Embedding and Similarity Search for Point Sets under Translation*

Minkyoung Cho
Department of Computer Science
University of Maryland
College Park, Maryland, USA
minkcho@cs.umd.edu

David M. Mount
Department of Computer Science and
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland, USA
mount@cs.umd.edu

ABSTRACT

Pattern matching in point sets is a well studied problem with numerous applications. We assume that the point sets may contain outliers (missing or spurious points) and are subject to an unknown translation. We define the distance between any two point sets to be the minimum size of their symmetric difference over all translations of one set relative to the other. We consider the problem in the context of similarity search. We assume that a large database of point sets is to be preprocessed so that given any query point set, the closest matches in the database can be computed efficiently. Our approach is based on showing that there is a randomized algorithm that computes a translation-invariant embedding of any point set of size at most n into the L_1 metric, so that with high probability, distances are subject to a distortion that is $O(\log^2 n)$.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*Pattern matching*

General Terms

Algorithms, Theory

Keywords

Point pattern matching, similarity search, embedding, randomized algorithms.

1. INTRODUCTION

Geometric pattern matching is a well studied computational problem with a wide variety of formulations and applications. The most common formulation considered in computational geometry involves determining the degree of similarity between two given point sets, subject to some group of allowable geometric transformations. The literature on this topic is vast. The interested reader is referred to the survey by Alt and Guibas [2]. An important formulation for many modern applications involves viewing the problem from the perspective of similarity search. We are given a large database of point sets, which is to be preprocessed so that, given a query set, it is possible to efficiently compute its closest neighbor(s) in the database.

We consider this problem in a relatively simple context but one that still leads to quite an interesting computational problem. We assume that point sets have integer coordinates, that they are to be matched subject to an unknown translation, and that there is a significant fraction of outliers, that is, points from one set may not match any point of the other set. We assume, however, the points that do match (subject to the optimum translation) identically. This is in contrast with measures such as the partial Hausdorff distance [11], where both outliers and near misses are tolerated. Outliers are challenging because global properties of the point sets, based for example on the identification of reference points such as centroids [1] are not applicable. The distance metric we use is the size of the symmetric difference of the two point sets, which is to be minimized through some translation of one set relative to the other. (Formal definitions are given below.)

Our approach is based on finding an embedding function that maps a point set into a metric space [12]. The distortion of such an embedding is the maximum multiplicative variation that distances might suffer in the mapping process. The objective is to produce an embedding of low distortion into a space in which similarity search can be performed efficiently (in particular, its dimension should be low).

In the context of database search for point sets undergoing transformations, a well-known solution is based on geometric hashing [20] which is to store all possible transformed sets so that the search will be processed efficiently. This approach is notably inefficient when outliers are present. A common approach in computer vision is to compute some property that is invariant under motion. A popular choice is the histogram of colors appearing in the image. These can be compared using, for example, the Earth-Mover's distance (EMD) [2,6]. The problem of computing low distortion embeddings of variants of EMD has been studied by Indyk and Thaper [13] and Wang, *et al.* [19]. Unfortunately, similarity of two color histograms provides no guarantees on the similarity of the

*This work was supported by the National Science Foundation under grant CCF-0635099.

SCG'08, June 9–11, 2008, College Park, Maryland, USA.

underlying images. The distance histogram is one example of a more general approach that is widely studied in computer vision based on computing some invariant feature of the underlying objects [5,15]. Although some invariants can provide good practical performance, they generally cannot provide guarantees on the quality of the match. One popular example is the set of inter-point distances, called the *distance histogram* [3, 8]. Although one would expect that similar sets would have similar distance histograms, there exist so called *homometric sets* [16,18], which have identical distance histograms but which may not be at all similar to one another. For example, the following two sets are homometric: $P = \{0, 1, 4, 10, 12, 17\}$ and $Q = \{0, 1, 8, 11, 13, 17\}$.

Our objective here is to compute a translation invariant of guaranteed performance. More formally, consider a point set consisting of at most n points on the d-dimensional integer grid, where d is a constant. We assume that the coordinates of each point are bounded above by a polynomial function of n. As usual, let $\mathbb{Z}$ denote the set of integers, and let $\mathbb{Z}_u$ denote $\{0, 1, 2, \ldots, u-1\}$. (We do not assume that u is prime.) Let $\mathbb{Z}^d$ denote the set of d-element vectors over $\mathbb{Z}$, and define $\mathbb{Z}_u^d$ analogously for $\mathbb{Z}_u$. Let $\mathbb{Z}_u^d(\leq n)$ denote the collection of point sets over $\mathbb{Z}_u^d$ that contain at most n points. Given two finite sets P and Q, let $P \Delta Q$ denote their *symmetric difference*, that is,

$$P \Delta Q \;=\; (P \setminus Q) \cup (Q \setminus P).$$

The cardinality of the symmetric difference is a well known metric on finite sets, which we denote by $|P \,\Delta\, Q|$. This can be generalized to multisets by summing the absolute differences of the multiplicities of corresponding elements.

Given a point set P and any $t \in \mathbb{Z}^d$, the *translate* $P + t$ is defined to be $\{p + t \mid p \in P\}$. Extending the symmetric difference, we define the *symmetric difference distance under translation*, denoted $\langle P \,\Delta\, Q \rangle$, to be

$$\langle P \,\Delta\, Q \rangle \;=\; \min_{t \in \mathbb{Z}^d} |(P + t) \,\Delta\, Q| \,.$$

It is easy to verify that this is a metric. (See Lemma 2.1 below.) Throughout, we will assume that P and Q are taken from $\mathbb{Z}_u^d(\leq n)$.

Let ℓ_1^d denote metric space consisting of real d-dimensional space $\mathbb{R}^d$ endowed with the L_1 metric. Given $x, y \in \ell_1^d$, we denote their L_1 distance by $\|x - y\|_1$. We use the terms *randomized embedding* and *randomized function* throughout to denote a function computed by a randomized algorithm that satisfies the given probability bounds. We also use log to denote logarithm base 2 and ln to denote the natural logarithm. Our main result is presented below.

Theorem 1. *Given sufficiently large integers n and u, where $u \leq n^{O(1)}$, a constant d, and failure probability β, there exists a randomized embedding $\Psi \colon \mathbb{Z}_u^d(\leq n) \to \ell_1^m$, where $m = O(n \log^2 n \log(1/\beta))$, such that for any $P, Q \in \mathbb{Z}_u^d(\leq n)$:*

(i) $\|\Psi P - \Psi Q\|_1 \leq (2 \log n)\, \langle P \,\Delta\, Q \rangle$.

(ii) $\|\Psi P - \Psi Q\|_1 \geq \frac{1}{17 \log n}\, \langle P \,\Delta\, Q \rangle$, *with probability at least* $1 - \beta$, *and*

This embedding can be computed in time $O(n \log^4 n \log(1/\beta))$.

Note that part (i) of the above theorem holds irrespective of randomization. It follows that the resulting embedding achieves a distortion of at most $34 \log^2 n$ (with probability at least $1 - \beta$).

