Research in Algorithms for Geometric Pattern Matching MIT2001-06

Progress Report: July 1, 2002 – December 31, 2002

Piotr Indyk

Project Overview

Geometric pattern matching is pervasive in many areas of computer science, e.g., in computer vision, computational drug design and computational biology.
The goal of this project is to develop efficient algorithms for key geometric pattern matching problems.

As mentioned earlier, the focus of this project is to design a fast similarity search algorithm for large data sets of images. To evaluate the similarity between images, we use the Earth-Mover Metric (EMD). It was experimentally verified to capture well the perceptual notion of a difference between images, in fact much better than other well-known metrics (e.g., Euclidean distance between the feature vectors). The basic idea behind EMD is as follows. Assume that the features of an image are represented by a set of points in low-dimensional space R^d. For example, an image could be represented by a set of pixels, where each pixel is a point in 3-dimensional color space or texture space. The distance between two sets of points (representing two different images) is defined as the minimum amount of work needed to transform one set into another. Formally, this corresponds to the minimum weight matching between the two sets of points.

Since EMD has been shown to outperform other measures for comparing color or texture similarity between images, it is of great interest to design efficient algorithms for pattern matching under this metric. In particular, the most interesting case occurs when one is given a “query” image, and wants to scan a large database of images, in order to find the image most similar to the query. The approach used so far is to compute the distances between the query image and *each* image stored in the database. This is highly inefficient, since the time needed to answer a query could be very large for large databases.

During the earlier stages of this project we designed and implemented a method which drastically reduces the time needed to solve this problem [IT’01]. The main idea of our approach is to *embed* the Earth Mover Distance into Manhattan space and use very efficient nearest neighbor data structure for the latter (well-studied) space. In other words, we show that one can represent each pixel set by a feature vector, in such a way that the EMD between two pixel sets is approximately proportional to the Manhattan distance between the feature vectors. The distortion induced by the embedding algorithm is provably bounded.

Since very fast nearest neighbor algorithms for normed spaces are known (e.g., see [IM’98, GIM’98]), our embedding method yields dramatic improvement in the running of nearest neighbor algorithms for EMD. However, as we mentioned above, the embedding is not exact – it introduces a small error which *could* in principle affect the quality of the retrieved images. Thus, for this approach to work in practice, it is crucial to verify that the *actual* error occurring in practice is low. This could require additional adjustments and fine-tuning of the algorithm, to minimize the embedding error.

In the next section we describe our progress on implementing and evaluating our method in the context of image retrieval in large image databases.

Progress Through December 2002

During the period of June-December 2002, we implemented and evaluated algorithms for fast nearest neighbor search in R^d. For this purpose, we implemented a variant of the Locality-Sensitive Hashing algorithm [GIM'99]. To this end, we needed to modify many parts of the original algorithm. In particular:

- The original algorithm worked efficiently only when the input vectors were binary; we introduced new hashing scheme that works directly on points in d-dimensional space
- The original algorithm could not handle sparse vectors efficiently. This was undesirable, since the vectors obtained by using our embedding tools were very high-dimensional but sparse. We adapted the algorithm to deal with sparse data.

The final algorithm enables very fast similarity search in large collections of high-dimensional images. In particular, for a set of 20,000 points obtained by extracting color features of Corel-Draw images, LSH returned the answers 10-20 times faster than linear scan (the best previous method). Although LSH is a probabilistic and approximate algorithm, it almost always returned exact nearest neighbor. Since the "EMD to R^d " embedding introduces some small error (about 15%), our algorithm incurs a small error with respect to the original EMD metric, while achieving an order of magnitude speedup over earlier methods. We mention that the LSH algorithm has not been fine-tuned yet, and thus we expect much larger gain in future experiments.

Research Plan for the Next Six Months

Our main goals for the nearest future are:

- Build an easy-to-navigate user interface
- Perform rigorous user experiments to measure the influence of the embedding error on the perceptual retrieval error
- Write a report

In addition (time permitting) we plan to investigate the (quite fortunate!) discrepancy between the theoretical error bounds and the error we achieve in practice. This could entail developing a model for color histogram data sets which are closer to reality than the current worst-case approach.

References

[EMD] Scott Cohen, "Computing Earth-Mover distance under transformations",
http://robotics.stanford.edu/~scohen/research/emdg/emdg.html

[GIM'99] Aris Gionis, Piotr Indyk and Rajeev Motwani, "Similarity Search in High Dimensions via Hashing", IEEE Symposium on Very Large Databases, 1999.

[IM'98] Piotr Indyk and Rajeev Motwani, "Approximate Nearest Neighbor – Towards Removing the Curse of Dimensionality", ACM Symposium on Theory of Computing, 1998.

[IT'01] Piotr Indyk and Nitin Thaper, "Embedding Earth-Mover Distance into the Euclidean space", manuscript, 2001.