of Partial Matching. The left and right images are semantically related because both contain zebras, but there are no sky and clouds in the right image, and there are no trees in the left image. Partial matching is crucial in any CBIR system that aims to capture semantic similarity.

query, and only part of the information in the query matches information in the database image. The images in Figure 1.1, for example, are semantically related because they both contain zebras, but there are no sky and clouds in the right image, and there are no trees in the left image. It is also possible for a semantic relationship to exist when only a very small fraction of the one image can be matched to the other image. Consider, for example, the Apple logo image and the Apple advertisement shown in Figure 1.2. The Apple logo is less than one half of one percent of the advertisement. The ability to find even very small partial matches is important in CBIR.

The *image matching problem* is to identify all pairs of visually similar subregions from two images. An efficient solution to this problem is a holy grail in CBIR. Such region similarities provide crucial information to a CBIR system that attempts to make reasonable guesses as to the similarity of the semantic content of images. These guesses may be based on the relative positions of the similar regions, as well as which regions have no similar region in the other image. The set of similar regions can also be displayed to the user to show why a particular database image was retrieved for a given query.

At the heart of virtually any CBIR system is its image distance measure. Such distance measures usually do not operate directly on the images themselves, but rather on image summaries or *signatures* that record information in a form more suitable for efficient comparison. The main distance measure discussed in this thesis is the *Earth Mover's Distance*

Figure 1.2: The Importance of a Small Partial Match. The Apple logo image on the left is semantically related to the Apple advertisement on the right even though the logo covers less than one half of one percent of the total area of the advertisement.

(EMD). The use of the EMD in image retrieval was pioneered by Rubner, Tomasi, and Guibas ([69, 67, 65, 68]). This distance measure compares image signatures which are distributions of mass or weight in some underlying feature space. The weight associated with a particular point in the feature space is the amount of that feature present in the image, and a distribution is a set of (point,weight) pairs. The EMD between two distributions is proportional to the minimum work required to change one distribution into the other. The morphing process involves moving around mass within the feature space (hence the name of the distance measure). The notion of work is borrowed from physics. One unit of work is the amount of work necessary to move one unit of weight by one unit of distance in the feature space. The EMD framework has been successfully applied in color-based ([69, 67, 65, 68]) and texture-based ([69, 68, 66]) retrieval systems. The differences in these two cases are simply the feature space and the distance measure in the feature space.

1.1 The Pattern Problem

This thesis is concerned with a slightly simplified version of the image matching problem which we call the *pattern problem.*

The Pattern Problem. Given an image and a query pattern, determine if the image contains a region which is visually similar to the pattern. If so, find at least one such image region.

In contrast to the image matching problem, in the pattern problem we search for the entire query as a single subregion of the image. A solution to the pattern problem can be used to build a solution to the image matching problem if a query image can be decomposed into atomic regions of interest. This is certainly a reasonable assumption in the CBIR context since the user can manually outline a few relevant regions in the query before submitting it to the system. A routine that solves the pattern problem can be called with each of the query regions as the pattern.

The pattern problem is very difficult because of the combination of partial matching and scaling. The pattern may occur at any location in the image, and at any size. It is not known a priori where to look in the image, or how much of the image area around a given location to examine. Without a good estimate of the scale, it is very difficult for an algorithm to conclude that a pattern does not exist at a particular image location. If the assumed scale of the pattern is too large, then the image location is unfairly penalized because too much information is being examined, much of which may have no matching information in the query pattern. If the assumed scale of the pattern is too small, then much of the pattern information may not be matched because not enough area around the hypothesized pattern location is being examined. These points are illustrated in Figure 1.3, where we compare the color signatures for a query pattern and various size rectangular regions around and within the occurrence of the pattern in a database image.

Another difficult issue in the pattern problem is efficiency, even when the scale of the pattern is known. If the pattern scale is very small, then there are many nonoverlapping (and therefore independent) image locations to check for the pattern. This leads to an efficiency problem in the CBIR context in which a pattern problem must be solved for many (query, database image) pairs. To compound the problem even further, it is difficult to prune a search for a pattern in one image because of a negative search result in another image. Such pruning is possible in CBIR systems which have a true metric as an image distance measure. If query Q is far from database image P, and database image P is close

