Approximate Nearest Neighbor Algorithms for Hausdorff Metrics via Embeddings

MARTIN FARACH-COLTON [*]
RUTGERS UNIVERSITY

PIOTR INDYK [†]
STANFORD UNIVERSITY

April 23, 1999

Abstract

Hausdorff metrics are used in geometric settings for measuring the distance between sets of points. They have been used extensively in areas such as computer vision, pattern recognition and computational chemistry. While computing the distance between a single pair of sets under the Hausdorff metric has been well studied, no results were known for the *Nearest Neighbor* problem under Hausdorff metrics. Indeed, no results were known for the nearest neighbor problem for any metric without norm structure, of which the Hausdorff is one.

We present the first nearest neighbor algorithm for the Hausdorff metric. We achieve our result by embedding Hausdorff metrics into l_∞ and using known nearest neighbor algorithms for this target metric. We give upper and lower bounds on the number of dimensions needed for such an l_∞ embedding. Our bounds require the introduction of new techniques based on superimposed codes and non-uniform sampling.

[*] Department of Computer Science, Rutgers University, Piscataway, NJ 08855, USA. (*farach@cs.rutgers.edu*, *http://www.cs.rutgers.edu/∼farach*). Supported by NSF Career Development Award CCR-9501942, NATO Grant CRG 960215 and an Alfred P. Sloan Research Fellowship.

[†] Supported by Stanford Graduate Fellowship and ARO MURI Grant DAAH04–96–1–0007. Part of this work was done while the author was visiting Bell Laboratories.

1 Introduction

The *Nearest Neighbor Search (NNS)* problem is: Given a set of n points $X = \{x_1, \ldots, x_n\}$ in a metric space with distance function d, preprocess X so as to efficiently answer queries for finding the point in P closest to a query point q. This problem has been well-studied in the case where d is the k-dimensional Euclidean space. The low-dimensional case is well-solved [9], though running times and space are exponential in the dimension. In [15] and [19], the approximate version of the problem was addressed in an effort to reduce the dependence on d. Recently, [14] considered the nearest neighbor problems in non-Euclidean metrics, in particular for the l_∞ norm.

All of these results are examples of nearest neighbor searching in *normed* spaces, that is, the metrics on the points form a norm. Such metrics have a good deal of structure which can be exploited algorithmically. While many metrics of interest for nearest neighbor searching are normed, not all are. One of the most interesting cases of a non-normed metric are the *Hausdorff metrics*. The Hausdorff metric is an example of a *derived metric*. Suppose we have an arbitrary underlying metric d on a set of points X. Then, for any two subsets (say A and B) of X the (directed) Hausdorff distance from A to B is defined as a maximum distance from any point from A to its nearest neighbor in B; the undirected Hausdorff distance between A and B is computed by considering both directions and taking the larger value.

Hausdorff metrics do not have the useful structure of norms. Even if the underlying metric is "well-behaved" (for example, when taking the Hausdorff metric of points in low-dimensional Euclidean space) no nearest neighbor algorithms are known. This state of the art is unfortunate, since the Hausdorff metric (over low dimensional Euclidean spaces) is a commonly used metric over geometric objects. Geometric point set matching in two and three dimensions is a well-studied family of problems with application to areas such as computer vision [22], pattern recognition [6, 13] and computational chemistry [11, 12, 23]. Thus the problem of computing (exactly or approximately) the Hausdorff distance between two point sets P and Q in two and three dimensions has been studied extensively [1, 5, 6, 13, 24] with the interesting problems being those where one set can be rotated or translated and one seeks the transform which minimizes the Hausdorff distance.

Unfortunately, no efficient algorithms have been designed for the case when we want to match P with *many* Q's and find the closest one. This problem is of crucial importance in many applications; in particular, computational chemistry [11, 12, 23] and pattern recognition [6, 13], require matching a pattern against a huge database of molecules or images, respectively.

Our results. Here we sketch out the flavor of the results, highlighting the main contributions.

Our first result is an algorithm for approximate nearest neighbor searching in Hausdorff metrics over low dimensional normed spaces l_p^d. The algorithm proceeds by approximately embedding the Hausdorff metric into l_∞ norm with dimension D roughly equal to $O(s^2/\epsilon^d)$, where s is the sets' size and ϵ is the distortion (see formal definitions in Section 2). Notice, that D does *not* depend on the database size, but only on size of database sets, which is much smaller. After the embedding, we apply the approximate nearest neighbor search algorithm in l_∞ of [14] to find the neighbor. In particular, for the most interesting case of $d = 2$ or $d = 3$ we get:

- A constant factor approximation algorithm with query time roughly $O(s^2 \log n)$ and mildly superpolynomial storage $n^{O(\log s)}$

- An $O(\log\log s)$-approximation algorithm with the same query time and roughly $s^2n^{1+\rho}$ space, for any $\rho > 0$.