We know of no prior work on this problem. In a 1-dimensional discrete setting, this problem is related to a version of edit distance on bit strings, where the number of replacements corresponds to the symmetric difference in the sets, but it is possible to shift one set relative to the other without cost. The most closely related work to ours is that of Cormode and Muthukrishnan [7], on embedding strings under edit distance with moves. They present an embedding with distortion $O(\log n \log^* n)$ into L_1 with an exponential number of dimensions.

Here is a brief outline of our algorithm. First, we observe that it is possible to reduce our problem to one involving 1-dimensional point sets of over $\mathbb{Z}_{u'}$, where $u' = n^{O(d)}$. We then apply a low distortion linear hash function, which maps the point set to $\mathbb{Z}_s$, where $s = O(n \log n)$. The linearity of the hash function is critical, since it preserves distances under translation. Such a set can be viewed as a bit-vector in $\mathbb{Z}_2^s$. In order to obtain an embedding that is invariant under translations, we select various sized *probes*, each of which is a random subset of $\mathbb{Z}_s$. The application of a probe of size ρ to a single placement of the vector produces an integer in $\mathbb{Z}_{2^\rho}$, where this integer is based on the bit pattern appearing in the bit-vector at each of the probed positions. We apply the probe at each of the s positions, and take the union of the probe results. The result can be viewed as a vector in $\mathbb{Z}_{s+1}^{2^\rho}$. We then apply another hash function to reduce the dimension from $\mathbb{Z}_{s+1}^{2^\rho}$ to $\mathbb{Z}_{s+1}^{O(s)}$. We show that, if ρ is chosen in a manner that is sensitive to the actual distance between the original point sets, then the distance between these vectors is related to the distance between the original point sets under translation. Since this distance is not known, we apply the construction to a series of exponentially increasing distance estimates. We show that, through an appropriate weighting of the components of this series, we obtain the desired distortion bounds in expectation. In order to produce results that apply with high probability, we repeat the process some number of times, using different random hash functions and different random probes.

2. PRELIMINARIES

Recall that a nonnegative function $d(P, Q)$ is a *metric* if $d(P, P) = 0$, $d(P, Q) = d(Q, P)$ and $d(P, R) \leq d(P, Q) + d(Q, R)$. The last condition is the well known *triangle inequality*.

Lemma 2.1. *The symmetric distance under translation is a metric.*

Proof. It is easy to see that the first two requirements of a metric hold. To establish the triangle inequality, let t_1, t_2, and t_3 denote the optimal translations for $\langle P \,\Delta\, Q \rangle$, $\langle Q \,\Delta\, R \rangle$, and $\langle P \,\Delta\, R \rangle$, respectively. We have

$$\begin{aligned}
&\langle P \,\Delta\, Q \rangle + \langle Q \,\Delta\, R \rangle - \langle P \,\Delta\, R \rangle \\
&\quad = \; |(P + t_1) \,\Delta\, Q| + |(Q + t_2) \,\Delta\, R| - |(P + t_3) \,\Delta\, R| \\
&\quad = \; |(P + t_1) \,\Delta\, Q| + |Q \,\Delta\, (R - t_2)| - |(P + t_3) \,\Delta\, R| \\
&\quad \geq \; |(P + t_1) \,\Delta\, Q| + |Q \,\Delta\, (R - t_2)| - |(P + t_1) \,\Delta\, (R - t_2)| \\
&\quad \geq \; 0,
\end{aligned}$$

where the last implication follows from the triangle inequality for symmetric difference (which is well known to be a metric). □

Next, we observe that the problem of computing distances for point sets in the d-dimensional space $\mathbb{Z}_u^d$ under translation can be reduced to computing distances under translation in a 1-dimensional space $\mathbb{Z}_{O(u^d)}$. The result is based on the simple observation that we can unravel the d-dimensional grid into a sufficiently large 1-dimensional grid to avoid wrap-around effects. Since the mapping is linear it preserves similarity under translation.

Lemma 2.2. *Consider a positive integer u and constant d. There exists a function $g\colon \mathbb{Z}_u^d \to \mathbb{Z}_{u'}$, where $u' = O(u^d)$ such that for any sets $P, Q \subseteq \mathbb{Z}_u$ we have $\langle gP \,\Delta\, gQ\rangle = \langle P \,\Delta\, Q\rangle$. This function is computable in $O(1)$ time (assuming that arithmetic operations on numbers of magnitude $O(u)$ can be computed in constant time).*

The proof of Lemma. 2.2 follows directly from the next two lemmas. Given integers $a \le b$, let $[a, b]$ denote the set of integers from a to b, and let $\pm\mathbb{Z}_u$ denote $[1-u, u-1]$.

Lemma 2.3. *For $v = (v_0, v_1, \ldots, v_{d-1}) \in [1-u, 2u-1]^d$, let $g(v) = v_0 + v_1(3u) + v_2(3u)^2 + \ldots + v_{d-1}(3u)^{d-1}$. Then, g is linear and injective.*

Proof. It is obvious that g is linear, that is, $g(v + v') = g(v) + g(v')$.

Now, we show that g is an injective function over the domain $[1-u, 2u-1]^d$. Suppose to the contrary that there existed x and x' from this domain such that $x \neq x'$ and $g(x) = g(x')$. Let i denote the largest coordinate index such that $x_i \neq x'_i$. Then,

$$\begin{aligned} g(x) - g(x') &= \sum_{j=0}^{d-1} x_j(3u)^j - \sum_{j=0}^{d-1} x'_j(3u)^j \\ &= \sum_{j=0}^{i} x_j(3u)^j - \sum_{j=0}^{i} x'_j(3u)^j = 0. \end{aligned}$$

Clearly, $|x_i - x'_i| \le 3u - 1$, and so

$$\begin{aligned} \left|(x_i - x'_i)(3u)^i\right| &= \left|\sum_{j=0}^{i-1} x_j(3u)^j - \sum_{j=0}^{i-1} x'_j(3u)^j\right| \\ &\le \sum_{j=0}^{i-1}(3u-1)(3u)^j \;=\; (3u)^i - 1. \end{aligned}$$

Since $x_i \neq x'_i$, the left side is at least $(3u)^i$ which yields the desired contradiction. □

Lemma 2.4. *Given P, Q and g as defined above*

$$\langle P \,\Delta\, Q\rangle = \langle gP \,\Delta\, gQ\rangle .$$

Proof. Let $t \in \mathbb{Z}^d$ denote an optimal translation between P and Q. We observe that each coordinate t_i satisfies $1 - u \le t_i \le u - 1$ since an optimal translation will succeed in aligning at least one point. Thus, we can assume that $t \in \pm\mathbb{Z}_u^d$. Analogously, let t_g denote an optimal translation between gP and gQ. Similarly, we can assume that $t_g \in \pm\mathbb{Z}_{(3u)^d}$. In order to prove that $\langle P \,\Delta\, Q\rangle = \langle gP \,\Delta\, gQ\rangle$ it suffices to show that

$$\min_{t \in \pm\mathbb{Z}_u^d} |(P + t) \,\Delta\, Q| \;=\; \min_{t_g \in \pm\mathbb{Z}_{(3u)^d}} |(gP + t_g) \,\Delta\, gQ| .$$

Let G denote a set of images from the valid translations, that is, $G = \{g(t) \mid\in \pm\mathbb{Z}_u^d\}$. Let $\overline{G}$ denote the complement of G, that is, $\overline{G} = \pm\mathbb{Z}_{(3u)^d} \setminus G$. For any $t_g \in \pm\mathbb{Z}_{(3u)^d}$ we consider two cases, $t_g \in G$ and $t_g \in \overline{G}$.

