Order-preserving Wasserstein Distance for Sequence Matching

Bing Su[1], Gang Hua[2]
[1]Science & Technology on Integrated Information System Laboratory,
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[2]Microsoft Research
{subingats, ganghua}@gmail.com

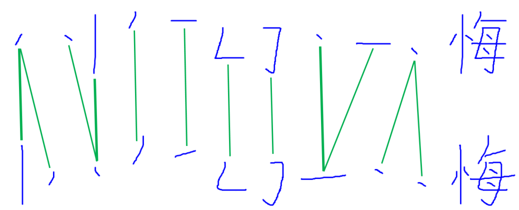
Abstract

We present a new distance measure between sequences that can tackle local temporal distortion and periodic sequences with arbitrary starting points. Through viewing the instances of sequences as empirical samples of an unknown distribution, we cast the calculation of the distance between sequences as the optimal transport problem. To preserve the inherent temporal relationships of the instances in sequences, we smooth the optimal transport problem with two novel temporal regularization terms. The inverse difference moment regularization enforces transport with local homogeneous structures, and the KL-divergence with a prior distribution regularization prevents transport between instances with far temporal positions. We show that this problem can be efficiently optimized through the matrix scaling algorithm. Extensive experiments on different datasets with different classifiers show that the proposed distance outperforms the traditional DTW variants and the smoothed optimal transport distance without temporal regularization.

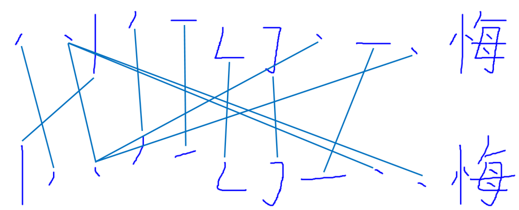
1. Introduction

Effectively measuring the distance between sequences plays a fundamental role in a wide range of sequence pattern analysis problems. Due to the inherent complexity of sequence data, defining distance measures for sequences can be much more difficult than for vectors. First of all, the evolution speed of instances in different sequences may be quite different. For example, different subjects may perform the same action with different speeds. The sampling rate for different sequences may also be different. As a result, although the instances in sequences have a fixed dimensionality, different sequences generally have different numbers of instances. Therefore, conventional distance measures for vectors such as Euclidean distance, L_p norms, and Mahalanobis distance cannot be directly applied to sequences.

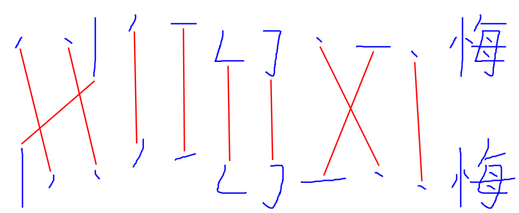
Secondly, the evolutions of instances in sequences are



(a)



(b)



(c)

Figure 1. The two stroke sequences of the same online character differ in local orders at the first and last three positions. (a) The alignments generated by DTW are disturbed by such local order reversal since DTW preserves the temporal positions strictly, where the misalignments are shown in bold. (b) If only the stroke shape is considered, Sinkhorn (smoothed OT) aligns the same stroke "dot" appeared in different temporal positions. (c) The proposed OPW is able to generate proper alignments that can tackle local order distortion by jointly considering the shape matching and temporal orders. The alignments shown in (b) and (c) are the largest transports associated with instances of the second sequence in the learned OT.

not uniform. For example, when performing the same action of "kicking a ball", one person may spend more time on the run-up and the other may keep the leg raised longer. As a result, instances of different sequences in the same temporal position may correspond to different poses. Therefore, temporal alignments are necessary when comparing different sequences. In addition, local temporal ordering of instances in sequences describing the same pattern may also vary.

We show an example in Fig. 1(a). The correct order of strokes to write the Chinese character is shown at the top. A very common wrong stroke order of this character is shown in the bottom, where the orders of the dots are reversed at the beginning and close to the end. This means that the strict rank preservation may not be imposed to the alignments.

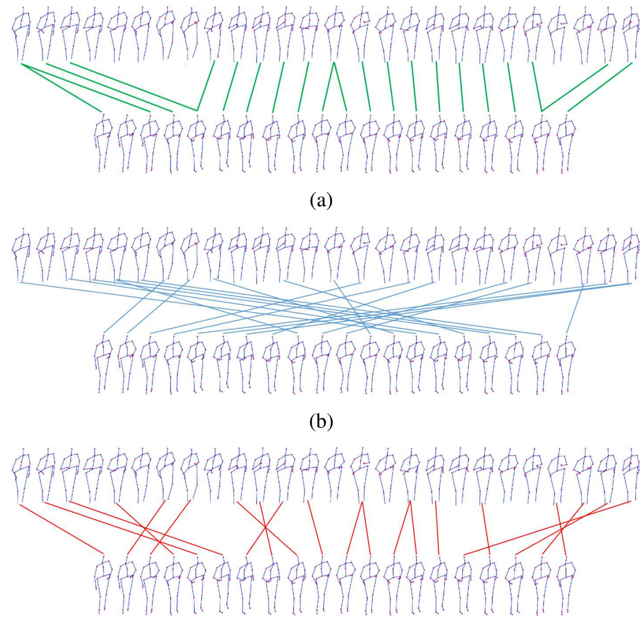
Thirdly, the instances in the same sequence are not independent samples, but are temporally related. For example, without considering the orders of frames, the two actions "stand up" and "sit down" performed by the same person cannot be distinguished. This prevents the application of distances between sets, distributions or bags such as the Jaccard index, the Chernoff distance [11, 28] and the KL divergence [15] to sequences.