Our algorithms can be generalized to minimum Hausdorff distance under *isometries* (i.e. rotations and translations). The time/space bounds remain essentially the same if only translations are allowed; for general isometries both query time and space are multiplied by s. Also, our approach (i.e. embedding Hausdorff metric into l_∞) has several additional benefits. One of them is that any future improvements of algorithms for l_∞ automatically yield improved algorithms for Hausdorff distance. Also, from the practical prespective, it gives a flexibility in choosing the l_∞ algorithm (from many existing implementations) which works best for particular applications.

Since the dimensionality D of the l_∞ space is crucial for the efficiency of the embedding, we further investigate the relationship between D and s, d, and ϵ. In particular, we show a lower bound for D of roughly s^2 for the case when the underlying norm is $l_\infty^{\log s}$. Since the corresponding upper bound in this case is $O(s^2O(1)^d) = s^{O(1)}$, we conclude that either the superlinear dependence on s or exponential dependence on d is neccessary[1] (we believe that the most likely case is that both of them occur). It is interesting that both the upper and lower bound uses *superimposed codes*; in particular, our lower bound proceeds by showing that an assumed embedding allows us to construct codes of small length, which contradicts known lower bounds. We believe this technique can be applicable to showing other lower bounds for derived metrics.

Our second line of research focus on *arbitrary* underlying metrics. We show the following result: a Hausdorff metric over any metric M can be embedded into l_∞ with roughly $D = s^2m^\rho$ dimensions and constant error (here m is the size of of the metric). Moreover, if efficient approximate nearest neighbor oracle exists for M, then the embedding can be performed by using D oracle calls. Thus we obtained the following surprising structural result: for any metric M for which a sublinear-time approximate nearest neighbor algorithm exists, the approximate nearest neighbor problem on the Hausdorff metric over M also has a sublinear-time algorithm.

Our embedding is randomized, requiring the selection of a collection of reference sets which are used in the embedding. In this regard, it resembles other embedding algorithms (see e.g. [4, 20, 21], etc.). However, in all of those results, the reference sets are selected uniformly at random, a scheme which turns out not to work in our case. Instead, we develop an embedding algorithm which relies on *non-uniform* sampling. We believe this technique could find further applications for other metric space problems.

Outline. In §2, we introduce notation and preliminary ideas. In §3, we give bounds on embedding Hausdorff metrics over normed spaces. In §4, we give bounds on embedding Hausdorff metrics over general space. Finally, in §5, we show how to use our embeddings for nearest neighbor searching.

2 Preliminaries

Metric spaces. Let $X = \{x_1, \ldots, x_n\}$ and let $d : X^2 \rightarrow \Re^+$ be a *metric*, that is $d(x,y) = d(y,x) \geq d(x,x) = 0$ and $d(x,y) \leq d(x,z) + d(y,z)$. The pair (X,d) forms a *finite metric space*. We extend d to pairs p, S where $S \subset X$ by defining $d(p,S) = \min_{q \in S} d(p,q)$. We also extend it to pairs S, S' for $S, S' \subset X$

[1] For the Hausdorff metric over l_p^d for $p < \infty$ the dependence on d can be seen to be exponential even when $s = 1$; this follows from lower bounds for embedding of l_p into l_∞. However, the same argument clearly does not apply to the case $p = \infty$.

by defining $d(S, S') = \min_{p \in S} d(p, S')$. Notice that the above extensions do not constitute metric spaces and should not be confused with the *Hausdorff metric* defined later.

Let $P \in \Re^d$. Then

$$l_k(P) = \left(\sum_{i=1}^{d} |P[i]|^k \right)^{\frac{1}{k}},$$

$$l_\infty(P) = \lim_{k \to \infty} l_k(P) = \max_{1 \le i \le d} |P[i]|.$$

When $X \subset \Re^d$, we refer to (X, l_k) as l_k^d, with X understood. We will sometimes refer to l_2^d as *Euclidean d-space*. When we wish to emphasize that the points in l_k^d are in some range $[0, R]^d$ rather than in $\Re^d$, we will write $[0, R]_k^d$.

For any $p \in X$ and $r > 0$ we define $B(p, r)$ to be the set of points $q \in X$ such that $d(p, q) \le r$.

For any pair of (X, d) and (X', d') of metric spaces, we say that a function $f : X \to X'$ is an (R, α, r)-*embedding* if for any points $p, q \in X$:

- if $d(p, q) \le r$ then $d'(f(p), f(q)) \le r$
- if $d(p, q) \ge R$ then $d'(f(p), f(q)) \ge \alpha R$.

Most of the embeddings introduced in this paper are, in fact, contractions; in such cases we call them (R, α)-embeddings (as the value of r is irrelevant). Furthermore, in many situations the actual value of R is not important; in such cases we assume $R = 1$ and call f an α-embedding.

