Distribution-based Similarity Measures for Multi-dimensional Point Set Retrieval Applications

Jie Shao†, Zi Huang†, Heng Tao Shen†, Jialie Shen‡, and Xiaofang Zhou†
†School of Information Technology & Electrical Engineering, The University of Queensland, Australia
‡School of Information Systems, Singapore Management University, Singapore
{jshao, huang, shenht, zxf}@itee.uq.edu.au jlshen@smu.edu.sg

ABSTRACT

Effective and efficient method of similarity assessment continues to be one of the most fundamental problems in multimedia data analysis. In case of retrieving relevant items from a collection of objects based on series of multivariate observations (e.g., searching the similar video clips in a repository to a query example), satisfactory performance cannot be expected using many conventional similarity measures based on the aggregation of element pairwise comparisons. Some correlation information among the individual elements has also been investigated to characterize each set of multi-dimensional points for ranked retrieval, by making use of an unwarranted assumption that the underlying data distribution has a particular parametric form. Motivated by this observation, this paper introduces a novel collective gauge of relevance ranking by evaluating the probabilities that point sets are consistent with the same distribution of the query. Two non-parametric hypothesis tests in statistics are justified to exploit the distributional discrepancy of samples for assessing the similarity between two ensembles of points. While our methodology is mainly presented in the context of video similarity search, it enjoys great flexibility and can be easily adapted to other applications involving generic multi-dimensional point set representation for each object such as human gesture recognition.

Categories and Subject Descriptors: H.3.3 [Information Store and Retrieval]: Information Search and Retrieval; G.3 [Probability and Statistics]: Multivariate statistics

General Terms: Algorithms, Measurement, Experimentation, Performance

Keywords: Similarity measures, multi-dimensional point set, non-parametric, hypothesis tests, minimal spanning tree, reproducing kernel Hilbert space

1. INTRODUCTION

A multi-dimensional sequence dataset can be treated a collection of series of multivariate observations. This kind of data is becoming popular in various multimedia applications as well as other scientific domains such as environmental, biomedical and biometric information systems. For example, a cyber-glove used in human-computer interaction usually has around 20 sensors on it, each of which can generate about 50 values per second [15, 20]. Another typical example is that a video sequence can be viewed as $X = \{x_1, x_2, \ldots, x_n\}$, where each element $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector point representing a frame in X, and n is the number of sampling frames. In general, each item in a collection of multi-dimensional point sets is normally stored as an $n \times d$ matrix, where n is the number of observations, and d is the number of variables, e.g., the dimensionality of feature vector space.

How to measure the similarity of multi-dimensional point sets is a long-standing and challenging research problem in large scale multimedia data analysis. During the past decade, a number of similarly measures have been proposed to compare multi-dimensional sequences. Mean distance [19] is the most generic among them and it firstly extends the traditional Euclidean distance for time series to support multi-dimensional sequence matching. Other classical examples, such as Dynamic Time Warping (DTW) [16], Longest Common Subsequence (LCSS) [27] and Edit Distance (ED), e.g., Edit Distance on Real sequence (EDR) [7] can also be adapted for measuring the similarity of multi-dimensional sequences. In essence, all of them are based on element-to-element comparison results by aggregating separate differences. However, these **element-based** approaches have some limitations and could be unreliable. This is due to the fact that, a slight dithering in the original space (e.g., photometric or geometric video transformations with different encoding parameters such as brightness, saturation, aspect ratio or cropping, etc.) may lead to a considerable change in the observed feature space. Moreover, in the presence of noise, these methods still accumulate all the differences of individual elements thus trigger performance degradation. As a consequence, element pairwise comparisons are virtually not tolerant to shift, scale or rotation, and often overly sensitive to noise.

In many circumstances, it is more reasonable to treat each multi-dimensional point set as a whole entity, since there are naturally some important correlations among the individual elements insensitive to these distortions which can be considered for retrieval. Recently, some **correlation-based** approaches, such as applying Principal Component Analysis (PCA) to the original matrices for deriving some correlation information (e.g., eigenvectors and associated eigenvalues)

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

as the global representations of multi-dimensional point sets, are demonstrated to improve the efficiency of retrieval applications [29, 24]. Nevertheless, accompanied with PCA, these methods implicitly bring an artificial restriction that the data points in each set are drawn from a particular parametric form - Gaussian normal distribution. In other words, the orientations of orthogonal principal components coincide with the independent axes of data scattering of any set, which is approximately a d-dimensional ellipsoid. Unfortunately, real data are rarely that well-behaved in practice.

From the maximum likelihood point of view, the problem of finding a good distance function comes down to the maximization of similarity probability. Distribution-based similarity measures have exhibited excellent performance in many applications, such as image retrieval with color and texture features [23, 21, 26]. In text retrieval, information-theoretically motivated distances such as Kullback-Leibler (KL) divergence of word frequency distributions are widely used in the language modelling framework. In bioinformatics, microarray data from different tissue types can be compared by their distributions as well, either to determine whether two sub-types of cancer can be treated as statistically indistinguishable from a diagnosis perspective, or to detect differences in normal and tumor tissues [4]. Interestingly, the data points of visually similar video clips also appear to inherit some intrinsic characteristics, such as constant inter-frame topological relationship. Therefore, vectorial distribution is often expected to be robust to a number of distortions in a properly defined feature space (we present some visualization examples in Section 5). This inspired us to exploit some invariant distribution information to determine their relevance.

This paper introduces a novel collective perspective of measuring the similarity of data point sets by **distribution-based** approaches. Several ideas of hypothesis tests in the literature of statistics are utilized to compare the multivariate distributions of two ensembles of points for assessing their similarity. The two specific methods adopted are *Friedman-Rafsky (FR) test* [10], which is a multivariate extension of Wald-Wolfowitz (WW) test [28], and a more efficient *Maximum Mean Discrepancy (MMD) test* [11], which employs the unit balls in a universal Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ [25] as its function class. The classical WW test is based on the intuition that if two samples are similar, the order of their univariate observations will be randomly distributed when combining together in an ordered list. FR test [10] is its multivariate generalization by examining how two samples yield different disjoint subtrees in an overall Minimal Spanning Tree (MST). MMD test [11] takes a distinctive design philosophy. It is a kernel method of computing the maximum difference of the mean function values of two samples mapped into an RKHS as their distributional discrepancy. Kernelization renders some linear statistics in RKHS able to capture the potential nonlinear distributions in input space [14]. The test statistic of either method can be expressed as the probability that two ensembles of points are consistent with a same multivariate distribution, which itself is not explicitly known. We further extend this notion as a collective gauge of relevance ranking especially tailored for similarity search in a collection of multi-dimensional point sets. The main advantage of our proposal is that, no prior knowledge about the underlying data distributions of the point sets under study has to be assumed, in contrast to the aforementioned correlation-based approaches. It provides a more comprehensive analysis that captures the essence of invariant distribution information for relevance ranking in the process of retrieval applications.