Last but not least, for periodic patterns or events, different sequence samples may start from different instances. For example, as shown in Fig. 2, both performing the action "jogging", one person may start with lifting the left leg while stretching the right arm and another person may start with lifting the right leg while stretching the left arm. The starting (ending points) of different sequences cannot be forced to align and flexible alignments are required.

A lot of efforts have been made to find meaningful distance measures between sequences that can tackle these issues. *Dynamic time warping (DTW)* [22] is perhaps the most widely adopted. DTW is able to align sequences with different lengths, speeds, and non-homogeneity. However, the alignments determined by DTW preserve the orders strictly, *i.e.*, no instance in one sequence is allowed to align to instances in the other sequence before the instance aligned in the previous step. As shown in Fig. 1, when local reverses or temporal distortions exist, erroneous alignments are unavoidable, which also affect other regular alignments. The boundary condition of DTW requires the starting and ending points of the two sequences be aligned. This will lead to erroneous alignments for periodic sequences with different starting points as shown in Fig. 2(a).

By viewing the instances in a sequence as empirical samples from a probability distribution or a set of vector points, *optimal transport (OT)* [31] or its smoothed and computationally efficient version, *the Sinkhorn distance* [9], provides a canonical way to automatically lift the geometry between instances to define a distance measure for two sequences, which has many excellent properties such as existence of optimal maps, separability, and completeness.

The learned transport naturally defines flexible alignments between two sets. Generally, the optimal transport only assigns large weights to the most similar instance pairs.



(a)

(b)

(c)

Figure 2. The two periodic sequences of the "jogging" action start from different states. The right arm is first lifted in the top sequence while the left arm is first lifted in the bottom sequence. (a) Misalignments are introduced by DTW due to the strict temporal constraint. (b) Sinkhorn aligns the frames representing the most similar poses even when the relative temporal positions of the two frames are very far. As a result, long connection lines occur frequently in the alignments. The first few frames of one sequence are aligned to the rear frames of the other sequence across a complete periodic cycle. (c) The proposed OPW aligns each frame in one sequence to the frame with the same pose of the temporally adjacent cycle in the other sequence. The generated alignments exhibit periodic segments.

Although OT can solve the local rank reverses and different starting points, it fully ignores the inherent temporal dependencies of instances. As shown in Fig 1(b), if only the shape features are considered, the two dots in the right parts are largely transported to the dot in the left radical. The poses are mainly transported to the most similar poses in distant cycles in Fig. 2(b).

To incorporate the advantages of flexibility of optimal transport and order preserving alignments, we develop a new distance measure between sequences by imposing temporal constraints on the optimal transport, namely *Order-Preserving Wasserstein (OPW) distance*. The resulting metric inherits some excellent properties of OT while capturing the general temporal orders of instances, leading to flexible temporal-sensitive alignments. As shown in Fig. 1(c), and Fig. 2(c), the proposed distance measure can properly address these issues which are problematic for DTW and OT.

The main contributions of this paper are summarized as follows: 1. we adapt the optimal transport as the basis for distance measures between sequences, which has many excellent mathematic properties and is robust to local rank reverses and periodic patterns with different starting points; 2. we propose two novel temporal regularizations to punish transports or alignments between instances with distant temporal positions, such that the learned optimal transport is capable of preserving the temporal dependencies of instances in sequences; 3. we show that the regularized OT problem can be efficiently solved by Sinkhorn's matrix scaling algorithm [26]. We empirically demonstrate that the proposed OPW distance outperforms DTW and Sinkhorn distances on different tasks with different classifiers.

2. Related Work

Most studies on distance measures for sequences mainly focus on improving DTW. In [19], a locality constraint is used to constrain the amount of warping. Various methods are developed either to speed up the computation of DTW, such as the FastDTW [23] and the SparseDTW [1], or to accelerate the calculation of all-pairwise DTW matrix [25]. Originally, the Longest Common Subsequence (LCSS) distance [6] and the edit distance [18] are designed to compare string sequences. Some efforts [32, 17] attempt to extend them to handle continuous multi-dimensional sequences, and lead to DTW-like algorithms. Canonical Time Warping [36] and generalized time warping [35] extend DTW to handle multi-modal sequences whose instances may have different dimensions. These methods are generally DTW-based and suffer from similar issues of DTW.

In [20], a new weight sequence is generated by mapping the original sequence to the learned semi-continuous HMM and extracting the mixture weight vectors of states. The distance between the original sequences is defined as the DTW distance between the weight sequences. In [27, 29], the HMM-based statistics are extracted from each set of sequences. The DTW distance between the statistics is used as the distance measure between sets of sequences or sequence classes. These methods are also DTW-based and a pre-training of HMM is needed.

In [30], based on the transported square-root vector field representation, a rate-invariant distance is derived for trajectories, which is further applied to action recognition in [3]. Similar to DTW, this distance is also strictly order-preserving. In contrast, the proposed method in this paper allows local reorders, which can tackle local temporal irregular evolutions and outlier frames, while does not affect the distinguish of sequences from different classes. Generally, the methods for learning alignments such as [34, 14] not only require supervised learning with ground-truth alignments or class labels, but also do not directly generate a distance measure. In this paper, we develop an unsupervised distance measure for any types of sequences. The distance between any two sequences can be directly computed without either supervised or unsupervised training from other sequences.

Recently, the optimal transport is receiving growing attentions, such as the fast computation [9, 4], the application of computing barycenters [10], generating of PCA [24], and loss definition [13]. To our knowledge, we are the first to adapt OT as the similarity measure for sequences.