Hausdorff metric. For any metric $D = (X, d)$ the Hausdorff metric over D (denoted by $H(D)$) is defined over the powerset of X. For any sets $A, B \subset X$ the Hausdorff distance d_D is equal to

$$d_D(A, B) = \max\{\max_{p \in A} d(p, B), \max_{p \in B} d(p, A)\}$$

In the following we often restrict the domain of $H(D)$ to subsets of X of cardinality upper bounded by a parameter s. We denote the resulting metric by $H^s(D)$. Also, we consider *generalized Hausdorff spaces* $H_T(D)$, parameterized by $T : 2^X \to 2^X$. In this case the distance function d_D^T is defined as:

$$d_D^T(A, B) = \min_{t \in T} d_D(t(A), B))$$

As long as T is closed under composition and inversion, the function $d_D^T(A, B)$ is a metric.

The problem of estimating the (generalized) Hausdorff distance between two point sets in 2 and 3 dimensions (usually under translations and rigid motions) has been studied extensively [7, 13, 24] (see also the survey by Alt and Guibas [3]). The approximate versions of the above problems have also been investigated [1, 16, 5]. In particular, the combination of the results of [1] and [5] results in an $O(s \log s)$-time algorithm for estimating (up to any constant factor) the Hausdorff distance of sets from $H_T^s(l_2^2)$, where T is the set of all rigid motions.

Finally, we consider another derived metric similar to the Hausdorff, defined as follows:

$$d_D(A, B) = \sum_{p \in A} d(p, B) + \sum_{p \in B} d(p, A)\}$$

We will commonly refer to this d_D as $H_1(D)$. Using this notation, the Hausdorff metric can be thought of as the $H_\infty(D)$ metric, however, we will use the simpler $H(D)$ for convenience throughout.

Approximate nearest neighbor algorithms. The approximate nearest neighbor problem was recently the subject of extensive research. The most recent results of [15] and [19] give algorithms for approximate nearest neighbor in d-dimensional Euclidean space with polynomial storage and query time polynomial in $\log n$ and d. These algorithms are of mainly theoretical interest, as their storage requirements are quite large. Indyk and Motwani [15] also gave another algorithm with small polynomial storage and sublinear query time. Unfortunately, the techniques used to achieve these results heavily exploit properties of the Euclidean norm and therefore do not seem applicable to other metric spaces. Subsequently, Indyk [14] gave an algorithm for the approximate nearest neighbor problem in l_∞^d. The algorithm achieves an approximation ratio of $4 \log_{1+\rho} \log 4d$ with $O(dn^{1+\rho} \log n)$ storage and $O(d \log n)$ query time. The latter result is crucial for our applications, as we obtain our results by embedding Hausdorff metrics into l_∞.

Definition 1 (c-Point Location in Equal Balls (c-PLEB)) *Given n unit balls centered at $P = \{p_1, \ldots, p_n\}$ in metric space (X, d), devise a data structure which for any query point $q \in X$ does the following:*

- *if there exists $p \in P$ with $q \in B(p, 1)$ then return* YES *and a point p' such that $q \in B(p', c)$,*
- *if $q \notin B(p, c)$ for all $p \in P$ then return* NO*,*
- *if for the point p closest to q we have $1 \leq d(q, p) \leq c$ then return either* YES *or* NO.

In [15] it was proved that given an algorithm for c-PLEB which uses $f(n)$ space on an instance of size n where f is convex, there is a data structure for $(c + \epsilon) - NNS$ problem requiring $O(f(n\text{poly}(\log n, 1/(c-1))))$ space and using $O(\text{poly}(\log n, 1/(c-1)))$ invocations to c-PLEB per query. Thus in this paper we will concentrate on solving the c-PLEB problem.

Superimposed codes. There exist several variants of superimposed codes. In this paper we assume the following definition (see [8]):

Definition 2 *An $N \times M$ binary matrix A is called a* superimposed (z, M)-code *(or (z, M)-code) of length N if the boolean sum of any z columns of A does not contain any other column. We refer to A's columns (which we denote by $A[0] \ldots A[M-1]$) as* codewords.

Notice that each codeword corresponds to a subset of $[M]$ (take the set of all coordinates set to 1) and therefore we can refer to codewords as sets.

Here, we are interested in the situation when M and z are given and the goal is to minimize N. Let $N_{\min}(M, z)$ denote the minimum length of any (z, M)-code. Dyachkov and Rykov [8] showed that $N_{\min}(z, M) = \Theta(z^2 \log_z M)$ (similar bounds were also obtained by Erdos, Frankl and Furedi [10]). However, their upper bound was obtained by a probabilistic argument and is non-constructive. The best explicit construction [18] (based on Reed-Solomon codes) achieves $N = O(z^2 \log_z^2 M)$.

3 Embeddings of Hausdorff metrics over normed spaces

We begin by showing a 1-embedding into l_∞. Our approximate embeddings will be based on this exact embedding. It is well known that any metric (X, d) can be 1-embedded into $l_\infty^{|X|}$. Thus any Hausdorff metric $H(d)$ can be 1-embedded into l_∞, with the number of dimensions equal to the number of sets. Here, we show that fewer dimensions suffice.