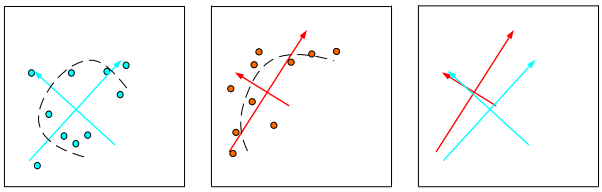
In order to evaluate the performance of distribution-based similarity search on multimedia information retrieval in a collection of multi-dimensional point set representations, we first perform some experiments to study one of its applications - video similarity search. In particular, the visual content of each video clip in our test repository can be characterized by its vectorial distribution in a properly defined feature space. Each comparison of video clips involves the computation of some test statistic which provides the basis for their distributional discrepancy. We evaluate which method assesses the similarity of relevant items more properly. The results of this dataset show the superiority of FR method as compared to the recently introduced correlation-based analysis, which in turn outperforms the aggregation of element pairwise comparisons such as DTW and ED. In addition, our methodology can be easily extended to support other applications wherever generic multi-dimensional sequence representations are involved such as motion capture retrieval. This can be validated with the experiments conducted on the high-quality recordings of Australian sign language (Auslan) data [15] obtained from UCI KDD archive[1]. The results of this dataset demonstrate that the superb retrieval quality with the efficiency close to the best performing alternative can be achieved by MMD method.

The rest of this paper is organized as follows: after briefly reviewing some closely related research efforts in Section 2, Section 3 justifies and tailors two hypothesis tests to better exploit the statistical distributions of data as similarity measures. The framework of our evaluation is described in Section 4, and the results are reported in Section 5. Finally, we conclude and mention some scheduled extensions of this work as future research objectives in Section 6.

2. RELATED WORK

Yang and Shahabi [29] introduced a similarity measure named Extended Frobenius norm (Eros) for multivariate time series based on PCA similarity factor [18]. They first apply PCA to pre-process each matrix of multi-dimensional point set regardless of its real form of distribution, and then summarize the distribution of individual elements to a global representation described by the top k ranked eigenvectors and associated eigenvalues. Specifically, the sum of squared cosine values of the angles between two eigenvectors, which is weighted based on the corresponding eigenvalues, is obtained as the distance between two sets of multivariate observations. More recently, Shen *et al.* [24] proposed a Bounded Coordinate System (BCS) method to globally summarize each multi-dimensional point set by the mean value of elements and a few top ranked principal components bounded by two furthermost projections. To be precise, each $n \times d$ matrix is characterized by a centering d-dimensional point and some d-dimensional vectors showing the major orientations and ranges of data scattering. The proposed similarity measure is subsequently transformed into a comparison of the corresponding compact representations, which intuitively can be estimated by performing translation, rotation and scaling operations to match two BCSs, as illustrated

[1]http://kdd.ics.uci.edu



(a) Plot one set of data points and apply PCA.

(b) Plot another set of data points and apply PCA.

(c) Compare based on their principal components.

Figure 1: A 2-dimensional illustration of BCS.

with a simple example in Figure 1.

The main advantage of these correlation-based approaches is the linear time complexity of online similarity measure (taking the advantage of generating the compact representations of individual sets offline). However, they inherit some weaknesses. First, PCA does not take local phenomena into account, which may be very relevant in determining the similarity between two objects represented by multi-dimensional point sets that could differ in means, variances or shapes. Although this kind of compact representations can facilitate rapid identification, it is generally not tolerant to the perturbation of various distortions. For example in the specific application of video similarity search, while being efficient for detecting the exact duplicate copies, it tends to overlook relevant items which are perturbed by factors such as photometric or geometric transformations, as suggested in [30]. Second, they implicitly assume the data points in each set are always Gaussian normal distributed in $\mathbb{R}^d$ space, which is statistically invalid in real world, especially for data residing in high-dimensions. It is generally acknowledged that at least in many applications, multiple clusters with mixture of Gaussian distributions are often present. As shown in Figure 1(a) and Figure 1(b), PCA can be ill-adapted to the sets of data points with intrinsic nonlinear correlations [6], since PCA is a linear model designed for the scattering of data which is approximately a (hyper-)ellipsoid. The real distribution of data points denoted by the dashed line in each figure actually calls for nonlinear modelling. There are some attempts to use multiple PCA representations [5] to cope with this nonlinearity. However, they will also involve more complex representations and comparisons.

In fact, the aforementioned Eros and BCS methods resorting to correlation information can be regarded as some attempts to roughly reflect the statistical distributions of data by some coarse content tendency in $\mathbb{R}^d$ space. Our proposal based on the criterion of distributional discrepancy is more direct, reliable (more descriptive local information will be exploited) and general (not can only fit a particular data distribution), which shows better retrieval quality in the experiments of two different applications. For easy reference, a list of notations used in this paper is shown in Table 1.

Notation	Description
X, Y	sets of points
x, y	data points
p, q	probability distributions
m, n	sample point numbers
$\mathbb{R}^d$	d-dimensional metric space
R	number of runs
μ, σ	mean, standard deviation
C	permutation parameter of MST
W	quantity of run test statistic
$\mathcal{H}$	reproducing kernel Hilbert space
f, $\mathcal{F}$	function, class of functions
Φ, $k(\cdot,\cdot)$	feature map, kernel function
$\mathbb{E}[f]$	expectation of f
MMD	quantity of maximum mean discrepancy

Table 1: A list of notations.