3. Background on Optimal Transport

Optimal transport [31], also known as *Wasserstein distance*, measures the dissimilarity between two probability distributions over a metric space. Intuitively, if each distribution is viewed as a way of piling up a unit amount of dirt on the space, the Wasserstein distance is the minimum cost of transporting the pile of one distribution into the pile of the other distribution. Therefore, the Wasserstein distance is also known as the earth mover's distance [21].

Formally, given a complete separable metric space (Ω, d), where $d : \Omega \times \Omega \to \mathbb{R}$ is the metric on the space Ω, let $P(\Omega)$ denotes the set of all Borel probability measures on Ω. Given two sets $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M)$ of sample points in Ω, their corresponding empirical probability measures can be estimated as $f = \sum_{i=1}^{N} \alpha_i \delta_{\boldsymbol{x}_i} \in P(\Omega)$ and $g = \sum_{j=1}^{M} \beta_j \delta_{\boldsymbol{y}_j} \in P(\Omega)$, respectively, where $\delta_{\boldsymbol{x}}$ is the Dirac unit mass on the position of $\boldsymbol{x}$ in Ω, N and M are the numbers of points in the two sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. α_i is the weight on the unit mass on $\boldsymbol{x}_i$.

Since f is a probability distribution, the corresponding weight vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)$ lies in the simplex $\Theta_N := \{\boldsymbol{\alpha} \in \mathbb{R}^N | \alpha_i \geq 0, \forall i = 1, \cdots, N, \sum_{i=1}^{N} \alpha_i = 1\}$. Similarly, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M) \in \Theta_M$. Without any prior knowledge on the samples, the points can be viewed as uniformly sampled from each distribution, and the weights can be estimated as $\boldsymbol{\alpha} = (\frac{1}{N}, \cdots, \frac{1}{N})$ and $\boldsymbol{\beta} = (\frac{1}{M}, \cdots, \frac{1}{M})$, respectively.

The empirical joint probability measure of $(\boldsymbol{X}, \boldsymbol{Y})$ can be estimated as

$$h = \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma_{ij} (\delta_{\boldsymbol{x}_i}, \delta_{\boldsymbol{y}_j}), \tag{1}$$

whose marginal measures w.r.t. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are f and g, respectively. Thus the weight matrix $[\gamma_{ij}]_{ij}$ is a $N \times M$ nonnegative matrix with row and column marginals $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The set of all the feasible weight matrices is defined as the transportation polytope $U(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$U(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \{\boldsymbol{T} \in \mathbb{R}_+^{N \times M} | \boldsymbol{T}\mathbf{1}_M = \boldsymbol{\alpha}, \boldsymbol{T}^T\mathbf{1}_N = \boldsymbol{\beta}\}. \tag{2}$$

An element t_{ij} of a feasible $\boldsymbol{T}$ can be viewed as the amount of mass transported from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$. The distance between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is measured by the metric d raised to the power p. All the pairwise distances between elements in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are stored in the matrix $\boldsymbol{D}$, *i.e.*,

$$\boldsymbol{D} := [d(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij} \in \mathbb{R}^{N\times M}. \tag{3}$$

The cost of transporting f to g given a transport $\boldsymbol{T}$ is $\langle \boldsymbol{T}, \boldsymbol{D}\rangle$, where $\langle \boldsymbol{T}, \boldsymbol{D}\rangle = tr(\boldsymbol{T}^T\boldsymbol{D})$ is the Frobenius dot product. The p-th Wasserstein distance raised to the power p between empirical probability measures f and g can be formulated as

$$W_p^p(f,g) = d_W(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{D}) = \min_{\boldsymbol{T}\in U(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D}\rangle, \tag{4}$$

where $W_p^p(f,g)$ is a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$, and hence can also be written as $d_W(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{D})$. We only consider $p=1$ in this paper and omit p hereafter for simplicity.

Computationally, it is quite expensive to obtain the optimal solution of (4). Recently, Cuturi [9] has proposed to add an entropy constraint to the transportation polytope, which turns out to be an entropy regularized optimal transport problem, resulting in *the Sinkhorn distance*, *i.e.*,

$$\begin{aligned} &d_S^\lambda(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{D}) = \langle \boldsymbol{T}^\lambda, \boldsymbol{D}\rangle \\ &s.t.\ \boldsymbol{T}^\lambda = \mathop{arg\min}_{\boldsymbol{T}\in U(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D}\rangle - \tfrac{1}{\lambda}h(\boldsymbol{T}) \end{aligned}, \tag{5}$$

where $h(\boldsymbol{T}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} t_{ij}\log t_{ij}$ is the entropy of $\boldsymbol{T}$. The optimal $\boldsymbol{T}^\lambda$ that minimizes (5) has a simple form, *i.e.*,

$$\boldsymbol{T}^\lambda = diag(\boldsymbol{\kappa}_1)\boldsymbol{e}^{-\lambda\boldsymbol{D}}diag(\boldsymbol{\kappa}_2), \tag{6}$$

where $\boldsymbol{e}^{-\lambda\boldsymbol{D}}$ is the element-wise exponential of the matrix $-\lambda\boldsymbol{D}$, $\boldsymbol{\kappa}_1 \in \mathbb{R}^N$ and $\boldsymbol{\kappa}_2 \in \mathbb{R}^M$ are the non-negative left and right scaling vectors which are unique up to multiplying a factor. $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ can be efficiently determined by the Sinkhorns fixed point iterations. Therefore, the computational complexity is greatly reduced compared with the original problem (4).

