Research in Algorithms for Geometric Pattern Matching MIT2001-06

Progress Report: January 1, 2002 – June 30, 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.

During the period of January-June 2002, the main focus of this project was implementing and evaluating algorithms for embedding Earth-Mover Distance into the Euclidean space. Earth-mover distance (EMD) is a recently proposed metric for computing distance between features of images (see [EMD] and references therein). 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 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 the Euclidean 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 Euclidean distance between the feature vectors. The distortion induced by the embedding algorithm is provably bounded.

Since very fast nearest neighbor algorithms for the Euclidean space 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 June 2002

We have implemented a system, which given a large collection of images, extracts color features of the images, embeds them into the Euclidean space and then enables to search for similar images. Due to efficient implementation of the algorithm, the embedding procedures take very little time compared to the time needed to *extract* color information from images, and thus time overhead of our method is essentially negligible.

Having the system ready, we performed preliminary experiments comparing similarities of images under the embedded EMD with the actual EMD. The experiments were performed on images taken from the CorelDraw image database, with varied number of images. The experiments revealed that the actual embedding error, as well as the error resulting from computing nearest neighbor under the "embedded EMD" as opposed to original EMD, was very small. In many cases, it was smaller than 10%. This discovery is quite significant, since it means that the actual error is much smaller than it would follow from our current theoretical bound [IT'01]. In addition, anecdotal *user* experiments indicate that this small error is negligible in the context of searching for *perceptually* similar images. Thus, our algorithm replaces the need of dealing with a fairly intractable metric (the well-studied Euclidean space is used instead), while essentially preserving the quality of the retrieval of the original EMD metric.

Research Plan for the Next Six Months

Our main goals for the nearest future are:

- Perform further experiments on additional data sets, to evaluate the embedding error, as a function of various embedding parameters. This will allow us to fine-tune the parameters to achieve the optimal retrieval performance
- Implement fast nearest neighbor algorithms for searching in the space of embedded color histograms; perform extensive timing experiments
- 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.