3. DISTRIBUTION-BASED COMPARISONS

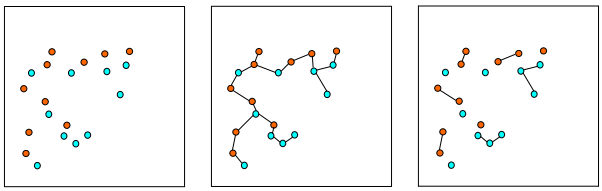
Suppose we have observations of $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, $\{x_i\}$ are random variables originating from probability distribution p, and independently, $\{y_i\}$ are random variables originating from probability distribution q, both defined in $\mathbb{R}^d$ space. There are a variety of non-parametric tests available in statistics to check the null hypothesis $H_0 : p = q$, against a general alternative hypothesis $H_1 : p \neq q$, i.e, to check whether they are consistent with a same distribution. In contrast to some existing approaches, their likeness can be directly estimated based on data points from unknown distributions. Since the hypothesis tests for such a *two-sample problem* are based on finite samples, it is possible that an incorrect answer will be returned. The requirement of such a hypothesis test is that it should have two-sided guarantees: in the large sample limit, it reports spurious differences with low probability (i.e., it should only have few false positives), and it detects true differences with 100% probability (i.e., it should never have false negatives). In statistical jargon, a test with the above ideal property is *consistent*.

The first test we considered for multi-dimensional point set retrieval applications exploits the geometric mixture of X and Y directly in $\mathbb{R}^d$ space, whereas the second test considered here takes a fundamentally different design philosophy of mean function value computations in another higher-dimensional $\mathcal{H}$ space along with the feature map Φ induced by some kernel $k(\cdot,\cdot)$. In practice, X can be interpreted as a baseline set representing a query, with which a search target set Y in a repository is compared. We make these two ideas of non-parametric hypothesis tests adapt to similarity search in a collection of multi-dimensional point set representations.

3.1 FR Test in $\mathbb{R}^d$ Space

For samples in $\mathbb{R}^d$, $d = 1$, Wald and Wolfowitz [28] proposed to begin with combining the $N = m + n$ *univariate* observations together and sort them in ascending order irrespective of sample identity. Each element is then replaced by a label of identity 'X' or 'Y' depending on the set to which it originally belonged, which yields a ranked label sequence. A *run* is defined as a consecutive sequence of identical labels. For example, a sequence in the form of '$XXXXYYYXYYXXXXXYYYYY$' has 6 runs, three of which consist of 'X's and the others consist of 'Y's. The test statistic of such a two-sample problem is the total number of runs R. The rejection of H_0 is for small values of R. The null distribution (i.e., the probability distribution of R when the null hypothesis H_0 is true) can be derived by a combinatorial analysis. If two labels alternate randomly, R is a random variable whose distribution has the mean

$$\mu = \frac{2mn}{N} + 1$$



(a) Pool the samples of two sets. (b) Construct the overall MST for the graph. (c) Remove the edges linking different samples.

Figure 2: A 2-dimensional illustration of FR test.

and the variance

$$\sigma^2 = \frac{2mn(2mn-N)}{N^2(N-1)}.$$

It has been proven a quantity

$$W = \frac{R-\mu}{\sigma} = \frac{R - \frac{2mn}{N} - 1}{\sqrt{\frac{2mn(2mn-N)}{N^2}}}$$

asymptotically approaches a standard normal distribution [28]. This test is consistent if $m, n \to \infty$ with the ratio m/n bounded away from 0 and ∞.

The challenge of generalizing the classical WW test to more than one-dimensional values is that there is usually no immediate total order for multivariate observations. Friedman and Rafsky [10] intelligently conjectured a multivariate extension of it. The proposed FR test employs the Minimum Spanning Tree (MST) of *multivariate* observations as a generalization of univariate sorted list. Suppose in total there are N pooled sample points $\{x_i\} \cup \{y_i\}$ in $\mathbb{R}^d$ space ($d \geq 2$), naturally they can be viewed as the vertices of a connected graph, while the inter-point differences (e.g., their Euclidean distances) are the weights of corresponding edges. First, from the 'complete graph' which has $N(N-1)/2$ edges, a MST can be constructed as a sorting scheme which connects all the elements in such a way that the total Euclidean distance is minimum. Then, all the edges from different samples are removed and only those edges connecting different labelled vertices are retained. Finally, the test statistic R is defined as the number of resulting disjoint subtrees. For the sake of clarity following the example in Figure 1, Figure 2(a) illustrates two samples of size $m = n = 10$ in the 2-dimensional space. The overall MST for the graph of pooled sample points can be constructed as shown in Figure 2(b). By removing the edges that link vertices from different samples, the number of disjoint subtrees can be obtained as the number of runs R, which is conceptually analogous to its counterpart in the univariate case. The test statistic in Figure 2(c) is $R = 12$. Note that, MST connects all the vertices with $N-1$ edges and vertex pairs defining the edges represent points that tend to be close to each other.

Number the $N-1$ edges of MST arbitrarily and define Z_i, $1 \leq i \leq N-1$ as:

$$Z_i = \begin{cases} 1 \text{ if the } i\text{th edge links vertices from different samples} \\ 0 \text{ otherwise,} \end{cases}$$

then the number of runs can be written as

$$R = \sum\nolimits_{i=1}^{N-1} Z_i + 1.$$

Let C be a permutation parameter defined as the number of edge pairs that share a common vertex, and d_i be the degree of ith vertex, then $C = \left(\sum_{i=1}^{N} d_i(d_i-1)\right)/2$. In Figure 2(b), the number of edges that share a common vertex is $C = 12+3\times 3 = 21$, as there are 12 pairs of edges that share a vertex, and 3 triples of edges that share a vertex. In other words, C depends on the topological configuration of MST. Under the null hypothesis H_0, the mean of R is the same as in the univariate case, and the variance is given by

$$\sigma^2 = \frac{2mn}{N(N-1)} \times \left\{ \frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)} \times [N(N-1) - 4mn + 2] \right\}.$$

Similarly, in conjunction with a permutational central limit theorem, a quantity

$$W = \frac{R-\mu}{\sigma}$$

is referred to a standard normal distribution [10]. Here we remark that the quantity of W can express how R deviates from its conditional expectation μ, where μ and σ are given in closed form based on m and n. In our running example, $\mu = 2mn/N + 1 = 2 \times 10 \times 10/20 + 1 = 11$, and $W = (R-\mu)/\sigma = (12-11)/\sqrt{4.644} = 0.464$. With the adjacent matrix of inter-point distances, an MST can be constructed in $O(N^2)$ time. Finding d_i in the constructed MST requires $O(N)$ operations per vertex, thus determining the value of C also requires $O(N^2)$ complexity, which is needed by computing σ. Obtaining W additionally takes $O(N)$ time.