4. Order-Preserving Wasserstein Distance

Given two sequences $\boldsymbol{X} = [\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1,\cdots,\boldsymbol{y}_M]$ with lengths N and M, respectively, the Wasserstein distance can be applied as a distance measure between them by viewing the elements in each sequence as independent samples, *i.e.*,

$$d_O(\boldsymbol{X},\boldsymbol{Y}) = \min_{\boldsymbol{T}\in U(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D}\rangle \tag{7}$$

In this case, each sequence is considered as a set of points sampled independently from a distribution, and hence $\boldsymbol{\alpha} = (\frac{1}{N},\cdots,\frac{1}{N})$ and $\boldsymbol{\beta} = (\frac{1}{M},\cdots,\frac{1}{M})$. The Wasserstein distance measures minimum cost of transporting the distribution of elements $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N$ in $\boldsymbol{X}$ into the distribution of elements $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_M$ in $\boldsymbol{Y}$, but the ordering relationship of these elements is totally ignored. As shown in Fig. 1(b) and Fig. 2(b), the first instance of one sequence may be matched (transported) to the last but one instance of the other sequence. Therefore, the Wasserstein distance can only measure the divergence between the spatial distributions of the elements, but is incapable of distinguishing the temporal orders of the elements. In cases two sequences only differ in the order of the elements, neither the Wasserstein distance nor the Sinkhorn distance can separate them.

To take the inherent temporal information into account, it is desired that the sample in one temporal position of one sequence can only be transported to the elements in the nearby temporal positions of the other sequence. That is, elements with relatively far temporal orders in the two sequences cannot be matched. We use $\frac{i}{N}$ to measure the relative temporal order or position of the element $\boldsymbol{x}_i$ in the sequence $\boldsymbol{X}$. Recall that N is the length of the sequence $\boldsymbol{X}$. All the increasing relative temporal positions of all elements in $\boldsymbol{X}$ form an order-prior sequence $\boldsymbol{O_X}$ associated with a sequence $\boldsymbol{X}$, *i.e.* ,

$$\boldsymbol{O_X} = [\frac{1}{N},\frac{2}{N},\cdots,\frac{N-1}{N},1].$$

If elements in one sequence are transported into those elements in the other sequence with similar relative temporal positions, the transport matrix $\boldsymbol{T}$ should show *local homogeneous* structures. That is, large values appear in the area near the diagonal of $\boldsymbol{T}$, while the values in other areas of $\boldsymbol{T}$ are zero or very small. The *inverse difference moment* [2] of the transport matrix $\boldsymbol{T}$ measuring such local homogeneity is:

$$I(\boldsymbol{T}) = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{t_{ij}}{(\frac{i}{N}-\frac{j}{M})^2+1}, \tag{8}$$

where $I(\boldsymbol{T})$ will have a large value if the large values of the transport $\boldsymbol{T}$ are mainly distributed along the diagonal. Ideally, maximizing $I(\boldsymbol{T})$ w.r.t. $\boldsymbol{T}$ over $U(\boldsymbol{\alpha},\boldsymbol{\beta})$ without any other constraint will result in a matrix whose non-zero values only appear in positions where $\frac{i}{N}=\frac{j}{M}$. To encourage temporally approached elements to match, the inverse difference moment $I(\boldsymbol{T})$ of the leaned $\boldsymbol{T}$ should be as large as possible.

A general ideal distribution of the values in $\boldsymbol{T}$ is that the peaks appear on the diagonal, and the values decrease gradually along the direction perpendicular to the diagonal. This can be modeled by a two-dimensional distribution, whose marginal distribution along any line perpendicular to the diagonal is a Gaussian distribution centered at the intersection on the diagonal, *i.e.*,

$$p_{ij} := \boldsymbol{P}(i,j) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\ell^2(i,j)}{2\sigma^2}}, \tag{9}$$

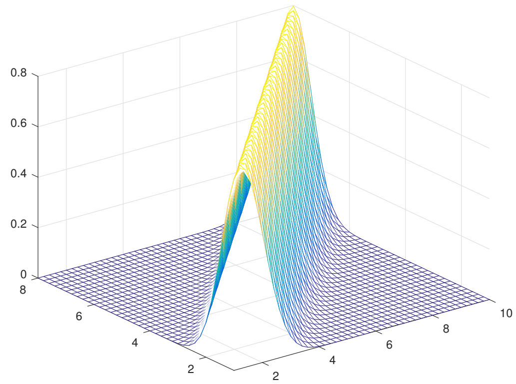


Figure 3. The Prior distribution of the transport matrix.

where $\ell(i, j)$ is the distance from the position (i, j) to the diagonal line, *i.e.*,

$$\ell(i, j) = \frac{|i/N - j/M|}{\sqrt{1/N^2 + 1/M^2}}.$$

We use Equation (9) as the *Prior distribution* of values in $\boldsymbol{T}$, as illustrated in Fig. 3. As can be observed, the farther one element in one sequence is from the other element in another sequence in terms of temporal orders, the less likely it is transported to that element. Since the values in $\boldsymbol{T}$ can be considered as representing the proportion of transporting the pile in one temporal position of the sequence to the corresponding temporal position in the other sequence, the distribution of values in the learned $\boldsymbol{T}$ and the prior distribution should be as similar as possible to encourage smooth and reasonable assignment or transportation.

To encourage that elements in different sequences with similar temporal positions be matched and the assignments of such matches be smooth, we introduce the following feasible set of the transport matrix $\boldsymbol{T}$ by imposing two additional constraints to the set $U(\boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$\begin{array}{l} U_{\xi_1,\xi_2}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\boldsymbol{T} \in R_+^{N\times M} | \boldsymbol{T}\mathbf{1}_M = \boldsymbol{\alpha}, \\ \boldsymbol{T}^T\mathbf{1}_N = \boldsymbol{\beta}, I(\boldsymbol{T}) \geq \xi_1, KL(\boldsymbol{T}||\boldsymbol{P}) \leq \xi_2\} \end{array}, \tag{10}$$

where $KL(\boldsymbol{T}||\boldsymbol{P}) = \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M} t_{ij} \log \frac{t_{ij}}{p_{ij}}$ is the Kullback-Leibler (KL) divergence between two matrices. This means that a transport matrix $\boldsymbol{T}$ is feasible only if its inverse difference moment is constraint to lie above a pre-defined threshold, while the KL divergence between $\boldsymbol{T}$ and the prior distribution $\boldsymbol{P}$ cannot exceed another pre-defined threshold.