If $t_g \in G$ then there exists t such that $t_g = g(t)$. From Lemma. 2.3, since g is linear and injective on $[1-u, 2u-1]^d$, we have

$$\begin{aligned} |(P + t) \,\Delta\, Q| &= |g(P + t) \,\Delta\, gQ| \\ &= |(gP + g(t)) \,\Delta\, gQ| \\ &= |(gP + t_g) \,\Delta\, gQ| . \qquad (1) \end{aligned}$$

Otherwise, if $t_g \in \overline{G}$, then there is no match between $gP + t_g$ and gQ. Suppose to the contrary that there existed at least one matched pair $(g(p), g(q))$ under translation t_g, that is,

$$g(q) \;=\; g(p) + t_g.$$

Clearly $p \in P$ and $q \in Q$ are the preimages of $g(p)$ and $g(q)$, respectively. Then there exists $t' \in \pm\mathbb{Z}_u^d$ such that $q = p + t'$. By linearity of g (from Lemma 2.3),

$$g(q) \;=\; g(p + t') \;=\; g(p) + g(t'),$$

and $t_g = g(t') \in G$, a contradiction.

Since the optimal translation has at least one matched point, the optimal translation t_g for gP and gQ is in G. Combining this with (1), we have

$$\begin{aligned} \langle gP \,\Delta\, gQ\rangle &= \min_{t_g \in \pm\mathbb{Z}_{(3u)^d}} |(gP + t_g) \,\Delta\, gQ| \\ &= \min_{t_g \in G} |(gP + t_g) \,\Delta\, gQ| \\ &= \min_{t \in \pm\mathbb{Z}_u^d} |(P + t) \,\Delta\, Q| \quad \text{by (1)} \\ &= \langle P \,\Delta\, Q\rangle . \end{aligned}$$

□

Next we present a couple of utility results. The first is a straightforward observation that a randomized function with a low collision probability produces a low distortion with respect to the size of the symmetric difference.

Lemma 2.5. *Let $0 \le \gamma \le 1$, and suppose that we are given a randomized function $h\colon \mathbb{Z} \to \mathbb{Z}$ such that for all distinct $x, y \in \mathbb{Z}$, $\Pr[h(x) = h(y)] \le \gamma$. We are also given a positive integer n, multisets $P, Q \in \mathbb{Z}(\le n)$, and failure probability β. Then*

(i) $|hP \,\Delta\, hQ| \le |P \,\Delta\, Q|$, and

(ii) $|hP \,\Delta\, hQ| \ge \left(1 - \frac{2n\gamma}{\beta}\right) |P \,\Delta\, Q|$ with probability at least $1 - \beta$.

Proof. Part (i) is trivial, since applying any function (which need not be 1–1) can only decrease the size of the symmetric difference. To prove (ii), let $\delta = |P \,\Delta\, Q|$. Define a collision to be any distinct pair $x, y \in \mathbb{Z}$, such that $h(x) = h(y)$. Observe that this collision can affect the size of the symmetric difference only if both x and y are in $P \cup Q$, and at least one is in $P \Delta Q$.

Let K denote the distinct elements in $P \cup Q$, and let k be $|K|$. Observe that $k \le 2n$. Let us consider a case for a fixed element x in K. Let δ_x denote the absolute difference between the multiplicities of x in P and Q. Let $I_{x,y}$ denote a random variable whose value is 1 if $h(x) = h(y)$ and 0 otherwise. Let

$$C_x \;=\; \sum_{y \in K, x \neq y} \min(\delta_x, \delta_y) I_{x,y}.$$

Then,

$$\mathrm{E}[C_x] \;=\; \sum_{y\in K, x\neq y} \min(\delta_x,\delta_y)\Pr[h(x)=h(y)] \;\le\; k\delta_x\gamma.$$

Each collision involving x can cause the symmetric difference to decrease by at most δ_x. (This occurs, for example, if $x \in P\backslash Q$ or $x \in Q\backslash P$ and $y \in P\cup Q$, causing the mismatches to disappear.) It is easy to see that

$$|P\,\Delta\,Q| - |hP\,\Delta\,hQ| \le \sum_{x\in P\Delta Q} C_x.$$

Thus, we have

$$\begin{aligned}&\mathrm{E}[\,|P\,\Delta\,Q| - |hP\,\Delta\,hQ|\,]\\ &\quad\le \sum_{x\in P\Delta Q}\mathrm{E}[C_x] \;\le\; k\gamma\sum_{x\in P\Delta Q}\delta_x \;=\; k\gamma\delta \;\le\; 2n\gamma\delta.\end{aligned}$$

By Markov's inequality,

$$\Pr\left[|P\,\Delta\,Q| - |hP\,\Delta\,hQ| \;\ge\; \frac{2n\gamma\delta}{\beta}\right] \;\le\; \beta.$$

Recalling that $\delta = |P\,\Delta\,Q|$, it follows that with probability at least $(1-\beta)$, we have

$$|hP\,\Delta\,hQ| \;\ge\; \left(1-\frac{2n\gamma}{\beta}\right)|P\,\Delta\,Q|,$$

as desired. □

The following lemma will be useful for compressing space. As mentioned in the introduction, because of the need to maintain invariance under translation, we will make use of a hash function that is linear, and in particular $h(x+t) = h(x)+h(t)$, for all x and t. Common choices for low-collision hash functions (such as the universal hash function $h(x) = (ax+b \mod u) \mod s$) do not satisfy this condition. Our next lemma presents such a function for the domain of values in interest. Note that a similar approach has also been used in the context of string pattern matching [14].

Lemma 2.6. *Consider positive integers n and u, where $u \le n^c$ for some constant $c \ge 1$. For all α and β, where $0 \le \alpha,\beta \le 1$, there exists $s = \Theta((n\log n)/(\alpha\beta))$ (with constant factors depending on c) and a randomized linear function $h\colon \mathbb{Z}_u \to \mathbb{Z}_s$ such that for any sets $P, Q \in \mathbb{Z}_u(\le n)$ we have*

(i) $|hP\,\Delta\,hQ| \le |P\,\Delta\,Q|$, *and*
(ii) $|hP\,\Delta\,hQ| \ge (1-\alpha)\,|P\,\Delta\,Q|$ *with probability at least* $1-\beta$.