As analyzed by Henze and Penrose [12], the number of removed edges normalized by the total vertex number $N = m+n$ (i.e., $(R-1)/N$) converges to

$$2ab \int \frac{pq}{ap+bq}$$

almost surely with respect to weights a and $b = 1-a$, when $m, n \to \infty$ for asymptotic with the ratio $m/(m+n) \to a \in (0,1)$. The above convergence of test statistic R can also be expressed as

$$\frac{R-1}{N} \to 1 - \delta(p,q,a) = 1 - \int \frac{a^2p^2 + b^2q^2}{ap+bq},$$

where $\delta(p,q,a)$ is termed Henze-Penrose (HP) divergence [12] between two probability distributions p and q.

The constructed MST of our first method investigated contains an edge linking each vertex and the vertex closest to it (i.e., it conveys the nearest neighbor information about each point), and if any edge of MST is removed, thereby dividing the graph into two disjoint subgraphs, the weight of removed edge is the smallest inter-point distance between two resulting disjoint subsets (i.e., it conveys the shortest linkage information about subset of points). It offers a nice and concise description of the locality variation of a multi-dimensional point set. Generally, the structure of MST is unchanged under some distortions preserving the ordering of edge lengths. Meanwhile, it is relatively insensitive to small amounts of noise widely and randomly spread over the field. For example, if someone has manipulated the original video content by just adding a few noise frames, while in human perception they are still quite similar, the primary part of MST representation for similarity comparison will not be

tampered with, and their vectorial distributions are still expected to be close to each other. Therefore, the structural relationship of MST can provide some robustness to simple transformations and noise degradations.

3.2 MMD Test in $\mathcal{H}$ Space

Another method we investigated in this paper is a sophisticated kernelization-based technique originally introduced by Gretton *et al.* [11], which exploits the difference between the expectations of two samples mapped into a universal Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ as the test statistic of their distributional discrepancy. The premise is that, for a compact metric space $\mathbb{R}^d$ ($\mathbb{R}^d$ is closed and bounded), $p = q$ if and only if $\mathbb{E}_p[f(x)] = \mathbb{E}_q[f(y)]$ for all $f \in C(\mathbb{R}^d)$, where $C(\mathbb{R}^d)$ is the space of continuous bounded functions in $\mathbb{R}^d$. As a result, no density estimations are required as an intermediate step but we can use mean function values in an RKHS $\mathcal{H}$ directly, provided that $\mathcal{H}$ is universal [25]. When the quantity of the maximum difference of mean function values is large, two samples are likely from different distributions.

More formally, let $\mathcal{F}$ denote a class of functions f, then Maximum Mean Discrepancy (MMD) between p and q and its empirical estimate can be defined as

$$MMD(p, q, \mathcal{F}) = \sup_{f \in \mathcal{F}} (\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(y)])$$

and

$$MMD(X, Y, \mathcal{F}) = \sup_{f \in \mathcal{F}} (\frac{\sum_{i=1}^{m} f(x_i)}{m} - \frac{\sum_{i=1}^{n} f(y_i)}{n}),$$

respectively, where the supremum operator in each equation signifies the *least upper bound* of a set of values. At first glance, this form is somewhat similar to the two-sample Kolmogorov-Smirnov (KS) test [9], which uses the maximum deviation between two empirical distributions

$$D_{m,n} = \sup_{x,y} (S_m(x) - S_n(y))$$

as its test statistic, where $S_m(x)$ and $S_n(y)$ are the step functions (some piecewise constant functions) of sets X and Y, respectively. KS test has a desirable property of being invariant to arbitrary strictly monotonic transformations of data. However, it is designed for one-dimensional values.

The quality of MMD as a test statistic critically relies on the class of functions $\mathcal{F}$ it employed. Although letting $\mathcal{F}$ be all the continuous bounded functions in $\mathbb{R}^d$ in principle allows $p = q$ to be identified uniquely, such a rich function class is unpractical to work with in the finite sample setting. $\mathcal{F}$ should be *rich* enough so that the value of MMD vanishes if and only if $p = q$, yet *restrictive* enough to provide useful finite sample estimates. If we employe a class of smooth functions $\mathcal{F} = \{f | \|f\|_{\mathcal{H}} \leq 1\}$ which are unit balls in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathbb{R}^d$ with associated kernel $k(\cdot, \cdot)$, $MMD(X, Y, \mathcal{F}) = 0$ if and only if $p = q$ [11]. For example, Gaussian and Laplace Radius Basis Function (RBF) kernels are universal [25], so both of them are members of function class $\mathcal{F}$. Different kernel functions have been designed based on their closure properties. The performance of MMD apparently may depend on the specific kernel used. In the experiments, we test different kernels in MMD.

A Hilbert space is a complete, normed vector space endowed with a dot product $\langle \cdot, \cdot \rangle$ giving rise to its norm via $\|x\| = \sqrt{(x, x)}$. In RKHS, function evaluations can be written in the dot product form of $f(x) = \langle k(x, \cdot), f \rangle$ [25]. Denote the expectation of $k(x, \cdot)$ by $\mu_p = \mathbb{E}_p[k(x, \cdot)]$, since $\mathbb{E}_p[f(x)] = \mathbb{E}_p[\langle k(x, \cdot), f \rangle] = \langle \mu_p, f \rangle$, MMD can be re-written in the norm form of

$$MMD(p, q, \mathcal{F}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_p - \mu_q, f \rangle = \|\mu_p - \mu_q\|_{\mathcal{H}}.$$

Considering

$$\mu_X = \frac{\sum_{i=1}^{m} k(x_i, \cdot)}{m}$$

and the squared form of MMD

$$\|\mu_p - \mu_q\|_{\mathcal{H}}^2 = \langle \mu_p, \mu_p \rangle_{\mathcal{H}} - 2 \langle \mu_p, \mu_q \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}},$$

the empirical estimate of MMD (again, in the squared form) can be expressed in a more convenient manner to compute:

$$MMD^2(X, Y, \mathcal{F}) = \frac{\sum_{i,j=1}^{m} k(x_i, x_j)}{m^2} - \frac{2 \sum_{i,j=1}^{m,n} k(x_i, y_j)}{mn} + \frac{\sum_{i,j=1}^{n} k(y_i, y_j)}{n^2}.$$

In a nutshell, this test is based on the maximum deviation of the expectation of a function evaluated on each random variable mapped into a universal RKHS $\mathcal{H}$, taken over a sufficiently rich function class $\mathcal{F}$. However, the above test statistic of MMD is biased, although it has an upper bound on the deviation between the expected value of empirical estimate and true value [11]. Based on the asymptotic normality of U-statistic [17], an unbiased empirical estimate of $MMD(p, q, \mathcal{F})$ can be obtained. With some kernel $k(\cdot, \cdot)$, the squared form of MMD can be modified as

$$MMD^2(p, q, \mathcal{F}) = \mathbb{E}_p \left[k(x, x')\right] - 2\mathbb{E}_{p,q} \left[k(x, y)\right] + \mathbb{E}_q \left[k(y, y')\right].$$