Consequently, the elements in two sequences with very different temporal positions are unlikely to be matched by $\boldsymbol{T}$, and the transport proportions of elements in two sequences prefer to decrease with the increase of their temporal distance according to a Gaussian function. It can be easily observed that $U_{\xi_1,\xi_2}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a convex set.

We define the *Order-Preserving Wasserstein (OPW) Distance* as a distance measure between two sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$d^O_{\xi_1,\xi_2}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{T}\in U_{\xi_1,\xi_2}(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D}\rangle \,. \tag{11}$$

From the modeling perspective, this formulation is similar to maximizing the inverse difference moment of $\boldsymbol{T}$, minimizing the KL-divergence of $\boldsymbol{T}$ and the prior $\boldsymbol{P}$, while requiring the transport cost to be constrained. Since only elements with similar relative positions are preferred to be matched, the ordering relationships of the two sequences are preserved when calculating the distance. Moreover, similar to the entropy-smoothed Wasserstein distance, the additional constraints greatly reduce the computational complexity of calculating the optimal transport.

We consider the dual order-preserving Wasserstein distance by introducing two Lagrange multipliers for the inverse difference moment constraint and the KL-divergence constraint, *i.e.*, $\lambda_1 > 0$ and $\lambda_2 > 0$, respectively:

$$\begin{array}{l} d_O^{\lambda_1,\lambda_2}(\boldsymbol{X}, \boldsymbol{Y}) := \left\langle \boldsymbol{T}^{\lambda_1,\lambda_2}, \boldsymbol{D} \right\rangle \\ s.t.\ \boldsymbol{T}^{\lambda_1,\lambda_2} = \underset{\boldsymbol{T}\in U(\boldsymbol{\alpha},\boldsymbol{\beta})}{arg\min}\ \langle \boldsymbol{T}, \boldsymbol{D}\rangle - \lambda_1 I(\boldsymbol{T}) + \lambda_2 KL(\boldsymbol{T}||\boldsymbol{P}) \end{array} \tag{12}$$

According to the duality theory, for each pair ξ_1, ξ_2 in Equation (11), there exists a corresponding pair $\lambda_1 > 0$, $\lambda_2 > 0$ such that $d_O^{\lambda_1,\lambda_2}(\boldsymbol{X}, \boldsymbol{Y}) = d^O_{\xi_1,\xi_2}(\boldsymbol{X}, \boldsymbol{Y})$ is gained for the pair $(\boldsymbol{X}, \boldsymbol{Y})$. The two constraints can be viewed as regularization terms in (12).

$\boldsymbol{T}^{\lambda_1,\lambda_2}$ denotes the optimal transport matrix of the constraint in Eq. (12), *i.e.*, it optimizes

$$\begin{array}{l} \min\limits_{\boldsymbol{T}\in\mathbb{R}_+^{N\times M}} \langle \boldsymbol{T}, \boldsymbol{D}\rangle - \lambda_1 I(\boldsymbol{T}) + \lambda_2 KL(\boldsymbol{T}||\boldsymbol{P}) \\ s.t.\ \boldsymbol{T}\mathbf{1}_M = \boldsymbol{\alpha}, \boldsymbol{T}^T\mathbf{1}_N = \boldsymbol{\beta} \end{array} . \tag{13}$$

Since both the objective and the feasible set in (13) are convex, the optimal $\boldsymbol{T}^{\lambda_1,\lambda_2}$ exists and is unique. To obtain the optimal $\boldsymbol{T}^{\lambda_1,\lambda_2}$, we start from the Lagrangian function of Equation (13)

$$\begin{array}{l} L(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M}(d_{ij}t_{ij} - \lambda_1 \frac{t_{ij}}{(\frac{i}{N}-\frac{j}{M})^2+1} \\ + \lambda_2 t_{ij} \log \frac{t_{ij}}{p_{ij}}) + \boldsymbol{\mu}^T(\boldsymbol{T}\mathbf{1}_M - \boldsymbol{\alpha}) + \boldsymbol{\nu}^T(\boldsymbol{T}^T\mathbf{1}_N - \boldsymbol{\beta}) \end{array}, \tag{14}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are the dual variables for the two equality constraints $\boldsymbol{T}\mathbf{1}_M = \boldsymbol{\alpha}$ and $\boldsymbol{T}^T\mathbf{1}_N = \boldsymbol{\beta}$, respectively. The derivative of $L(\boldsymbol{T}, \boldsymbol{\mu}, \boldsymbol{\nu})$ w.r.t. t_{ij} for a couple (i, j) is

$$\begin{array}{l} \frac{\partial L(\boldsymbol{T},\boldsymbol{\mu},\boldsymbol{\nu})}{\partial t_{ij}} = d_{ij} - \frac{\lambda_1}{(\frac{i}{N}-\frac{j}{M})^2+1} + \lambda_2 \log \frac{t_{ij}}{p_{ij}} \\ \qquad + \lambda_2 + \mu_i + \nu_j \end{array} . \tag{15}$$