The function h is computable in $O(1)$ time (assuming that arithmetic operations on numbers of magnitude $\max(n, 1/(\alpha\beta))$ can be computed in constant time).

Proof. As before (i) is trivial, and so we concentrate on proving (ii). Let $f = 6 + (8c/(\alpha\beta))$ and let $r = fn\ln(fn)$. Let R denote the set of prime numbers in the range $n\ln n$ to r. The Prime Number Theorem (see, e.g., [17]) implies that, for all sufficiently large n the number of primes less than or equal to n is at least $n/(2\ln n)$ and at most $3n/(2\ln n)$. It follows that for all sufficiently large n we have

$$\begin{aligned}|R| \;&\ge\; \frac{fn\ln(fn)}{2\ln(fn\ln(fn))} - \frac{3n\ln n}{2\ln(n\ln n)} \;\ge\; \frac{fn\ln(fn)}{4\ln(fn)} - \frac{3n}{2}\\ &\ge\; \left(\frac{f-6}{4}\right)n = \frac{2c}{\alpha\beta}n.\end{aligned}$$

For any $s \in R$, define $h_s(x) = x \mod s$. Linearity follows since $h_s(x+t) = (x+t) \mod s = (x \mod s)+(t \mod s) = h_s(x) + h_s(t)$ (where addition before the mapping is done over $\mathbb{Z}_u$ and addition after mapping is done over $\mathbb{Z}_s$). For any distinct $x, y \in \mathbb{Z}_u$, observe that $h_s(x) = h_s(y)$ if and only if $|x-y|$ has s as a factor. Since $x, y \le u \le n^c$, and $s \ge n$, it follows that this can be true for at most c choices of s. We define $h(x)$ to be $h_s(x)$, where s is a random element of R. Observe that for any fixed $x, y \in \mathbb{Z}_u$,

$$\Pr[h(x)=h(y)] \;\le\; \frac{c}{|R|} \;=\; \frac{\alpha\beta}{2n}.$$

By applying Lemma 2.5(ii) with $\gamma = \alpha\beta/(2n)$ and failure probability β, it follows that

$$\begin{aligned}|hP\,\Delta\,hQ| \;&\ge\; \left(1-\frac{2n\gamma}{\beta}\right)|P\,\Delta\,Q|\\ &=\; \left(1-\frac{2n(\alpha\beta/(2n))}{\beta}\right)|P\,\Delta\,Q| \;=\; (1-\alpha)\,|P\,\Delta\,Q|\end{aligned}$$

holds with probability at least $(1-\beta)$, as desired. □

The linearity of the hash function implies that the results of the previous lemma hold under translation.

Corollary 2.1. *The results of Lemma 2.6 apply to the symmetric difference distance under translation. That is,*

(i) $\langle hP\,\Delta\,hQ\rangle \le \langle P\,\Delta\,Q\rangle$, *and*
(ii) $\langle hP\,\Delta\,hQ\rangle \ge (1-\alpha)\,\langle P\,\Delta\,Q\rangle$ *with probability* $\ge 1-\beta$.

Proof. It is easy to see that for every $x \in \mathbb{Z}_s$, $h(x) = x$, and therefore the preimage $h^{-1}(\mathbb{Z}_s)$ is $\mathbb{Z}_u$. Combining this with the linearity of h and Lemma 2.6(i) we have

$$\begin{aligned}\langle hP\,\Delta\,hQ\rangle \;&=\; \min_{t\in\mathbb{Z}_s}|(hP+t)\,\Delta\,hQ|\\ &=\; \min_{h(t)\in\mathbb{Z}_s}|(hP+h(t))\,\Delta\,hQ|\\ &=\; \min_{t\in h^{-1}(\mathbb{Z}_s)}|(h(P+t)\,\Delta\,hQ|\\ &\le\; \min_{t\in\mathbb{Z}_u}|(P+t)\,\Delta\,Q| \;=\; \langle P\,\Delta\,Q\rangle.\end{aligned}$$

Part (ii) follows analogously. □

The following lemma will be applied in a context where the linearity of the hash function is not required, and this additional flexibility allows us to remove the restriction in the size of u. It follows by applying Lemma 2.5 to a suitable universal hash function [4]. The choice of the hash function given in the proof will be discussed later when we consider execution times. This lemma applies more generally to multisets. We may readily generalize the symmetric difference distance to multisets by counting not just the number of mismatched elements, but counting the absolute differences in the multiplicities of each distinct element. We extend the definition of $\mathbb{Z}_u(\le n)$ to include multisets in which the total cardinality (counting multiplicities) is at most n.

Lemma 2.7. *Consider positive integers n and u. For all α and β, where $0 \le \alpha,\beta \le 1$, there exists $s = O(n/(\alpha\beta))$ and a randomized function $h\colon \mathbb{Z}_u \to \mathbb{Z}_s$ such that for any multisets $P, Q \in \mathbb{Z}_u(\le n)$ we have*

(i) $|hP\,\Delta\,hQ| \le |P\,\Delta\,Q|$, *and*
(ii) $|hP\,\Delta\,hQ| \ge (1-\alpha)\,|P\,\Delta\,Q|$ *with probability* $\ge 1-\beta$.

PROOF. Let s be the smallest power of 2 that is at least as large as $2n/(\alpha\beta)$. For the purposes of defining the function, it will be convenient to express each element of $\mathbb{Z}_u$ (resp., $\mathbb{Z}_s$) as a bit vector of length $\lceil \log u \rceil$ (resp., $\log s$). Thus, P and Q can be viewed as multisets of cardinality at most n where elements are drawn from $\mathbb{Z}_2^{\lceil \log u \rceil}$.

Given a matrix $M \in \{0,1\}^{\log s \times \lceil \log u \rceil}$ and a vector $b \in \{0,1\}^{\lceil \log u \rceil}$, define $h_{M,b} \colon \mathbb{Z}_2^{\lceil \log u \rceil} \to \mathbb{Z}_2^{\log s}$ to be

$$h_{M,b}(x) \;=\; Mx + b,$$

where arithmetic operations are performed over $\mathbb{Z}_2$. It is well known that if M and b are randomly generated, the resulting randomized function $h = h_{M,b}$ is a universal hash function [4, 9], and for any fixed $x, y \in \mathbb{Z}_u$, $\Pr[h(x) = h(y)] \le 1/s \le \alpha\beta/(2n)$. Given this bound on the collision probability, the rest of the proof follows the same structure used in the proof of Lemma 2.6. □

3. TRANSLATION-INVARIANT MAPPING

The purpose of the section is to present a transformation of a point set to a vector that is invariant under translations of the point set. In the next section we will apply the results of this section to produce the embedding function described in Theorem 1.

Before presenting this transformation, it will be convenient to explain that we may interpret the point sets P and Q in $\mathbb{Z}_s (\le n)$ as bit-vectors in an s-dimensional space, in particular, as elements of $\mathbb{Z}_2^s$. That is, for $1 \le p \le s$ we set the point's component of the bit-vector to 1 if $p \in P$ and 0 otherwise.