Especially let $m = n$, its unbiased empirical estimate

$$MMD^2(X, Y, \mathcal{F}) = \frac{\sum_{i \neq j}^{m} h\left[(x_i, y_i), (x_j, y_j)\right]}{m(m-1)} = \frac{\sum_{i \neq j}^{m} \left[k(x_i, x_j) - k(x_i, y_j) - k(x_j, y_i) + k(y_i, y_i)\right]}{m(m-1)}$$

can be obtained, which is a one-sample U-statistic with

$$h[(x_i, y_i), (x_j, y_j)] = k(x_i, x_j) - k(x_i, y_j) - k(x_j, y_i) + k(y_i, y_i).$$

This test is based on the asymptotic distribution of MMD. Under the alternative hypothesis H_1, it has been proven in [11] that $MMD(X, Y, \mathcal{F})$ has an asymptotic normality with the variance $4\sigma^2/m$, where σ^2 can be estimated from data by

$$\mathop{\mathbb{E}}_{x,y} \left[\left[\mathop{\mathbb{E}}_{x',y'} \left[k(x, y), (x', y')\right] \right]^2 \right] - \left[\mathop{\mathbb{E}}_{x,y,x',y'} k\left((x, y), (x', y')\right) \right]^2.$$

The null distribution under H_0 can be found in the appendix of [2]. Note that, this test is consistent, and its test statistic takes $O(N^2)$ time to compute [11]. In terms of time complexity, it is expected that adopting MMD test is more efficient than FR test.

MMD test also aims at directly comparing two multi-dimensional point sets with arbitrary data distributions, since no density estimations are involved in its operational procedure. This is in contrast to some indirect approaches [2, 3] that first attempt to plug in the kernel density estimates $\hat{p} = \sum_{i=1}^{m} k(x_i, x)/m$ and $\hat{q} = \sum_{i=1}^{n} k(y_i, y)/n$ of two point sets and then compute some distance (e.g., L_1 used in [3] or

L_2 used in [2]) between $\hat{p}$ and $\hat{q}$. Those indirect approaches not only suffer from *empirical estimate bias* but also *dimensionality curse* which implies they cannot be applied to high-dimensional problems. Along with the feature map Φ induced by some kernel $k(\cdot,\cdot)$ from $\mathbb{R}^d$ to $\mathcal{H}$, some linear statistics in RKHS (i.e., the computations of mean function values via linearity) are equipped with the ability of capturing nonlinear data distributions in input space, without the need of explicitly estimating their densities. Therefore, it could provide more reliable results compared with PCA-based methods that take no local information into account and only exploit global linear correlations.

3.3 Distribution-based Similarity Search

The quantities of W and MMD in the above FR and MMD tests are originally proposed in the community of statistics to compare two multivariate distributions for the two-sample problems. In these tests, for distinguishing two hypotheses H_0 and H_1 of binary comparison results, there is usually a certain threshold which is in turn based on null distribution. If this threshold is exceeded, H_0 is rejected. Otherwise, it holds. For example in MMD, this test threshold can be derived either by using the permutation technique of bootstrap on the pooled sample points [8] or fitting Pearson curves to the first four moments [1] in order to evaluate a parameter α, which is a quantile of the null distribution of unbiased test statistic. The acceptance region of such a hypothesis test is defined as any real number below the certain threshold.

However, we can also extend this kind of test statistics as a collective gauge of relevance ranking specifically tailored for similarity search in a collection of multi-dimensional point sets, by only evaluating the probabilities that they are consistent with the same (but unknown) distribution of the query. When the quantity of difference is smaller, they are expected to be more similar. From this point of view, the objects in a repository can be directly ranked according to their distributional discrepancies as compared with a query baseline, without referring to the procedure of threshold inferring that depends on null distribution. A good distributional discrepancy can measure the degree of difference between two probability distributions, and is desired to truly quantify the intuition of relevance or not. In the sequel, we perform some experiments to compare different methods.

It is worthwhile to note that, our distributional discrepancy as a criterion of retrieving multi-dimensional point sets is not a metric, since it violates the metric properties. Nevertheless, it is not necessarily a disadvantage because the application of similarity search itself does not rely on the properties of metric distance functions. Indeed, non-metric distances can be more accurate for measuring the similarity of complex objects, as highlighted by Jacobs *et al.* in [13].

4. EVALUATION METHODOLOGY

In this section, we present the detailed experimental settings of our quantitative evaluation to study the performance of distribution-based similarity measures.

4.1 Test Collections and Query Sets

Our experiments are conducted on two different multimedia data collections, both of which contain series of multivariate observations. Note that each item in a dataset has the same number of columns, but may have different number of rows.

Transformation	Description
1	brightness decreased by 20%
2	brightness increased by 20%
3	saturation decreased by 20%
4	saturation increased by 20%
5	20% border pixels cropped
6	a small logo inserted at the top left corner
7	10% frames randomly dropped
8	10% noise frames randomly added

Table 2: Transformations applied to video clips.

The first one is a video clip collection originally consisting of 1,084 video clips which are TV commercials captured from free-to-air TV programmes. They were recorded at PAL frame rate of 25fps. The time length of each video clip is about 60 seconds, thus each of them consists of about 1,500 frames. We select 20 distinctive video clips which have one and only one similar clip existing in the original collection as the queries to be used for similarity search, and process each of them to make eight kinds of transformations with VirtualDub[2] utility, as listed in Table 2. The various distortions further produces a diversity of visual similarity that videos often exhibit. Therefore, by adding these 8×20 transformed video clips into the original collection, we have a test repository of 1,244 video clips in total. Meanwhile, the ground-truth set of each query can be collected as the one inherent similar and eight transformed video clips.

Without loss of generality, the most prominent representation of visual information - color histogram, is selected as a generally applicable method of extracting homologous low-level features to characterize distribution information, for the sake of achieving high performance without resorting to complicated frame representations. Statistically, it denotes the joint probability of RGB intensities describing the global color distribution of each video frame. In general, color histograms provide useful clues for similarity of frames, due to its robustness to background complications and object distortions. In addition, they are translation, scale and rotation invariant, and very simple to implement [26]. Searching similar videos can be achieved by the ranking of some distributional discrepancy, and thereby inherits these invariant characteristics. Four feature datasets in 8-, 16-, 32- and 64-dimensional RGB color spaces for this repository were generated for test purpose, which are similar to the datasets used in [24].