Theorem 1 *For any finite metric $D = (X, d)$ the space $H(D)$ can be 1-embedded into $l_\infty^{|X|}$.*

Proof: Assume $X = \{p_1, \ldots, p_n\}$. For any $S \subset X$, the value $f(S)$ is defined as

$$f(S) = (d(p_1, S), \ldots, d(p_n, S)).$$

Notice that this mapping is a contraction; therefore it is sufficient to show $|f(S) - f(S')|_\infty \geq 1$ for any $S, S' \subset X$ such that $d_D(S, S') \geq 1$. To this end note that if $d_D(S, S') = t$, then there exists $p \in S$ such that $d(p, S') = t$ (or $p' \in S'$ such that $d(p', S) = t$, we will assume the first case without loss of generality). Then

$$|f(S) - f(S')|_\infty \geq |d(p, S) - d(p, S')| \geq |0 - t| = t,$$

thus establishing the claim. ■

This theorem is complimented by the following:

Theorem 2 *Any finite metric $D = (X, l_\infty^d)$ can be 1-embedded into $H^d(l_p^1)$.*

The choice of p does not matter since all l_p norms are the same in 1 dimension.

Proof: Let Δ be the diameter of some set X of points in l_∞^d, that is, let $\Delta = \max_{i,j}\{l_\infty(x_i - x_j)\}$. Then the mapping of l_∞^d into $H^d(l_p^1)$ is as follows. Let $x_i = [x_{i,1}, \ldots, x_{i,d}]$. Then map x_i into the set $S_i = \{S_{i,1}, \ldots, S_{i,d}\}$, where $S_{i,j} = 2(j-1)\Delta + x_{i,j}$.

The points defined by dimension j are all between $2(j-1)\Delta$ and $(2j-1)\Delta$ and so, the distance from the $S_{i,j}$ to S_k is simply $|S_{i,j} - S_{k,j}|$. Thus the Hausdorff distance between S_i and S_j is simply $l_\infty(x_i - x_j)$. ■

Therefore, Hausdorff and l_∞ metrics are closely linked.

3.1 Upper bound, $H^s(l_\infty^d)$

In this section we modify the techniques used above to obtain embeddings which are *approximate* but which require smaller number of dimensions. More specifically:

Theorem 3 *For any $I > 1$, $s > 0$ and $0 < c < 1$ the space $H^s([0, I]_\infty^d)$ can be c-embedded into l_∞^E where*

$$E = O\left(ds^2 \left\lceil \frac{1+c}{1-c} \right\rceil^{2d} \left(\log I + \log \frac{1}{1-c}\right)\right).$$

Proof: As before, we need a mapping $f : H^s([0, I]_\infty^d) \to l_\infty^E$ such that for any $A, B \subset [0, I]^d$ of cardinality s and $p \in A$ such that $d(p, B) \geq 1$ we have $|f(A) - f(B)|_\infty \geq c$. The embedding will be defined by a sequence of E subsets $S_1 \ldots S_E$ of $[0, I]^d$ as

$$f(A) = (d(A, S_1), \ldots, d(A, S_E)).$$

In the following we will provide the sets $S_1 \ldots S_E$ such that for any A, B and q as above there exists S_i such that the following conditions are true:

- $d(p, S_i) \leq a$
- $d(B, S_i) \geq b$
- $a - b \geq c$

which is clearly sufficient for our purpose. To this end, we let $a = \frac{1-c}{2}$ and $b = \frac{1+c}{2}$. Note that $a + b = 1 \leq d_D(A, B)$. Impose a regular cubic grid on $\Re^d$ with a side $2a$. Among all cells intersecting $B(p, a)$ choose the one which contains p and call it $\overline{p}$. Notice that the center of $\overline{p}$ belongs to $B(p, a)$. Define $\overline{C}$ to be the set of all cubes intersecting $[0, I]^d$ and let $\overline{B}$ be the set of cubes whose centers belong to $\cup_{p \in B} B(p, b)$. Notice that $\overline{p} \notin \overline{B}$. Also, we can bound

- $|\overline{C}| \leq (\frac{I}{2a} + 2)^d$
- $|\overline{B}| \leq u(c) = s\left(\left\lceil \frac{1+c}{1-c} \right\rceil\right)^d$

By using superimposed codes we know there exists $E = u(c)^2 \log |\overline{C}|$ sets $\overline{S_1} \ldots \overline{S_E}$ such that for any p, B as above there exists $\overline{S_i}$ containing $\overline{p}$ but none of then elements from $\overline{B}$. We then construct S_i from $\overline{S_i}$ by replacing each grid cell by its center. These sets satisfy the above requirements. ■

Remark 1 *By using the probabilistic method one can in fact improve the dependence on c from $\left(\lceil \frac{1+c}{1-c} \rceil\right)^{2d}$ to $\left(\lceil \frac{1+c}{1-c} \rceil\right)^{d}$. This is due to the fact that the sets $\overline{B}$ are not arbitrary but have special structure. Unfortunately, we do not have any explicit construction which yields such a bound.*