For a translation $t \in \mathbb{Z}_s$, we will still use the notation $P + t$ to denote the translation of P by t, which in this context involves a right circular shift of the elements of this bit vector by t positions.

Given a positive integer ρ, define an (s,ρ)-*probe* to be a ρ-element vector $\pi = (i_1, i_2, \ldots, i_\rho)$, where $i_j \in \mathbb{Z}_s$ for $1 \le j \le \rho$. We say that such a probe is *random* if each element i_j is sampled independently at random from $\mathbb{Z}_s$. (Note that duplicates are possible.) Consider $P \in \mathbb{Z}_2^s$, where $P = (p_1, p_2, \ldots, p_s)$. Define $P[\pi]$ to be the integer whose bit representation is $\langle p_{i_1} p_{i_2} \ldots p_{i_\rho} \rangle$. Define the multiset

$$\widehat{\Phi}_\pi P \;=\; \{(P+t)[\pi] \mid t \in \mathbb{Z}_s\}$$

(This is a multiset because different translations may generate the same bit pattern. See Fig 1) The cardinality of $\widehat{\Phi}_\pi P$ (counting multiplicities) is s, and its elements are from $\mathbb{Z}_{2\rho}$. Because the probe is applied uniformly to all translations in $\mathbb{Z}_s$ we have:

Lemma 3.1. *Given any (s,ρ)-probe π, $\widehat{\Phi}_\pi$ is invariant under translation. That is, for all $t \in \mathbb{Z}_s$, $\widehat{\Phi}_\pi(P+t) = \widehat{\Phi}_\pi P$.*

Our first observation regarding this invariant transformation shows that the distance between the transformed objects is a function of the relationship between the probe length and the actual distance between the point sets under translation. Intuitively, as the probe length ρ increases, the likelihood of encountering a mismatch increases (depending on the distance between the point sets). Part (i) asserts that if ρ is sufficiently small, the distance between the resulting vectors will also be small. Part (ii) asserts that if ρ is sufficiently large, the probability of encountering a mismatch is so high that the distance between the resulting vectors will be almost as high as the maximum possible value of $2s$.

Lemma 3.2. *Consider a positive integer s and two sets $P, Q \in \mathbb{Z}_s (\le n)$, and let $\delta^* = \langle P \,\Delta\, Q \rangle$. Given a positive integer ρ, let π be a random (s,ρ)-probe. Then*

(i) $\left\| \widehat{\Phi}_\pi P - \widehat{\Phi}_\pi Q \right\|_1 \le 2\rho\delta^*$, *and*

(ii) if $\rho \ge (3s \ln s)/\delta^*$ *then* $\left\| \widehat{\Phi}_\pi P - \widehat{\Phi}_\pi Q \right\|_1 \ge 2s - 2$ *with probability at least* $\left(1 - \frac{2}{s}\right)$.

PROOF. Since $\delta^* = \langle P \,\Delta\, Q \rangle$, this means that for some translation of P there are δ^* elements of $\mathbb{Z}_s$ in the symmetric difference $P \Delta Q$. Because $\widehat{\Phi}$ is invariant under translations, there is no loss in generality if we assume that this translation is the identity. Since π has ρ elements, there are at most $\rho\delta^*$ choices of translations t such that the result of $(P+t)[\pi]$ accesses one of these mismatched elements. Each of these may produce a probe value that fails to match any of the probe results of $\widehat{\Phi}_\pi Q$. All the others placements will match the corresponding probe of Q. Similarly, there are $\rho\delta^*$ choices of r such that $(Q+r)[\pi]$ fails to match any probe value of $\widehat{\Phi}_\pi P$, but all others will match. Thus we have

$$\left\| \widehat{\Phi}_\pi P - \widehat{\Phi}_\pi Q \right\|_1 \;\le\; 2\rho\delta^*,$$

which establishes (i).

In order to establish (ii), given any translations $t, r \in \mathbb{Z}_s$, let $\delta_{t,r} = |(P+t) \,\Delta\, (Q+r)|$. Clearly, for any $t, r \in \mathbb{Z}_s$, $\delta_{t,r} \ge \delta^*$. For any particular probe placement π, the probe values $(P+t)[\pi]$ and $(Q+r)[\pi]$ match if and only if each of the selected positions match. Since the indices of π are chosen at random from $\mathbb{Z}_s$, each position matches with probability $(1 - \delta_{t,r}/s)$. Since the elements of π are chosen independently, we have

$$\begin{aligned} \Pr[(P+t)[\pi] = (Q+r)[\pi]] \;&=\; \left(1 - \frac{\delta_{t,r}}{s}\right)^\rho \le \left(1 - \frac{\delta^*}{s}\right)^\rho \\ &\le\; e^{-\delta^*\rho/s}. \end{aligned}$$

By the union bound we have

$$\Pr\left[(P+t)[\pi] \in \widehat{\Phi}_\pi Q\right] \;\le\; \sum_{r \in \mathbb{Z}_s} e^{-\delta^*\rho/s} \;=\; s e^{-\delta^*\rho/s}.$$

Let X be a random variable whose value is the number of probe placements t such that $(P+t)[\pi] \in \widehat{\Phi}_\pi Q$. By the linearity of expectation and since $\rho \ge (3s \ln s)/\delta^*$, we have

$$\begin{aligned} \mathrm{E}[X] \;&=\; \sum_{t \in \mathbb{Z}_s} \Pr\left[(P+t)[\pi] \in \widehat{\Phi}_\pi Q\right] \;\le\; s^2 e^{-\delta^*\rho/s} \\ &\le\; s^2 e^{-(3s\ln s)\rho/s\rho} \;=\; s^2 e^{-3\ln s} \;=\; \frac{1}{s}. \end{aligned}$$

By Markov's inequality, it follows that $\Pr[X \ge 1] \le \frac{1}{s}$. By symmetry, the number of probe placements such that $\widehat{\Phi}(Q+r)[\pi] \in \widehat{\Phi}_\pi P$ satisfies the same bound. Since $\widehat{\Phi}_\pi P$ and $\widehat{\Phi}_\pi Q$ each contain s values (including multiplicities), their L_1 distance is $2s$ minus the number of matches. And so with probability at least $\left(1 - \frac{2}{s}\right)$ we have

$$\left\| \widehat{\Phi}_\pi P - \widehat{\Phi}_\pi Q \right\|_1 \;=\; 2s - 2X \;\ge\; 2s - 2,$$

as desired. □

$P \in \mathbb{Z}_u (\leq n)$	$u = 24$	{3, 6, 10, 14, 22}
$h'P \in \mathbb{Z}_s (\leq n)$	$s = 11$	1 0 0 1 0 0 1 0 0 0 1
$\pi \subset \mathbb{Z}_s$	$\rho = 4$	{0, 3, 6, 7}
$\widehat{\Phi}_\pi$	$\lvert\widehat{\Phi}_\pi\rvert = s$	{1110, 0000, 0000, 1101, 0011, 0010, 1000, 0101, 0110, 0000, 1001}
h''	$s' = 2^4$	3 0 1 1 0 1 1 0 1 1 0 0 0 1 1 0

Figure 1: Example of our invariant transformation. (For simplicity we have chosen h'' to be the identity)

Using the above lemma, we present next the main utility result upon which our translation invariant transformation is based. This lemma shows that given two points sets over $\mathbb{Z}_u$ each containing at most n elements and an estimate δ on their distance, there exists a function Φ that maps a point set to a vector. The function has the property that, after applying this function, the distance between the two resulting vectors reveals whether δ is significantly greater or significantly smaller than the optimal distance δ^*. Intuitively, this is done by randomly probing the point sets such that, under the assumption that $\delta = \delta^*$, the probability of encountering a mismatch will be a constant. Thus, if δ is either significantly larger or significantly smaller than δ^*, the distance between the resulting vectors will reflect this.