Setting $\frac{\partial L(\boldsymbol{T},\boldsymbol{\mu},\boldsymbol{\nu})}{\partial t_{ij}}$ to zero, we get:

$$t_{ij} = p_{ij} e^{-\frac{1}{2}-\frac{\mu_i}{\lambda_2}} e^{\frac{1}{\lambda_2}(s_{ij}^{\lambda_1} - d_{ij})} e^{-\frac{1}{2}-\frac{\nu_j}{\lambda_2}}, \tag{16}$$

where $s_{ij}^{\lambda_1} = \frac{\lambda_1}{(\frac{i}{N}-\frac{j}{M})^2+1}$. We denote $\boldsymbol{K} = [p_{ij}e^{\frac{1}{\lambda_2}(s_{ij}^{\lambda_1}-d_{ij})}]_{ij}$, then $t_{ij} = e^{-\frac{1}{2}-\frac{\mu_i}{\lambda_2}}\boldsymbol{K}_{ij}e^{-\frac{1}{2}-\frac{\nu_j}{\lambda_2}}$, and $\boldsymbol{T}^{\lambda_1,\lambda_2} = \boldsymbol{e}^{diag(-\frac{1}{2}-\frac{\boldsymbol{\mu}}{\lambda_2})}\boldsymbol{K}\boldsymbol{e}^{diag(-\frac{1}{2}-\frac{\boldsymbol{\nu}}{\lambda_2})}$.

All elements of $\boldsymbol{K}$ are strictly positive, since $\boldsymbol{e}^{\boldsymbol{S}}$ is the element-wise exponential of a matrix $\boldsymbol{S}$. According to the Sinkhorn's theorem (Theorem 1), there exist diagonal matrices $diag(\boldsymbol{\kappa}_1)$ and $diag(\boldsymbol{\kappa}_2)$ with strictly positive diagonal elements such that $diag(\boldsymbol{\kappa}_1)\boldsymbol{K}diag(\boldsymbol{\kappa}_2)$ belongs to $U(\boldsymbol{\alpha},\boldsymbol{\beta})$, $\boldsymbol{\kappa}_1 \in \mathbb{R}^N$ and $\boldsymbol{\kappa}_2 \in \mathbb{R}^M$. $diag(\boldsymbol{\kappa}_1)\boldsymbol{K}diag(\boldsymbol{\kappa}_2)$ is unique and the two diagonal matrices are also unique up to a scalar factor.

Theorem 1. *[26, 7, 8]: For any $N \times M$ matrix $\boldsymbol{A}$ with all positive elements, there exist diagonal matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ such that $\boldsymbol{B}_1\boldsymbol{A}\boldsymbol{B}_2$ belongs to $U(\boldsymbol{\alpha},\boldsymbol{\beta})$. $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ have strictly positive diagonal values, and are unique up to a positive scalar factor.*

The optimal $\boldsymbol{T}^{\lambda_1,\lambda_2}$ of Eq. (13) in $U(\boldsymbol{\alpha},\boldsymbol{\beta})$ has the same form with $diag(\boldsymbol{\kappa}_1)\boldsymbol{K}diag(\boldsymbol{\kappa}_2)$, and hence is exactly the unique matrix in $U(\boldsymbol{\alpha},\boldsymbol{\beta})$ that is a rescaled version of $\boldsymbol{K}$. $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ are the unique non-negative left and right scaling vectors up to a scaling factor. Consequently, they can be efficiently obtained by the Sinkhorn-Knopp iterative matrix scaling algorithm:

$$\boldsymbol{\kappa}_1 \leftarrow \boldsymbol{\alpha}./\boldsymbol{K}\boldsymbol{\kappa}_2, \tag{17}$$

$$\boldsymbol{\kappa}_2 \leftarrow \boldsymbol{\beta}./(\boldsymbol{K})^T\boldsymbol{\kappa}_1. \tag{18}$$

We use only 20 iterations in this paper, because a small fixed number of iterations for the Sinkhorn's algorithm works well as reported in [9, 12]. The complexity of calculating the OPW distance is then $O(d'NM)$, where d' is the dimensionality of vectors in sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$. OPW avoids the expensive computation of optimization methods such as interior point method to solve the conventional optimal transport problem.

5. Experiments

5.1. Datasets

MSR Sports Action3D dataset [16]. This dataset consists of 402 depth action sequences from 20 sport actions. Ten subjects perform each action for three times. These sequences contain 23,797 frames in total. Following [33, 34], the dataset is split into training and testing set according to the subjects, where action sequences performed by half of the ten subjects are used for training and the rest sequences are used for testing.

MSR Daily Activity3D dataset [33]. This dataset contains 320 daily activity sequences from 16 activity types. Generally, ten subjects perform each activity in the living room in two poses: "sitting on sofa" and "standing". When "sitting on sofa" or the subject stands close to the sofa, the 3D joint positions obtained by the skeleton tracker are quite noisy. Human-object interactions are involved in most activities. The dataset is split into training and testing set following the experimental setup in [33, 34].

Figure 4. Samples of similar online Chinese characters.

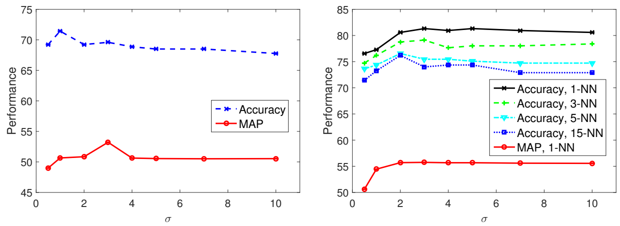
"Spoken Arabic Digits (SAD)" dataset from the UCI Machine Learning Repository [5]. This dataset contains 8,800 vector sequences from ten spoken Arabic digits. The vectors in sequences are the mel-frequency cepstrum coefficients (MFCCs) features extracted from speech signals. Each digit class has 880 sequence samples spoken by 44 male and 44 female Arabic native speakers for ten times. The dataset is split into training and testing sets. 660 samples of each class are used for training and the remaining are used for testing.