Remark 2 *We need not use superimposed codes in the construction. If we pick each cell with probability $1/2u(c)$, then with probability $\Theta(1/u(c))$ we get a set S such that $d(p, S) \leq a$ and $d(B, S) \geq b$. Choosing $\Theta(u(c) \log^2 I^s)$ such sets gives, w.h.p., the needed code words.*

3.2 Upper bound, $H^s(l_k^d)$

Our embedding algorithm in this case closely follows that of $H^s(l_\infty)$. However, instead of having a grid of size $2a = 1 - c$, we have a grid of size $2a/d^{1/p}$. We need only bound $|\overline{B}|$ and $|\overline{C}|$, which we do as follows.

- $|\overline{C}| \leq (\frac{Id^{1/p}}{2a} + 2)^d$
- $|\overline{B}| \leq u_p(c) = s\left(\left\lceil \frac{1+c}{1-c} \right\rceil\right)^d$

The first bound is direct, and the second comes from [15]. We once again use superimposed codes and get $E_p = u_p(c)^2 \log |\overline{C}|$, thus achieving the following theorem:

Theorem 4 *For any $I > 1$, $s > 0$ and $0 < c < 1$ the space $H^s([0, I]_p^d)$ can be c-embedded into l_∞^E where*

$$E = O\left(ds^2 \left\lceil \frac{1+c}{1-c} \right\rceil^{2d} \left(\log I + \log \frac{1}{1-c} + \frac{\log d}{p}\right)\right).$$

3.3 Lower bound

In this section we prove the following lower bound theorem.

Theorem 5 *For $\frac{1}{2} = R < \alpha r \leq r = 1$ let f be an (r, α, R)-embedding of $H^s([-1/2, 1/2]^d_\infty)$ into l^E_∞. Then there exists a $(s-1, 2^d)$-code of length $O(\frac{E}{\alpha r - R})$.*

In particular, this implies that if $\frac{1}{\alpha r - R} = O(1)$ and $2^d \geq s^2$, then $E = \Omega(s^2 d / \log s)$.

Proof: Let D be the l_∞ metric over $[-1/2, 1/2]^d$ and assume the existence of f as above. For any element a from the universe $U = \{-1/2, 1/2\}^d$ we will define its codeword $C(a)$ in a sequence of steps. Firstly, let 0 denote the origin (i.e. the point $(0, \ldots, 0)$). For any $A \subset U$ we define:

- $L(A) = A \cup \{0\}$
- $R(A) = A$

Claim 1 *For any $A, B \subset U$ we have*

- *if $A \subset B$ then $d_D(L(A), R(B)) = \frac{1}{2}$*
- *otherwise $d_D(L(A), R(B)) = 1$*

Proof: Assume first that $A \subset B$. In this case for any $a \in A$ the distance from a to $R(B)$ is 0 (since $a \in B$ as well). Also, the distance from 0 to any point is at most $1/2$. On the other hand, the distance from any point in $R(B)$ to 0 (which belongs to $L(A)$) is $1/2$. Therefore, $d_D(L(A), R(B)) = \frac{1}{2}$.

On the other hand, assume that $A \not\subset B$. In this case there exists $a \in A - B$. The distance from a to any point in B is 1. This implies $d_D(L(A), R(B)) = 1$. ■

From the above claim and the properties of f we know that for any $A, B \subset U$ with cardinalities at most $s - 1$ we have

- if $A \subset B$ then $|f(L(A)) - f(R(B))|_\infty \leq R$
- otherwise $|f(L(A)) - f(R(B))|_\infty \geq \alpha r$.

Claim 2 *Let P be the set of all points $f(L(A))$ and $f(R(B))$ where A, B as above. Then the diameter of P is at most 2.*

Proof: Since the empty set is included in any set, we have $d_D(f(L(\emptyset)), f(R(B))) = 1/2$ for any B. Also, for any set A we have $A \cup \emptyset \subset A$, and therefore

$$d_D(f(L(A)), f(L(\emptyset))) \leq d_D(f(L(A)), f(R(A))) + d(f(R(A)), f(L(\emptyset))) \leq 1/2 + 1/2 = 1$$

Therefore (by triangle inequality) all pairwise distances are at most 2. ■

We can therefore assume that $P \subset [0, 2]^E$. Impose a uniform grid on $[0, 2]^E$ of side $\Delta = (\alpha r - R) - \epsilon$ for arbitrary $\epsilon > 0$. For any point $p \in [0, 2]^E$ let $g_i(p)$ be the ith coordinate of the cell containing p; for convenience we assume that the coordinate values for distinct i's are distinct. Then we define:

- $L'(A) = \{g_1(f(L(A))), \ldots, g_E(f(L(A)))\}$
- $R'(A) = \cup_i g_i(B(f(R(A)), R))$

Claim 3 *For any A and B as above we have $L'(A) \subset R'(B)$ iff $A \subset B$.*

For any $a \in U$ we define $C(a) = L'(\{a\})$. We will show that $C(a)$ forms an $(s-1, 2^d)$-code. Consider any a, B such that $a \notin B$. Notice that

- for each $b \in B$, we have $C(b) = L'(\{b\}) \subset R'(B)$. Therefore, $\cup_{b \in B} C(b) \subset R'(B)$.
- $C(a) = L'(\{a\}) \not\subset R'(B)$, as $a \notin B$.