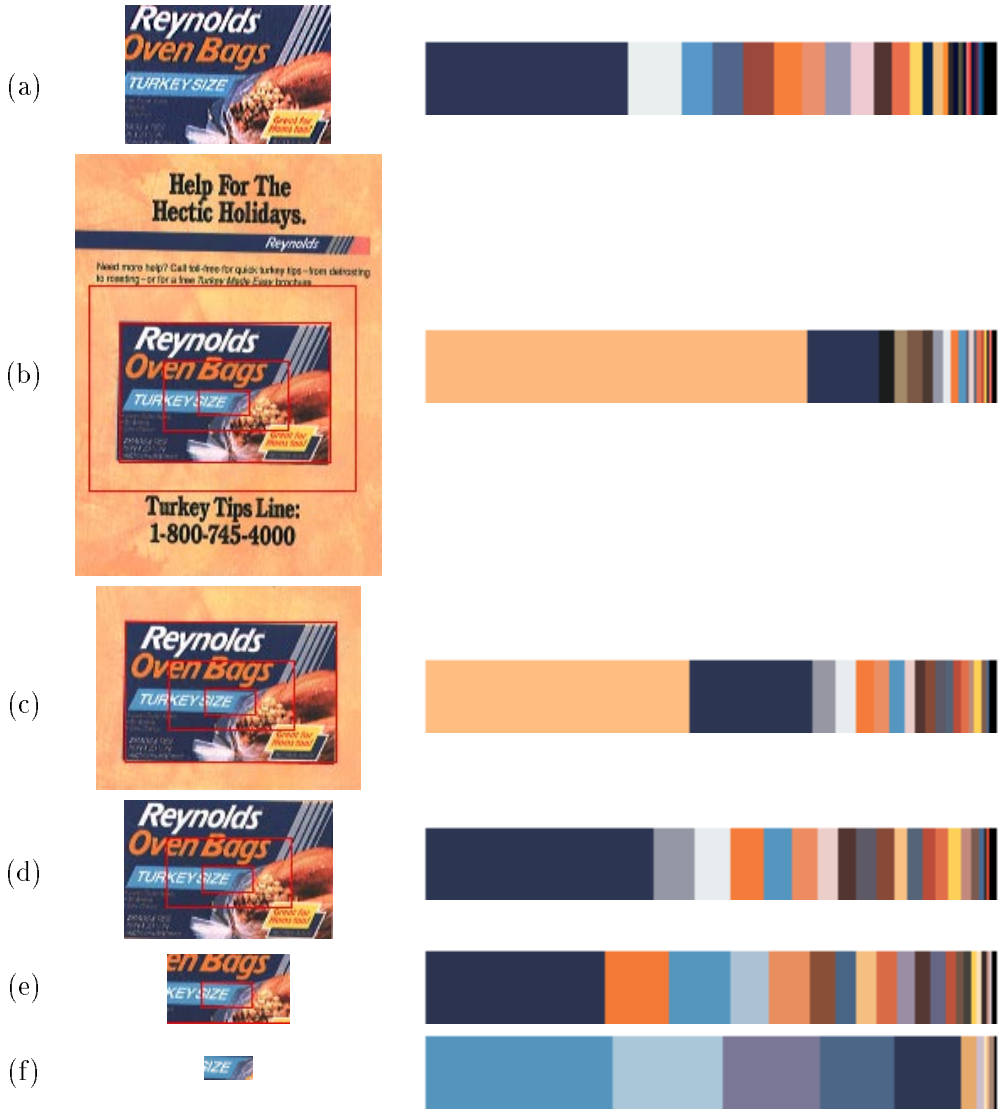

Figure 1.3: The Importance of Scale. The estimated scale at which a pattern appears is important because it determines the amount of information in the database image to compare to the information in the pattern. In row (a), we show the pattern and its color signature. In row (b), we show the database image and its color signature. In rows (c)-(f), we show various subregions of the image in (b) along with their color signatures. The EMD between the pattern signature and each of the image signatures (in CIE-Lab color space units) is (b) 27.7, (c) 19.8, (d) 5.5, (e) 9.4, (f) 20.8. Note that the subregion shown in (d) is almost exactly the occurrence of the pattern in the database image.

to database image R, then the triangle inequality implies that Q is also far from R. Any distance measure which allows for partial matching, however, will not be a metric because the triangle inequality can be violated. A query pattern Q may not occur in image P ($d(Q, P)$ is large), and image P may have many regions in common with image R ($d(P, R)$ is small), but the pattern may still occur in image R ($d(Q, R)$ is small) in a region of R which is not similar to a region in P.