Another application we considered is that, given a series of multivariate observations generated by cyber-glove sensing device, finding the similar behaviors performed. Our second data collection is the publicly available high-quality recordings of Australian sign language dataset [15]. It consists of signs collected from a native Auslan signer using high-quality position trackers and instrumented gloves on both hands. Each position tracker provided six degrees of freedom - roll, pitch and yaw as well as x, y and z, and the gloves also provided a full five fingers of data (one right and one left, thus 22 attributes in total). The original dataset from UCI KDD archive contains 95 distinct signs, and each has 27 examples. A recording set from each of following 10 words were extracted as our test data: 'Norway', 'cold', 'crazy', 'eat', 'forget', 'happy', 'innocent', 'later', 'lose' and

[2]http://www.virtualdub.org

'spend'. In total, the number of signs gathered is 270. The average length of each sign is around 65. This is the same experiment also conducted in [27, 7], but there only two attributes x and y were used as low-dimensional trajectory data. We use all the 22 channels of information tracked in our experiments. Therefore, the number of variables of Auslan dataset is 22. Each item can be taken as a query sign, thus in total we have 270 different queries.

Note that although this labelled dataset is used, the classification is not of primary interest to us, which can be accomplished by model-based machine learning approaches, such as Hidden Markov Model (HMM) or Support Vector Machines (SVM). Generally, in most cases of similarity search, each item may not be associated with an explicitly prescribed label, since there is no natural underlying class information that can be easily assigned, e.g., usually the objects in multimedia information retrieval are not clustered.

4.2 Evaluation Metrics and Configurations

A 'leaving-one-out' kNN search is used in our experiments. That is, each time we take one item out as a query q and perform k nearest neighbor search in the test repository with varying k until all the r relevant items of q are retrieved. Relevant items are those with the same labels as q. Recall that for video clip repository each query has 9 relevant items, and for Auslan dataset each query sign has 26 relevant items. Besides the standard *precision-call* curve which is a user-oriented performance measure frequently used in information retrieval community, we report the system-orientated measure *Mean Average Precision* (MAP), and the average time of each similarity comparison. Given a query q with r relevant items (q itself is excluded from the ground-truth set), and let $rank_i$ be the rank of the ith retrieved relevant item ($1 \leq i \leq r$), then Average Precision (AP) over all relevant items is defined as

$$AP = \frac{1}{r} \sum_{i=1}^{r} \frac{i}{rank_i}.$$

It is average of precisions computed after truncating the ranked retrieval result after each of the relevant items in turn. AP favors returning more relevant item earlier. The mean value of the average precisions computed for an appropriate set of queries is obtained as the non-interpolated Mean Average Precision (MAP). MAP is a single-valued measure incorporating both precision and recall aspect information, thus can indicate the effectiveness of a retrieval method.

In order to clearly show the performance of distribution-based similarity measures, we compare FR and MMD methods with two representative element-based methods, Dynamic Time Warping (DTW) and Edit Distance (ED), and a correlation-based method Bounded Coordinate System (BCS). Next, we give some implementation details and parameter settings used in our experiments.

For *FR* method, Prim's algorithm is chosen for constructing MST, as it is more appropriate for dense graphs than another established alternative - Kruskal's algorithm. For *MMD* method, in effect, the original data are mapped into a higher-dimensional space using some possibly nonlinear kernel function $k(\cdot, \cdot)$. RBF and polynomial kernels are what kernel method practitioners widely pick from a set of standard kernels [14], thus they are implemented as two different members of function class $\mathcal{F}$ and tested in MMD. The first scheme we tested is to use Gaussian RBF kernel with automatic 'good' kernel width estimation. The width parameter σ which determines the radius of influenced area can be dynamically adjusted related to the expression of Gaussian RBF kernel

$$k(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2})$$

subject to that the exponent part $\|x - y\|/2\sigma^2$ equals 1 for the median distance between two points x and y of all distances from the pooled sample points of sets X and Y. The second scheme is to globally use the custom polynomial kernel given by

$$k(x, y) = (\langle x \cdot y \rangle + c)^m,$$

where the degree of polynomial m is set to be 2 and the offset c is set to be 1. Based on our observations, RBF kernel function should be used in preference to polynomial kernel for the experiments with video clip data, while MMD method with polynomial kernel performs better in the experiments conducted on Auslan dataset. Due to page limit, only the one with better results is presented in the corresponding set of experiments for MMD method. For *DTW* method, dynamic time warping distances with unconstrained width of warping path are computed, where the distance between two points, i.e., the local distance, is the square of Euclidean distance, as in [16]. For *ED* method, the threshold of inter-element Euclidean distance ϵ is set to be 0.3. Therefore, once ϵ is larger than 0.3, the elements being compared will be penalized as mismatched and require some edit operation. In addition, for these two methods of element pairwise comparisons based on Euclidean distance, Auslan dataset is normalized to render each column unit norm. Note that the value of each dimension of video clip datasets has already been normalized in extracting feature vectors. The computational cost of correlation-based analysis can be divided into two parts: generating a compact representation of the original $n \times d$ matrix which could be pre-processed offline, and online measuring similarity based on principal components. For *BCS* method, the elapsed time of different parts for each comparison will be reported separately.

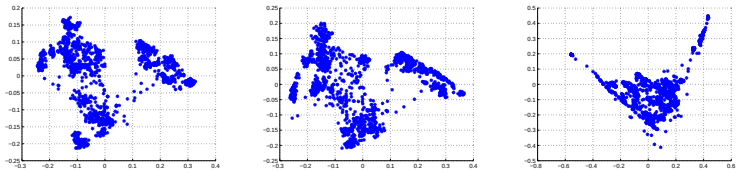
These five methods are all implemented in MATLAB™. Our experiments were performed on Window XP platform with Intel Core 2 CPU (2.4 GHz) and 2.0 GB RAM. All results reported are average numbers obtained over performing 20 and 270 queries to search against video clip repository and Auslan dataset, respectively.

5. EXPERIMENTAL RESULTS

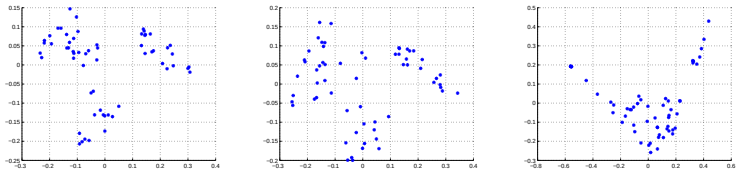
We report two sets of experiments to evaluate our proposal in different applications. In each of them, we study the performance with respect to several evaluation criteria, and the results show distribution-based similarity search methodology outperforms other competing methods in terms of retrieval quality, with the efficiency close to the best performing alternative. Some visualization examples are also presented.