Therefore $C(a) \not\subset \cup_{b \in B} C(b)$. Thus $C(a)$ is an $(s-1, 2^d)$ code. Moreover, let $U' = \cup_a C(a)$. It is easy to see that $|U'| = O(\frac{E}{\alpha r - R})$. The theorem follows. ■

4 Embeddings of Hausdorff metrics over general spaces

Let (X, d) be an arbitrary metric space. Then our method for embedding a Hausdorff of a normed space does not work anymore. The crucial problem is that $|B|$ cannot be bounded, that is, $|B(p, r)|/|B(p, r(1-\epsilon))|$ can be unbounded. Here, we give a randomized embedding procedure which is a modification of the randomized embedding given above. Our main modification will be that we select our reference sets via non-uniform sampling according to local properties of the metric.

In order to achieve an $1/\alpha$-Nearest Neighbor, we need to be able to generate (r, α)-embeddings, for many values of r. Instead, we will be able to generate (r', α)-embeddings, for some $r' < r$. Notice, that an (r', α)-embedding is (by definition) also an $(r, \frac{r'}{r}\alpha)$-embedding. Thus the dependence on α in the nearest neighbor search algorithm will become worse by a constant factor.

Theorem 6 *For any $\epsilon > 0$, $r > 0$ and $0 < \alpha < 1$, there is $r' < r$ such that $r' \geq \frac{r}{(1+\frac{2}{\alpha})^{1/\epsilon+1}}$ such that any metric X, $|X| = n$ can be $(r', 1-\alpha)$-embedded into l_∞^d with $d = O(s^2 n^\epsilon \log n/\epsilon)$.*

Proof: Let $\beta = \frac{\alpha+2}{\alpha}$ and let $r_i = r/\beta^i$ for $i = 0 \ldots 1/\epsilon + 1$. Our embedding will have a set of dimensions for each r_i. We will show that for each pair A, B, some dimension corresponding to some r_i will correctly approximately represent their Hausdorff distance. Since our embedding is into l_∞, we can simply concatenate all dimensions for all r_i in order to achieve our final embedding.

First note that, $\forall p \in X$ there is an $r(p) = r_i$ such that:

$$|B(p, \alpha r_i)| > 1/n^\epsilon \cdot |B(p, r_i(\alpha+2))| = 1/n^\epsilon \cdot |B(p, \alpha r_{i-1})|$$

since the volume of the ball around B can grow by a factor n^ϵ at most $1/\epsilon$ times. Now consider some $p \in A$ such that $d_D(A, B) = D(p, B)$, and let $r' = r(p)$ and $t = 2r'$. As before, we need a mapping $f : H^s([0, I]_\infty^d) \rightarrow l_\infty^E$ such that for A, B and $p \in A$ such that $d(p, B) \geq 1$ we have $|f(A) - f(B)|_\infty \geq c$. The embedding will be defined by a sequence of E subsets $S_1 \ldots S_E$ of X as

$$f(A) = (d(A, S_1), \ldots, d(A, S_E)).$$

Once again, we will provide the sets $S_1 \ldots S_E$ such that for any A, B as above there exists S_i such that the following conditions are true:

- $d(p, S_i) \leq \alpha r'$
- $d(B, S_i) \geq r'$

Call the reference set S_i a *witness* to A, B if it satisfies these conditions. We will select points to go into the reference sets according to the density of their neighborhood, that is, node v is selected with probability $P(v) = 1/s|B(v,t)|$. Let $P(A)$ be the probability of selecting some element in $A \subset X$. We show the following:

Claim 4 *The probability that S_i is a witness to A, B is $\Omega(1/sn^{\epsilon})$*

Proof of Claim: The proof proceeds by showing that $P(B(p, \alpha r')) = \Omega(1/sn^{\epsilon})$ and $1 - P(\cup_{q \in B} B(q, r')) = \Omega(1)$. Consider $P(B(p, \alpha r'))$ first. We know that for all $x \in B(p, \alpha r')$, we have $B(x,t) \subseteq B(p, t + \alpha r')$, so $|B(x,t)| \leq |B(p, r'(\alpha + 2))| \leq |B(p, \alpha r')| n^{\epsilon}$. Therefore $P(x) = 1/s|B(x,t)| \geq 1/sn^{\epsilon}|B(p, \alpha r')|$, and thus $P(B(p, \alpha r'))$ is $\Omega(1/sn^{\epsilon})$.