As mentioned in the introduction, the process involves three stages, first the application of a randomized linear hash function to reduce the domain size to $O(n \log n)$, then the application of a random probe sequence to produce translation invariance, and finally the application of a second hash function to again reduce the domain size. The entire process is repeated and the results are averaged in order to obtain the desired distortion bounds with sufficiently high probability of success.

Lemma 3.3. *Consider positive integers n and u, where $u \leq n^c$ for some constant $c \geq 1$, a distance estimate $1 \leq \delta \leq 2n$ and failure probability $0 < \beta \leq 1$. There exists $s = \Theta(n \log n)$ and a randomized function $\Phi \colon \mathbb{Z}_u (\leq n) \to \ell_1^m$, where $m = O(n \log n \log(1/\beta))$ that satisfies the following property. Given any two sets $P, Q \in \mathbb{Z}_u (\leq n)$, and letting δ^* denote $\langle P \,\Delta\, Q \rangle$ we have*

(i) $\|\Phi P - \Phi Q\|_1 \leq s\delta^*/\delta$, and

(ii) if $\delta \leq \frac{\delta^*}{8 \ln s}$ then $\|\Phi P - \Phi Q\|_1 \geq s$, with probability at least $1 - \beta$.

For any P, ΦP can be computed in $O(n \log^3 n \log(1/\beta))$ time.

The remainder of this section is devoted to proving this lemma. We first establish (ii). Let us assume that $\delta \leq \langle P \,\Delta\, Q \rangle / (8 \ln s)$, and let $\alpha_0 = \beta_0 = 1/16$. We begin by applying Lemma 2.6 with $\alpha = \alpha_0$ and $\beta = \beta_0$. Let $h' \colon \mathbb{Z}_u \to \mathbb{Z}_s$ denote the resulting function, where $s = \Theta(n \log n)$. We may assume that $s \geq 40$. If we let $P' = h'P$ and $Q' = h'Q$, by Corollary 2.1(ii) we have

$$\left\langle P' \,\Delta\, Q' \right\rangle \;\geq\; (1 - \alpha_0) \left\langle P \,\Delta\, Q \right\rangle,$$

with probability at least $1 - \beta_0$.

Next, we wish to apply Lemma 3.2 to the sets P' and Q' with $\rho = \lfloor s/(2\delta) \rfloor$. Observe that since $s = \Theta(n \log n)$ and $\delta \leq 2n$, we have $\rho = \Omega(\log n)$. We may assume, therefore that $\rho \geq 6$. Letting $\widehat{\Phi}_\pi$ denote the resulting translation invariant transformation, and let $\widehat{P} = \widehat{\Phi}_\pi P$ and $\widehat{Q} = \widehat{\Phi}_\pi Q$. Each is a multiset of cardinality s over $\mathbb{Z}_{2^\rho}$, which can be interpreted as a vector in $\mathbb{R}^{2^\rho}$ in which the ith component is the number of occurrences element i. Observe that the (multiset) symmetric difference distance is equivalent to the L_1 distance in this vector interpretation. Applying these two different interpretations of $\widehat{P}$ and $\widehat{Q}$, we have

$$\left| \widehat{P} \,\Delta\, \widehat{Q} \right| = \left\| \widehat{P} - \widehat{Q} \right\|_1.$$

Since $\rho \geq 6$, with probability at least $1 - \beta_0$ we have:

$$\begin{aligned} \left\langle P' \,\Delta\, Q' \right\rangle \;&\geq\; (1 - \alpha_0) \left\langle P \,\Delta\, Q \right\rangle \;>\; \frac{7}{8} \left\langle P \,\Delta\, Q \right\rangle \;\geq\; 7\delta \ln s \\ &\geq\; \frac{7s \ln s}{2(\rho + 1)} \;\geq\; \frac{3s \ln s}{\rho} \end{aligned}$$

Therefore, we may apply Lemma 3.2(ii) to obtain that with probability at least $(1 - \beta_0)\left(1 - \frac{2}{s}\right)$:

$$\left\| \widehat{P} - \widehat{Q} \right\|_1 \;\geq\; 2s - 2.$$

Given our assumption that $s \geq 40$, the probability of this holding is at least $(1 - \beta_0)^2$.

Finally, we apply Lemma 2.7 to the multisets $\widehat{P}$ and $\widehat{Q}$, where $n = s$, $u = 2^\rho$, $\alpha = \alpha_0$ and $\beta = \beta_0$. Let the resulting function be $h'' \colon \mathbb{Z}_{2^\rho} \to \mathbb{Z}_{O(s)}$, and let $P'' = h''\widehat{P}$ and $Q'' = h''\widehat{Q}$. (Given that ρ may be as large as $\Theta(n \log n)$, we will not compute this function explicitly. Computational issues will be discussed later.) We have

$$\left\| P'' - Q'' \right\|_1 \;\geq\; (1 - \alpha_0) \left\| \widehat{P} - \widehat{Q} \right\|_1 \;\geq\; (1 - \alpha_0)(2s - 2),$$

with probability at least $(1 - \beta_0)^3$.

To summarize, given point sets P and Q, by composing the three randomized functions we have shown the existence of a randomized function $\Phi'' \colon \mathbb{Z}_u (\leq n) \to \ell_1^{O(s)}$, such that with probability at least $(1 - \beta_0)^3$

$$\left\| \Phi'' P - \Phi'' Q \right\|_1 \;\geq\; (1 - \alpha_0)(2s - 2),$$

To increase the probability of success to $1 - \beta$, we repeat the above procedure for $k = \lceil 8 \ln(1/\beta) \rceil$ trials. (Each trial is performed with different random choices, but the overall sequence of random choices is the same for all point sets.) We then concatenate the resulting vectors. There is a subtlety to be noted. Thus far, we have assumed that there is a fixed value of s. Each invocation of Lemma 2.6 generates different random prime s (although all are $\Theta(n \log n)$). To correct for the bias to the distance resulting from larger or smaller values of s, we take s to be maximum possible value produced by the lemma, and for any smaller value s' produced by invoking lemma, we weight the associated vector by s/s'. For simplicity, we may assume that all invocations of this

lemma produce a results of uniform length s. We define $\Phi\colon \mathbb{Z}_u(\leq n) \to \ell_1^m$, where $m = O(sk) = O(n \log n \log(1/\beta))$, to be

$$\Phi P \;=\; \frac{1}{k}\left\langle \Phi_1'', \Phi_2'', \ldots, \Phi_k'' \right\rangle.$$

Because our upper bound on $\|\Phi'' P - \Phi'' Q\|_1$ holds unconditionally (irrespective of the randomization), assertion (i) follows immediately.