5.1 Results of Video Clip Data

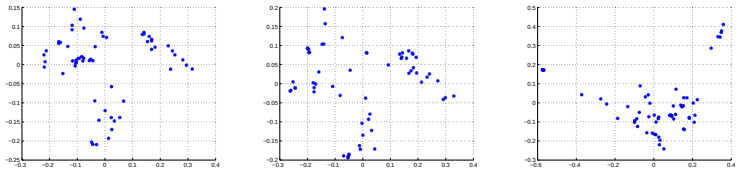
First, three vectorial distributions of video clips in feature space are exemplified in Figure 3(a). For facilitating visualization, the original 32-dimensional feature vectors are projected with the first two PCA coefficients. Each point in the figures stands for a frame. Actually video clips A



(a) Similar video clips A and B, and a dissimilar video clip C.



(b) Systematic sampling, uniformly select every 25th element.



(c) Simple random sampling, select 60 elements (without replacement).

Figure 3: Vectorial distributions in feature space.

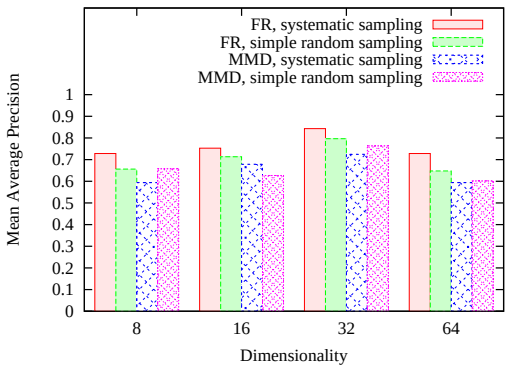


Figure 4: MAP vs. dimensionality.

and B are similar ones, while video clip C is different. It becomes evident that the distributions of similar videos are with much likeness, while dissimilar one is quite different. This example also confirms that the relevance of objects can be judged by distribution comparison. Considering the continuity of videos, content evolution within a second is often negligible. For video clip data, we consider a straightforward *systematic sampling* which uniformly selects the every 25th element from the original sequences of video frames, as well as a *simple random sampling* (without replacement) in which each element has an equal probability to be selected. Figure 3(b) and Figure 3(c) plot the selected points of two different sampling techniques, respectively. Experimentally, we observe that satisfactory comparison results can be achieved even when small to moderate sample size of representative points are used, thus the computational cost of distribution-based methods can be alleviated significantly.

Method \ Dimensionality	8	16	32	64
FR	0.0488	0.0492	0.0500	0.0516
MMD	0.0117	0.0121	0.0129	0.0145

Table 3: Average comparison time (sec.)

Next, we test the effect of different sampling techniques used in FR and MMD methods. Meanwhile, we are interested in studying the performance with respect to dimensionality by using all four feature datasets. The dimensionality of video clip data is up to 64. We have conducted some tests to investigate the practical effectiveness of measuring set similarity from samples by comparing the MAP values achieved based on different sample sizes. The results suggest that large sample setting may not necessarily improve the quality of relevance ranking. Therefore, we either uniformly select the every 25th element or randomly select 60 samples among each original sequence of frames. From the MAP values shown in Figure 4, it seems that systematic sampling is a more appropriate technique here than simple random sampling, especially for FR method which consistently shows higher MAP than MMD method for all dimensions. This is probably because systematic sampling can decrease the dependence of frames and select a more representative set of sample points from the full ensemble. Based on the selected elements, the set similarity can be understood. Another interesting observation from Figure 4 is that, the MAP value for 64-dimensional dataset is smaller than 32-dimensional. This suggests that for higher dimensions, the number of samples in each set could be even smaller than the dimensionality d, which implies that the space $\mathbb{R}^d$ is nearly empty. Therefore, distribution-based similarity measures are reliable in high-dimensions only with moderate sample size.

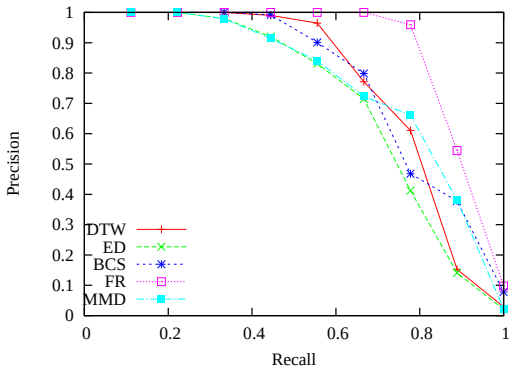


Figure 5: Precision vs. recall (video clip).

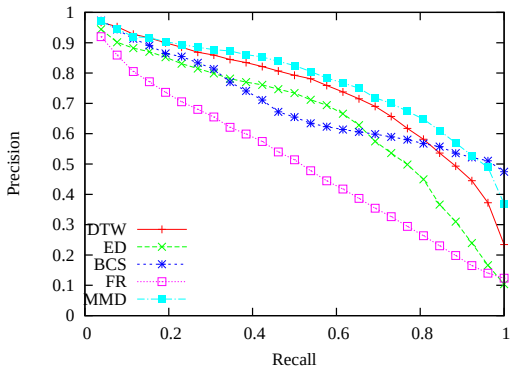


Figure 6: Precision vs. recall (Auslan).

Method	MAP	Average comparison time (sec.)
DTW	0.7243	0.0106
ED	0.6682	0.0809
BCS	0.7351	0.0527 (offline) + 0.0015 (online)
FR	**0.8431**	0.0500
MMD	0.7246	0.0129

Table 4: Performance comparison (video clip).

Method	MAP	Average comparison time (sec.)
DTW	0.7307	0.0136
ED	0.6397	0.0663
BCS	0.6942	0.0169 (offline) + 0.0032 (online)
FR	0.4925	0.0461
MMD	**0.7682**	0.0150

Table 5: Performance comparison (Auslan).

Table 3 reports the computational time per comparison of FR and MMD methods (independent of the underlying sampling technique). FR method is slower than MMD method as expected. However, for the experiments with video clip data, its retrieval quality is better.