Online Chinese character dataset. We select a set of similar online Chinese characters from a collected dataset [29]. The set consists of 12 similar characters sharing the same or similar radical. 107 persons write each character once. Examples and the corresponding GBK codes of these characters are shown in Fig. 4. We divide the samples of each character into five subsets to perform five-fold cross validation.

5.2. Experimental setup

Frame-wide features. For the Action3D dataset and the Activity3D dataset, each action sample is represented by a sequence of frame-wide features. We employ the features provided by the authors of [34] and [33], respectively. The features are the relative positions of all 3D joints in the frame. The dimensionality for the frame-wide features is 192 and 390 for the Action3D dataset and the Activity3D dataset, respectively. In the SAD dataset, each digit sample is represented by a sequence of 13-dimensional MFCC features. For the similar character dataset, we extract 10 features from each stroke of a character sample. The features include the position, shape, orientation, correlation and corner points.

Classification methods and evaluation measures. Two distance-based classifiers, the nearest mean (NM) classifier and the k nearest neighbor (k-NN) classifier, are adopted to perform classification. For NM, the distances between



(a) (b)

Figure 5. Performances with the increase of σ by (a) the NM classifier; (b) the NN classifier.

all the pairwise sequences in each class are calculated. The sequence with the minimum sum of distances with all the other sequences in the same class is determined as the mean of this class. For a test sequence, its distance to the mean sequence of a class is used as the dissimilarity between it and the class. The class with the minimum dissimilarity is determined as the label of the test sequence.

The accuracy and mean average precision (MAP) are used as evaluation measures. For k-NN, the distance between the test sequence and all the training sequences are calculated. The test sequences are classified by a majority vote of its k nearest neighbors. We consider $k = 1, 3, 5, 15$ respectively and the accuracy is used as the evaluation measure. When $k = 1$, the test sequence is also viewed as a query to retrieval the training sequences, and the MAP is calculated.

5.3. Influence of parameters

The proposed OPW distance has three parameters: the standard deviation σ of the prior distribution controlling the expected bandwidth of warping, λ_1 controlling the weight of the inverse difference moment regularization, and λ_2 controlling the balance of the regularization in terms of the KL-divergence with prior distribution. We first evaluate the performances of both the NM and NN classifiers with the increase of σ on the MSR Action3D dataset. λ_1 and λ_2 are fixed to 50 and 0.1, respectively. The results are shown in Fig. 5. We can find that for different evaluation measures and classifiers, the optimal values of σ are different. But generally the accuracies decrease when $\sigma > 5$, because large σ means that transports between distant relative temporal positions are allowed and hence the more temporal information is lost.

We then fix σ to 1 and evaluate the performances of OPW with different λ_1 and λ_2. When changing λ_2 (λ_1), λ_1 (λ_2) is fixed to 50 (0.1). The results are shown in Fig. 6. We can find that OPW is not very sensitive to λ_1 when $\lambda_1 > 1$. Large λ_2 reduces the performances because the prior distribution is used to bound the width of temporal flexibility and the learned transport should not be too close to it. The sudden drops appeared when $\lambda_1 > 500$ and $\lambda_2 \leq 0.01$ are because some entries of $\mathbf{K}$ exceed the machine-precision limit.

Distance	DTW	lDTW	nDTW	Sinkhorn	OPW
Accuracy	71.06	73.63	70.70	66.67	74.36
MAP	50.77	58.05	56.55	51.43	59.10

(a) Results with the NM classifier

Distance	DTW	lDTW	nDTW	Sinkhorn	OPW
MAP	58.95	56.67	56.52	54.58	58.70
1-NN	81.32	82.78	79.85	78.02	84.25
3-NN	81.32	82.05	79.12	77.66	82.78
5-NN	80.95	79.12	76.92	74.73	80.22
15-NN	82.78	75.82	76.19	69.96	77.29

(b) Results with the NN classifier

Table 1. Results on the MSR Sports Action3D dataset. The best results among all distance measures are shown in red, and the second position results are shown in blue.

Method	DTW	lDTW	nDTW	Sinkhorn	OPW
Accuracy	40.63	33.75	38.75	37.50	38.75
MAP	33.25	37.47	31.37	32.14	33.35

(a) Results with the NM classifier

Distance	DTW	lDTW	nDTW	Sinkhorn	OPW
MAP	33.79	28.81	30.56	30.66	34.62
1-NN	58.75	50.00	55.63	54.37	58.13
3-NN	50.62	43.75	47.50	48.13	50.62
5-NN	49.38	50.00	52.50	50.62	53.75
15-NN	43.13	38.75	40.00	41.25	44.37

(b) Results with the NN classifier

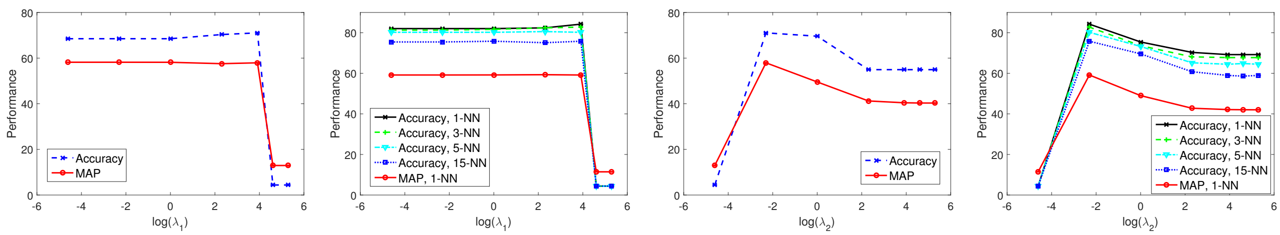
Table 2. Results on the MSR Daily Activity3D dataset.