Consider now $1 - P(\cup_{q \in B} B(q, r'))$. Let $q \in B(q, r')$. Since $t = 2r'$,for any $x \in B(q, r')$, we know that $B(q, r') \subset B(x,t)$. Therefore $|B(x,t)| \geq |B(q, r')|$, and $P(x) \leq 1/s|B(q, r')|$. We get $P(B(q, r')) \leq 1/s$. The probability that $d(B, S_i) \geq r'$ is then $\Omega(1)$.

Therefore, we get both events with probability $\Omega(1/sn^{\epsilon})$, as required. □

In order to have a witness for all $n^{O(s)}$ pairs A and B such that $r(p) = r'$ we need to repeat the above procedure $O(s \log n)$ times. Taking all $1/\epsilon$ values of r' gives a total of $O(s^2 n^{\epsilon} \log n / \epsilon)$ dimensions. ■

5 Approximate Nearest Neighbor Problem

In this section we first describe how to apply the embedding results proved so far to achieve a fast algorithm for the Approximate Nearest Neighbor problem. Then we point out how to generalize our algorithms to work for the minimum Hausdorff distance under isometries.

Algorithms. One can easily observe the following fact.

Fact 1 *If X is can be β-embedded in Y and there is a c-PLEB algorithm for Y, then there is a c/β-PLEB algorithm for X.*

Thus we can plug in our embedding result of Theorems 3 and 4 to the ϵ-PLEB algorithm for l_∞ and obtain algorithms for searching in $H^s(l_p^d)$ as stated in the Introduction. In principle we could also do the same for $H^s(X)$ using the result from the Section 4. However, a direct approach would require advance knowledge of all queries, as we would need to embed all points (including query points) during the preprocessing. In order to avoid this problem, we show that embedding of a query point q can be done quickly "on-line" provided we can perform approximate nearest neighbor query in the underlying metric X. To this end, observe that the proof of Theorem 6 is still valid if the embedding f is computed use c-approximations of the distances $d(A, S_i)$ instead of real distances; the only difference is that we obtain a $(r', 1 - c\alpha)$-embedding instead of $(r', 1 - \alpha)$ one. Thus we can reduce one approximate nearest neighbor query in $H^s(X)$ to $s^2 n^{\epsilon} \log n / \epsilon$ approximate queries in X (one for each set S_i) plus one query in $l_\infty^{s^2 n^{\epsilon} \log n / \epsilon}$.

Isometries. When the underlying metric is l_2, the above results can be easily extended to the Hausdorff metric under translations and rotations. To this end, we apply the following algorithm. For each set S_i from the database compute its centroid c_i and then translate all sets such that their centroids overlap (say at some point c). Moreover, for any i let $p_i \in S_i$ be the point in S_i with the largest distance from c_i. Rotate each S_i around c_i such that the vector $p_i - c_i$ is parallel to the X axis. Then build a nearest neighbor data structure DS for the resulting sets S_i.

In order to process the query set S, align its centroid to c. Then, for each $p \in S$, rotate S around c such that the vector $p - c$ is parallel to X axis and query the data structure DS with the rotated S as an argument. Return the answer with the smallest distance from S.

The correctness of these procedure (i.e. the fact that it returns an $O(1)$-approximate nearest neighbor w.r.t. Hausdorff distance under translations and rotations) follows from the results of [2, 1] and we omit the proof here. As mentioned in [2], by exploring $O(1/\epsilon^2)$ points close to c_i and trying $O(1/\epsilon)$ different rotations, one can guarantee that the approximation guarantee is $(1+\epsilon)\cdot C$, where C is the approximation guarantee for the static Hausdorff data structure; this increases the number of queries by a factor of $O(1/\epsilon^3)$.

6 Extensions

The embeddings of $H(l_p^d)$ shown earlier has the property that the dimensionality of the l_∞ space grows exponentially in the dimension of the space underlying the Hausdorff metric. One possible way to avoid this problem could be to give an embedding of $H^s(l_p^d)$ into (say) $H^{s'}(l_p^{\log s})$, i.e. reduce the dimensionality of the underlying space, thus replacing the factor exponential in d by $s^{O(1)}$. By the Johnson-Lindenstrauss Lemma [17] such an embedding is indeed possible for the special case $s = 1$ and $p = 2$, thus it might be possible to prove it for general s. Unfortunately, we are not aware of any such result. However, we can prove a slightly weaker but similar result for $H_1^s(l_2^d)$.

Theorem 7 *For any $n > 1$ there exists a family F of functions $f : H_1^s(l_2^d) \rightarrow H_1^{n^{O(1/\log s)}}(l_2^{O(\log s)})$ such that for any pair of sets $A, B \in H_1^s(l_2^d)$ if we choose f uniformly at random from F, then the probability that the distance between A and B is within a constant factor of the distance between $f(A)$ and $f(B)$ is at least $1/n$.*

Notice that this theorem would be sufficient to obtain a sublinear time approximate nearest neighbor algorithm for $H_1^s(l_2^d)$, provided the embedding theorem from the previous section held for H_1 (we do not have a proof that they do).