To establish (ii), consider the random variables $X_i = \|\Phi_i'' P - \Phi_i'' Q\|_1$, for $1 \leq i \leq k$. Clearly these variables i.i.d., and $0 \leq X_i \leq 2s$. With probability at least $(1-\beta_0)^3$ we have $X_i \geq (1-\alpha_0)(2s-2)$. Therefore we have

$$\mathrm{E}[X] = \mathrm{E}\left[\, \left\|\Phi'' P - \Phi'' Q\right\|_1 \,\right] \geq (1-\beta_0)^3(1-\alpha_0)(2s-2) \geq \frac{3s}{2},$$

where the last inequality holds by our definitions of α_0 and β_0 and of our assumption that $s \geq 40$. Let $X = \frac{1}{k}\sum_{i=1}^{k} X_i$. By Hoeffding's inequality [10], for any $\varepsilon > 0$ we have

$$\Pr\left[\mathrm{E}[X] - X > \varepsilon\right] \;\leq\; \exp\left(-\frac{2k\varepsilon^2}{(2s)^2}\right).$$

By setting $\varepsilon = s/2$, and by our choice of k we have

$$\begin{aligned}
\Pr\left[\, \|\Phi P - \Phi Q\|_1 \leq s\right] &= \Pr[X \leq s] \\
&\leq \Pr\left[\, \mathrm{E}[X] - X \geq \frac{s}{2}\,\right] \\
&\leq \exp\left(-\frac{2(8\ln(1/\beta))(s/2)^2}{(2s)^2}\right) = \beta,
\end{aligned}$$

which establishes (ii).

Next, we establish (i). By part (i) of both Lemmas 2.6 and 2.7, the functions that they produce cannot increase distances. By Lemma 3.2 therefore, we have

$$\|\Phi P - \Phi Q\|_1 \;\leq\; 2\rho\delta^* \;\leq\; 2\frac{s}{2\delta}\delta^* \;\leq\; \frac{s\delta^*}{\delta},$$

as desired.

Finally, we present the time complexity to compute the function. The first hash function $h'\colon \mathbb{Z}_u \to \mathbb{Z}_{O(n\log n)}$ takes time $O(n)$ for a set $P \in \mathbb{Z}_u(\leq n)$. A naive implementation of the construction of $\widehat{\Phi}P$ would require excessive time and space. (To see this, observe that each probe involves $O(s)$ elements and must be applied to $O(s)$ distinct positions, which would yield a total time bound of $O(n^2 \log^2 n)$ to compute the results of even a single probe.)

To achieve greater efficiency, we perform the probing and the second hashing functions as a single operation. Recall that P' denotes the point set after h', and recall that $s = \Theta(n \log n)$ is the size of its domain. Let π denote a probe of some size ρ. The invocation of Lemma 2.7 in this context maps points from $\mathbb{Z}_{2^\rho}$ to $\mathbb{Z}_{s'}$ where $s' = \Theta(s)$ and is a power of 2. This involves a random matrix $M \in \{0,1\}^{\log s' \times \rho}$ and random column vector $b \in \{0,1\}^\rho$.

Constructing $\widehat{\Phi}P'$ for each $t \in \mathbb{Z}_s$ involves computing $x_t = (P'+t)[\pi]$, and then applying the function $h''(x_t) = Mx_t + b$ (with operations performed over $\mathbb{Z}_2$). To do this, we decompose this operation into $\log s'$ operations. Let m_i denote the i-th row of M, and let $r_{i,t}$ be the value of the boolean inner product $(m_i \cdot x_t)$. Our objective is to compute $r_{i,t}$ for $i \in \{1, 2, \cdots, \log s'\}$. Since m_i is a boolean bit-vector we have

$$r_{i,t} \;=\; \left(m_i \cdot (P'+t)[\pi]\right) \;=\; \sum_{j=1}^{\rho} \left(m_i[j] \cdot \left((P'+t)[\pi]\right)[j]\right).$$

Observe that this a part of a convolution. Thus, given a row m_i, we can compute $r_{i,t}$ for all t as $P' \otimes [\pi \wedge m_i]$, where $\otimes$ denotes boolean convolution and $\wedge$ denotes the bitwise *and* operation.

Therefore, for all $i \in \mathbb{Z}_{\log s'}$ and $t \in \mathbb{Z}_s$ $r_{i,t}$ can computed through $\log s'$ convolution operations. It is easy to see that $h''(x_t) = r_{*,t} + b$ (interpreted now as a binary number). Since each convolution takes time $O((s+\rho)\log(s+\rho)) = O(s \log s)$, the total running time is $O((n + s\log^2 s)\log(1/\beta)) = O(n \log^3 n \log(1/\beta))$. This completes the proof.

4. EMBEDDING

In this section we present a proof of Theorem 1. First, let $\beta_0 = \beta/(2\log n)$. By Lemma 2.2 we may assume that the point sets have already been mapped from $\mathbb{Z}_u^d(\leq n)$ to the 1-dimensional space $\mathbb{Z}_{u'}(\leq n)$, where $u' = O(u^d)$. We apply Lemma 3.3 repeatedly with failure probability β_0 and distance estimates δ from $\{1, 2, 4, \ldots, 2^k\}$, where $k = \lceil \log(2n) \rceil$. (Note that some of this notation overlaps that used in the proof of Lemma 3.3, but the meanings here are quite different.) Let $\delta_i = 2^i$, and let $\Phi_i P$ denote the result of applying Lemma 3.3 with $\delta = \delta_i$. We apply a scalar weight to each of the resulting vectors and concatenate them to produce:

$$\Psi P = \left\langle \frac{1}{s}\Phi_0 P,\; \frac{2}{s}\Phi_1 P,\; \frac{4}{s}\Phi_2 P,\; \ldots,\; \frac{2^i}{s}\Phi_i P,\; \ldots,\; \frac{2^k}{s}\Phi_k P \right\rangle.$$

Observe that $\Psi\colon \mathbb{Z}_{u'}(\leq n) \to \ell_1^m$, where the dimension of the range of m is $O(kn \log n \log(1/\beta)) = O(n \log^2 n \log(1/\beta))$.

