Research in Algorithms for Geometric Pattern Matching MIT2001-06

Progress Report: November 1, 2001—December 31, 2001

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.

Progress Through December 2001

We highlight progress made during this period on several ongoing projects.

1. Frechet distance.

Frechet distance is one of the most fundamental measures used to compute dissimilarity between two *sequences* of points (e.g., between two signatures recorded by an electronic pen). It is often referred to as the *"dog-leash" distance*. This is because it corresponds to the minimal length of a "leash" so that a dog and its owner can traverse their respective sequences from beginning to end, while respecting the constraint that at any moment of time, the distance between them cannot exceed the leash length. See Project Proposal for formal definition.

In [I'02] we present a novel algorithm for searching a large database of sequences in order to find the sequence closest to a given "query" sequence. All previous algorithms for this problem required either a linear scan over the whole database (which is prohibitive for large databases), or additional storage of size *exponential* in the sequence length. In our paper we present the first algorithm which avoids the two aforementioned bottlenecks. In particular, the query time is very fast (poly-logarithmic in the database size) while the additional storage is arbitrarily small (see the paper for detailed tradeoffs). The amount of required storage is, however, still too large for most practical scenarios. More work is needed in order to obtain a more practical version of the algorithm.

2. Earth-mover distance.

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.

We proposed 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 small.

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. In the next section we describe our plans to evaluate our method in the context of image retrieval in large image databases.

Research Plan for the Next Six Months

Our main goal for the next six months is to investigate our algorithm for embedding EMD into Euclidean space. In the first stage, we are planning to build a software system, which given a large collection of images, extracts color features of the images and embeds them into the Euclidean space. When such system is ready, we will perform extensive experiments comparing similarities of images under the embedded EMD with other metrics, including original EMD as well as Euclidean distance between the color histograms. The experiments are going to be performed for several large image collections, including the CorelDraw image database.

After experimental verification of the embedding method, we are planning to augment the system with efficient implementation of a nearest neighbor search algorithm. Most likely this will involve re-implementation of the algorithm of [IM'98, GIM'99] for this particular application. We expect that the resulting system will be able to perform similarity queries in large image databases, orders of magnitude faster than the current methods.

In addition to the above plans, we intend to continue basic research on geometric pattern matching algorithms. We believe that designing efficient algorithms is the best way to obtain dramatic improvements in performance, and that geometric pattern matching is an ideal area for this approach to succeed.

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.

[I’02] Piotr Indyk, “Approximate Nearest Neighbor under Frechet Distance via Product Metrics”, ACM Symposium on Computational Geometry, 2002, to appear.