The idea of the proof is to create $v = n^{O(1/\log s)}$ mappings $f_1 \ldots f_v$, of the form $f_i : H_1^s(l_2^d) \rightarrow H_1^s(l_2^{O(\log s)})$. Each such mapping is induced by a random projection of l_2^d onto $l_2^{O(\log s)}$ as in the proof of the Johnson-Lindenstrauss Lemma. The function f is then obtained by placing all sets $f_i(A)$ in the same space $l_2^{O(\log s)}$ but sufficiently far away from each other so that there is no "interaction" between them when computing the distances. As the result does not yet have any algorithmic applications, we defer the full proof to the final version of this paper.

Acknowledgments

The authors would like to thank Sampath Kannan and Jiri Matousek for helpful comments and discussions.

References

[1] H. Alt, O. Aichholzer, and G. Rote. Matching shapes with a reference point. *Proceedings of the 10th Annual Symposium on Computational Geometry*, 1994.

[2] H. Alt, B. Behrends, and J. Blomer. Approximate matching of polygonal shapes. *Annals of Mathematics and Artificial Intelligence*, 13(3–4):251–65, 1995.

[3] H. Alt and L. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. a survey. Technical Report B96-11, Freie Universität Berlin, December 1996.

[4] J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

[5] David Cardoze and Leonard Schulman. Pattern matching for spatial point sets. *Proc. of the 39th IEEE Annual Symp. on Foundation of Computer Science*, 1998.

[6] L. Chew, M.T. Goodrich, D.P. Huttenlocher, K. Kedem, J.M. Kleinberg, and D. Kravets. Geometric pattern matching under euclidean motion. In *Proceedings of the Fifth Canadian Conference on Computational Geometry*, pages 151–156, 1993.

[7] L. Chew, M.T. Goodrich, D.P. Huttenlocher, K. Kedem, J.M. Kleinberg, and D. Kravets. Geometric pattern matching under euclidean motion. *Proceedings of the Fifth Canadian Conference on Computational Geometry*, pages 151–156, 1993.

[8] A. G. Dyachkov and V. V. Rykov. A survey of superimposed code theory. *Problems of Control and Information Theory*, 12(4), 1983. English translation.

[9] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Spinger Verlag, 1987.

[10] P. Erdos, P. Frankl, and Z. Furedi. Families of finite sets in which no set is covered by the union of r others. *Israel Journal of Mathematics*, 51:79–89, 1985.

[11] P. Finn, D. Halperin, L. E. Kavraki, J. C. Latombe, R. Motwani, C. Shelton, and S. Venkatasubramanian. Geometric manipulation of flexible ligands. *LNCS Series - 1996 ACM Workshop on Applied Computational Geometry*, 1148:67–78, 1996.

[12] P. Finn, L. E. Kavraki, J. C. Latombe, R. Motwani, C. Shelton, S. Venkatasubramanian, and A. Yao. Rapid: Randomized pharmacophore identification for drug design. In *Proceedings of the Thirteenth Annual ACM Symposium on Computational Geometry*, 1997.

[13] D. P. Huttenlocher, K. Kedem, and J. M. Kleinberg. On dynamic voronoi diagrams and the minimum Hausdorff distance for points sets under euclidean motion in the the plane. *Proceedings of the Eighth Annual ACM Symposium on Computational Geometry*, pages 110–120, 1992.

[14] Piotr Indyk. On approximate nearest neighbors in non-euclidean spaces. *Proc. of the 39th IEEE Annual Symp. on Foundation of Computer Science*, 1998.

[15] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proc. of the 30th Ann. ACM Symp. on Theory of Computing*, 1998.

[16] Piotr Indyk, Rajeev Motwani, and Suresh Venkatasubramanian. Geometric matching under noise: Combinatorial bounds and algorithms. *Proc. of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[17] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[18] W. H. Kautz and R. C. Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10:363–377, 1964.

[19] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *Proc. of the 30th Ann. ACM Symp. on Theory of Computing*, 1998.

[20] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Proc. of the 35th IEEE Annual Symp. on Foundation of Computer Science*, pages 577–591, 1994.

[21] Jiri Matousek. On embedding expanders into l_p spaces. *Israel Journal of Mathematics*, 1994.

[22] D. Mount, N. Netanyahu, and J. LeMoigne. Improved algorithms for robust point pattern matching and applications to image registration. In *Fourteenth ACM Symposium on Computational Geometry*, June 1998.

[23] R. Norel, D. Fischer, H. J. Wolfson, and R. Nussinov. Molecular surface recognition by a computer vision-based technique. *Protein Engineering*, 7(1):39–46, 1994.

[24] W.T. Rucklidge. Lower bounds for the complexity of the hausdorff distance. *Proceedings of the Fifth Canadian Conference on Computational Geometry*, pages 145–150, 1992.