We first establish part (i) of Theorem 1. Observe that

$$\begin{aligned}
\|\Psi P - \Psi Q\|_1 &= \sum_{i=0}^{k} \left\| \frac{2^i}{s}\Phi_i P - \frac{2^i}{s}\Phi_i Q \right\|_1 \\
&= \sum_{i=0}^{k} \frac{2^i}{s} \|\Phi_i P - \Phi_i Q\|_1 . \qquad (2)
\end{aligned}$$

Let $\delta^* = \langle P \,\Delta\, Q \rangle$. By Lemma 3.3(i) we have

$$\|\Phi_i P - \Phi_i Q\|_1 \leq s\delta^*/\delta_i = s\delta^*/2^i.$$

Also, $\Phi_i P$ and $\Phi_i Q$ each have at most s elements, and therefore $\|\Phi_i P - \Phi_i Q\|_1 \leq 2s$. Thus we have

$$\begin{aligned}
\|\Psi P - \Psi Q\|_1 &\leq \sum_{i=0}^{k} \frac{2^i}{s} \min\left(2s, \frac{s\delta^*}{2^i}\right) \\
&\leq \sum_{i=0}^{k} 2^i \min\left(2, \frac{\delta^*}{2^i}\right).
\end{aligned}$$

Observe that $\delta^*/2^i \leq 2$ for $i \geq \log\delta^* - 1$. Letting $k' = \lfloor \log\delta^* \rfloor$, we have

$$\begin{aligned}
\|\Psi P - \Psi Q\|_1 &\leq \sum_{i=0}^{k'-1} 2^{i+1} + \sum_{i=k'}^{k} \delta^* \;\leq\; 2^{k'+1} + \sum_{i=0}^{k} \delta^* \\
&\leq 2\delta^* + \delta^*(2 + \log 2n) \;\leq\; 2\delta^* \log n
\end{aligned}$$

for sufficiently large n. This establishes part (i).

To establish part (ii), Let $k'' = \lfloor \log(\delta^*/(8\ln s)) \rfloor$. Observe that for all $i \leq k''$, we have $\delta_i \leq \delta^*/(8\ln s)$. Therefore, we may apply Lemma 3.3(ii). Starting with Eq. (2)

with probability $(1-\beta)$ we have

$$\begin{aligned}\|\Psi P - \Psi Q\|_1 &= \sum_{i=0}^{k} \frac{2^i}{s} \|\Phi_i P - \Phi_i Q\|_1 \\ &\geq \sum_{i=0}^{k''} \frac{2^i}{s} s \;=\; 2^{k''+1} - 1 \\ &\geq \frac{\delta^*}{8 \ln s} - 1 \;\geq\; \frac{\delta^*}{16 \ln s}\end{aligned}$$

The last inequality is satisfied for $\frac{\delta^*}{16 \ln s} \geq 1$. Let us consider the case of $\frac{\delta^*}{16 \ln s} \leq 1$. We observe that

$$\|\Psi P - \Psi Q\|_1 \;\geq\; 2^{k''+1} - 1 \;\geq\; 1.$$

Because δ^* is less than $16 \ln s$ and $\|\Psi P - \Psi Q\|_1 \geq 1$, the distortion is at most $16 \ln s$ ($\leq 17 \log n$) with sufficiently large n. This establishes part (ii).

Finally, to establish the execution time, we observe that we invoke Lemma 3.3 $k = O(\log n)$ times. Each invocation takes $O(n \log^3 n \log(1/\beta))$ time. Thus, the total execution time is $O(n \log^4 n \log(1/\beta))$. This completes the proof.

5. CONCLUSIONS

We have presented a randomized algorithm that embeds an n-element point set over the multidimensional grid $\mathbb{Z}_u^d$, where u is $n^{O(1)}$, to a single point in a multidimensional space under the L_1 distance. We assume that distances over $\mathbb{Z}_u^d$ are measured using the symmetric difference under translation. This embedding has the property that with some given probability, it achieves a distortion of $O(\log^2 n)$. Our existing work applies to points with integer coordinates in arbitrary dimensions and is robust to missing and spurious points. The conditions under which our embedding applies are admittedly restrictive, but to our knowledge this is the first result in embeddings that are invariant under geometric transformations and robust to outliers.

In addition to the obvious questions of improving the distortion bounds and eliminating randomization, there are two significant issues. First, the symmetric difference distance function that we use requires that points match identically. Second, our embedding is invariant only under the group of translations. In practice, point coordinates are the result of measurements, and will be subject to the presence of noise and digitization errors. Handling noise is one important extension, which we would like to consider. Similarly, it would be very useful to consider allowing more extensive groups of geometric transformations, including, for example, rotation and/or scaling.

6. REFERENCES

[1] H. Alt, O. Aichholzer, and G. Rote. Matching shapes with a reference point. In *Proc. 10th Annu. ACM Sympos. Comput. Geom.*, pages 85–92, 1994.

[2] H. Alt and L. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In *Handbook of Computational Geometry*, pages 121–153. 1999.

[3] M. Boutin and G. Kemper. Which point configurations are determined by the distribution of their pairwise distances? *Int. J. Comput. Geometry Appl.*, 17(1):31–44, 2007.

[4] L. Carter and M. N. Wegman. Universal classes of hash functions. *J. Comput. Syst. Sci.*, 18(2):143–154, 1979.

[5] F. H. Cheng. Point pattern matching algorithm invariant to geometrical transformation and distortion. *Pattern Recogn. Lett.*, 17(14):1429–1435, 1996.

[6] S. D. Cohen and L. J. Guibas. The earth mover's distance under transformation sets. In *Proc. 4th Annu. IEEE Int'l Conf. on Computer Vision*, pages 1076–1083, 1999.

[7] G. Cormode and S. Muthukrishnan. The string edit distance matching problem with moves. *ACM Trans. Algorithms*, 3(1):2, 2007.

[8] T. Dakic. *On the turnpike problem*. PhD thesis, Simon Fraser University, 2000.

[9] O. Goldreich and A. Wigderson. Tiny families of functions with random properties: a quality-size trade-off for hashing. *Random Struct. Algorithms*, 11(4):315–343, 1997.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[11] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:850–863, 1993.

[12] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.*, page 10, 2001.

[13] P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *3rd Int'l Workshop on Stat. and Comput. Theories of Vision*, 2003.

[14] R. M. Karp and M. O. Rabin. Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.*, 31(2):249–260, 1987.

[15] H. Ling and D. W. Jacobs. Deformation invariant image matching. In *Proc. 10th Annu. IEEE Int'l Conf. on Computer Vision*, pages 1466–1473, 2005.

[16] J. Rosenblatt and P. Seymour. The structure of homometric sets. In *SIAM J. Alg. Disc. Methods*, volume 3,3, pages 343–350, 1982.

[17] B. Rosser. Explicit bounds for some functions of prime numbers. *Amer. J. Mathematics*, 63(1):211–232, 1941.

[18] S. S. Skiena, W. D. Smith, and P. Lemke. Reconstructing sets from interpoint distances (extended abstract). In *Proc. 6th Annu. ACM Sympos. Comput. Geom.*, pages 332–339, 1990.

[19] Z. Wang, W. Dong, W. Josephson, Q. Lv, M. Charikar, and K. Li. Sizing sketches: a rank-based analysis for similarity search. In *SIGMETRICS Perform. Eval. Rev.*, volume 35, pages 157–168, 2007.

[20] H. J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comput. Sci. and Eng.*, 4:10–21, 1997.