In the following comparison study with other competing methods, the default dimensionality is 32. Since DTW and ED take temporal order into account, only systematic sampling technique can be applied. Therefore, in computing DTW and ED feature vectors are obtained by uniformly selecting the every 25th element as above, which also has an effect of reducing computational cost. The performance comparison results of 32-dimensional video clip dataset are summarized in Figure 4 and Table 4. A notable impression of Figure 4 is that the precision of FR method is much higher compared with others for a relatively large recall. From Table 3, we see more clearly even though FR does not gain much in efficiency, its retrieval quality outperforms other methods. As verified by comparing the precision-call curves as well as the values of MAP, BCS using solely linear correlation information cannot reflect the real distributions of all video clip data very accurately, and indeed retrieval quality can benefit from exploiting more descriptive local information. FR and BCS methods are both robust to temporal changes. Although this dataset does not contain excessive noise, suboptimal results of DTW and ED methods indicate that element-to-element comparisons could be unreliable due to the accumulation of all the pairwise differences.

5.2 Results of Auslan Data

Distribution-based similarly measures are generally applicable to multi-dimensional point set retrieval applications, which is exemplified by our additional experiments with Auslan dataset. Since the average length of signs is 65, for this relatively short dataset we do not use sampling. Therefore, the five methods to be compared directly deal with all the elements of individual sets for ranked retrieval.

Similar to video clip data, the performance comparison results of Auslan data are summarized by the precision-call curves in Figure 6 as well as the values of MAP and average comparison time in Table 5. Figure 6 shows MMD method dominates the conventional similarity measures DTW and ED as well as correlation-based analysis BCS. This implies that for this dataset, kernel method can better capture the nonlinear distributions in input space. From Table 5, we can further see MMD outperforms other competing methods in terms of MAP, with the efficiency comparable to the best performing alternative DTW. We observe that BCS is less effective when used for this dataset. This is reasonable since unlike video clip data, each sign of Auslan dataset does not exhibit high degree of correlation (i.e., tendency). Therefore, the MAP of BCS is not as good as that of DTW in this application. The performance of FR method is unappealing for this dataset. This could be explained as this dataset actually contains 22 different channels of information, and $\mathbb{R}^d$ is not an orthogonal vector space. FR method which directly exploits $\mathbb{R}^d$ space does not perform well.

6. CONCLUSIONS AND PERSPECTIVES

This paper presents a novel collective gauge of multi-dimensional point set similarity based on the distributional discrepancy of (selected) elements. We leverage two rigorous non-parametric tests in the literature of multivariate statistics for checking the hypothesis whether two ensembles of points are from a same distribution. The main contribution of our work is to make them adapt to ranked retrieval in a collection of point set representations as more direct, reliable and general similarity measures, which is particularly devoted to addressing fitting any arbitrary form of data distribution. The experimental results demonstrate the superiority of our proposal, while the potentials of our methodology go beyond these two multimedia applications.

In the future, we plan to carefully consider the strengths and weaknesses of various other hypothesis tests. For exam-

ple, an exact distribution-free multivariate test proposed by the statistics community is recently published [22], which is based on minimum distance matching over a graph. However, it requires $O(N^3)$ time complexity, thus seems too computationally laborious to be utilized for retrieval scenarios. Further investigation of the scalability of these distribution-based similarity measures to large datasets using indexing structures such as inverted files is also needed.

7. ACKNOWLEDGMENTS

This work is supported by ARC grant DP0663272. The authors would like to thank the anonymous reviewers for their insightful comments, which led to improvements of this paper.

8. REFERENCES

[1] Pearson curves. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*. Springer-Verlag, 2002. http://eom.springer.de/P/p071920.htm.

[2] N. H. Anderson, P. Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivar. Anal.*, 50(1):41–54, 1994.

[3] G. Biau and L. Györfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

[4] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *ISMB (Supplement of Bioinformatics)*, pages 49–57, 2006.

[5] R. Cappelli, D. Maio, and D. Maltoni. Multispace kl for pattern representation and classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(9):977–996, 2001.

[6] B. Chalmond and S. C. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):422–432, 1999.

[7] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD Conference*, pages 491–502, 2005.

[8] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993.

[9] W. Feller. On the kolmogorov-smirnov limit theorems for empirical distributions. *Ann. Math. Statist.*, 19(2):177–189, 1948.

[10] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist.*, 7(4):697–717, 1979.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2006.

[12] N. Henze and M. D. Penrose. On the multivariate runs test. *Ann. Statist.*, 27(1):290–298, 1999.

[13] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(6):583–600, 2000.

[14] T. Joachims. Support vector and kernel methods. In *SIGIR, Tutorial*, 2003.

[15] M. W. Kadous. *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series.* PhD thesis, The University of New South Wales, October 2002. http://www.cse.unsw.edu.au/~waleed/phd.

[16] E. J. Keogh. Exact indexing of dynamic time warping. In *VLDB*, pages 406–417, 2002.

[17] V. S. Korolyuk and Y. Borovskich. *Theory of U-Statistics.* Kluwer Academic Publishers, Dordrecht, 1994.

[18] W. J. Krzanowski. Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703–707, 1979.

[19] S.-L. Lee, S.-J. Chun, D.-H. Kim, J.-H. Lee, and C.-W. Chung. Similarity search for multidimensional data sequences. In *ICDE*, pages 599–608, 2000.

[20] C. Li, P. Zhai, S.-Q. Zheng, and B. Prabhakaran. Segmentation and recognition of multi-attribute motion sequences. In *ACM Multimedia*, pages 836–843, 2004.

[21] J. Puzicha, T. Hofmann, and J. M. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *CVPR*, pages 267–272, 1997.

[22] P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of The Royal Statistical Society Series B*, 67(4):515–530, 2005.

[23] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84(1):25–43, 2001.

[24] H. T. Shen, X. Zhou, Z. Huang, and J. Shao. Statistical summarization of content features for fast near-duplicate video detection. In *ACM Multimedia*, pages 164–165, 2007.

[25] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[26] C. Theoharatos, V. K. Pothos, N. A. Laskaris, G. Economou, and S. Fotopoulos. Multivariate image similarity in the compressed domain using statistical graph matching. *Pattern Recognition*, 39(10):1892–1904, 2006.

[27] M. Vlachos, D. Gunopulos, and G. Kollios. Discovering similar multidimensional trajectories. In *ICDE*, pages 673–684, 2002.

[28] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *Ann. Math. Statist.*, 11(2):147–162, 1940.

[29] K. Yang and C. Shahabi. An efficient k nearest neighbor search for multivariate time series. *Inf. Comput.*, 205(1):65–98, 2007.

[30] W. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia*, 9(5):1037–1048, 2007.