In addition to the sheer number of region comparisons that a brute force approach with or without accurate scale information would perform, there is also the difficulty that each such comparison is not straightforward. Even if the system knows the scale at which the pattern may occur in an image, the pattern may occur rotated in the image with respect to the example presented to the system. Under general imaging assumptions, the comparison between regions must allow for a projective transformation between the image region and the query. In addition to this geometric transformation in image position space, in the color pattern case we might need to account for a photometric transformation in judging visual region similarity. The same object imaged under different light sources may have very different pixel colors, but will still appear very similar to a human observer. In terms of the underlying imaging system, unknown pattern scale, pose, and color appearance are the result of unknown camera-to-object distance, unknown camera viewpoint, and unknown lighting conditions. Also, perceptual color similarity is a complex and ill-understood notion which depends on context and many other factors; these issues are beyond the scope of this work.

Once geometric and photometric factors have been accounted for, the match between images of the same object from different viewpoints and under different illumination conditions will be nearly exact. However, we need to measure similarity and allow for inexact pattern matches. This raises the difficult problem of how to combine color and position information in judging the visual similarity of two color patterns.

1.2 Thesis Overview

This thesis is devoted to the pattern problem in the context of content-based image retrieval. Four main themes are present:

(i) partial matching,

(ii) matching under transformation sets,

(iii) combining (i) and (ii), and

(iv) effective pruning of unnecessary, expensive distance/matching computations.

In **chapter 2**, we give some background information to help put this thesis in the context of previous research. This includes brief descriptions of works with similar goals and that discuss similar problems to those in our work, as well as a discussion of high level differences in motivation, approach, and technique. In **chapter 3**, we solve a 1D shape pattern problem which seeks all (possibly scaled and rotated) approximate occurrences of a pattern polyline shape within another polyline.

The thesis shift gears a bit in **chapter 4** where we discuss the Earth Mover's Distance and a couple of modifications which make it more amenable for use in partial matching settings. One of these modifications is the *partial* EMD in which only a given fraction of the total weight in a distribution is forced to match weight in the other distribution. The other modification is the τ-EMD which measures the amount of weight that cannot be matched when weight moves are limited to at most τ units. Also included in this chapter is an algorithm that uses the EMD to estimate the scale at which a pattern may occur in an image. The issue of efficiency is the central theme of **chapter 5** in which we present efficient lower bounds on the EMD that often allow a system to avoid many more expensive, exact EMD calculations. These lower bounds were developed and are illustrated within the context of the color-based retrieval system described in [65].

In **chapter 6**, we extend the Earth Mover's Distance to allow for unpenalized distribution transformations. We consider the problem of computing a transformation of one distribution which minimizes its EMD to another, where the set of allowable transformations is given. The previously mentioned scale estimation problem is phrased and efficiently solved as an EMD under transformation ($\text{EMD}_{\mathcal{G}}$) problem in which transformations change the weights of a distribution but leave its points fixed. For $\text{EMD}_{\mathcal{G}}$ problems with transformations that modify the points of a distribution but not its weights, we present a monotonically convergent iteration called the *FT iteration*. This iteration may, however, converge to only a locally (cf. globally) optimal EMD value and transformation. The FT iteration is very general, and is modified to work with the partial EMD mentioned above, as well as in some cases when transformations modify both distribution points and weights. We also discuss cases of the $\text{EMD}_{\mathcal{G}}$ problem which can be solved directly, without our iteration.

In **chapter 7**, we describe the SEDL (Scale Estimation for Directed Location) content-based image retrieval system for the pattern problem. The SEDL framework is general enough to be applied to both the color and shape pattern problems. In the shape case, images are sets of curves such as might be produced by edgel detection and linking, or by any standard drawing program. Excellent results for a color database of product advertisements

and a shape database of Chinese characters are shown.

A key component in SEDL is the previously mentioned scale estimation module. The output of this module is either an estimate of the pattern scale in the database image or an assertion that the pattern does not appear in the image. In the initial placement phase that follows scale estimation, SEDL efficiently determines a handful of places in the image where the pattern might occur at the previously estimated scale. This small set of promising locations mark the starting points for the final verification and refinement phase. For each initial placement of the query at the estimated scale, SEDL checks for positional consistency of the underlying attributes (for example, colors), modifying the attribute locations by some transformation if this will help improve the match. In recognition of the difficulty of combining attribute and position information, a final check using the τ-EMD helps eliminate false positives.

Finally, we conclude in **chapter 8** with a thesis summary, main insights, and suggestions for future work.