5.4. Comparison with different distance measures

We compare the performances of the NM and NN classifiers by using different distance measures between sequences. The results on the four datasets are shown in Tab. 1 to Tab. 4, respectively. lDTW and nDTW are variations of DTW normalized by the length of the test sequence and the matching steps, respectively. For the Sinkhorn distance and the proposed OPW distance, the results are reported using the optimal parameters.

We can find that the proposed OPW distance outperforms the most widely adopted DTW distance and its variants, as well as the Sinkhorn distance, on most datasets with different classifiers and evaluation measures. In cases where the performances of OPW are not the best, the best results are obtained by different distances and OPW achieves the second best results that are very close to the best ones.

On three out of four datasets, OPW achieves the highest accuracies, and MAPs among all the classifiers. In partic-



(a) (b) (c) (d)

Figure 6. Performances with the increase of (a) λ_1 by the NM classifier; (b) λ_1 by the NN classifier; (c) λ_2 by the NM classifier; (d) λ_2 by the NN classifier.

Method	DTW	lDTW	nDTW	Sinkhorn	OPW
Accuracy	45.01	46.10	41.31	34.97	57.75
	(2.21)	(3.92)	(3.40)	(1.84)	(2.62)
MAP	40.02	40.50	35.21	28.76	47.40
	(1.75)	(2.76)	(2.07)	(1.12)	(2.64)

(a) Results with the NM classifier

Distance	DTW	lDTW	nDTW	Sinkhorn	OPW
MAP	28.27	27.09	23.87	19.41	32.12
	(0.30)	(0.40)	(0.34)	(0.32)	(1.14)
1-NN	62.44	66.08	60.31	57.54	72.25
	(3.03)	(2.31)	(1.97)	(3.34)	(3.00)
3-NN	59.95	69.78	63.02	57.97	73.62
	(2.40)	(1.84)	(1.89)	(1.62)	(2.01)
5-NN	62.36	69.98	64.39	60.53	76.99
	(2.18)	(2.16)	(2.31)	(2.04)	(2.01)
15-NN	63.77	72.19	64.24	61.29	78.38
	(1.76)	(4.33)	(2.77)	(2.49)	(2.49)

(b) Results with the NN classifier

Table 3. Results on similar online Chinese character recognition. Standard deviations are shown in brackets.

ular, OPW outperforms the sub-optimal results by a margin of 7% on accuracy and MAP with the NM classifier on similar character recognition and 5% on MAPs with the NM and 1-NN classifiers on the SAD dataset. Generally, the Sinkhorn distance leads to worse results than the DTW distance, while the proposed OPW achieves much better results. This means that directly applying the OT-based distances to sequences is not favorable and the temporal regularizations have played an important role in employing the temporal information.

6. Conclusions

In this paper, we have presented order-preserving Wasserstein distance, a new distance measure between sequences. OPW distance adapts the well-known optimal transport for sequence data, where the distribution of instances in one sequence is transported to match the distribution of another sequence with the minimal cost. The learned transport also preserves the temporal orders of instances such that each instance in one sequence should transport to instances in the other sequence with similar relative temporal positions. To this end, two novel regularization terms, *i.e.*, the inverse difference moment regularization and the K-L divergence with prior distribution regularization, are impose to the transport to encourage transports to nearby instances and punish alignments between distant instances. We show that OPW distance can be efficiently calculated by matrix scaling. Experiments on four different datasets demonstrate that OPW distance is able to achieve flexible order-preserving alignments and hence tackle periodic patterns with arbitrary starting points and local temporal reverse or distortion problems.

Method	DTW	lDTW	nDTW	Sinkhorn	OPW
Accuracy	82.55	80.18	76.27	77.00	87.27
MAP	73.23	76.29	67.75	65.11	82.23

(a) Results with the NM classifier

Distance	DTW	lDTW	nDTW	Sinkhorn	OPW
MAP	56.58	56.03	48.47	43.27	62.71
1-NN	96.36	96.73	95.05	87.95	96.68
3-NN	96.91	96.82	95.73	89.05	97.45
5-NN	97.23	96.73	96.09	89.23	97.14
15-NN	97.36	96.50	95.91	90.73	97.41

(b) Results with the NN classifier

Table 4. Results on the SAD dataset.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 61603373. Dr. Gang Hua is partly supported by National Natural Science Foundation of China (NSFC) Grant 61629301.

References

[1] G. Al-Naymat, S. Chawla, and J. Taheri. Sparsedtw: a novel approach to speed up dynamic time warping. In *AusDM*, 2009. 3

[2] F. Albregtsen. Statistical texture measures computed from gray level coocurrence matrices. *Image processing laboratory, department of informatics, university of oslo*, 5, 2008. 4
[3] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *T-PAMI*, 38(1):1–13, 2016. 3
[4] G. Aude, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. 2016. 3
[5] K. Bache and M. Lichman. *UCI Machine Learning Repository*. School of Information and Computer Sciences, University of California, Irvine, 2013. 6
[6] L. Bergroth, H. Hakonen, and T. Raita. A survey of longest common subsequence algorithms. In *International Symposium on String Processing and Information Retrieval*, 2000. 3
[7] A. Borobia and R. Cantó. Matrix scaling: A geometric proof of sinkhorn's theorem. *Linear algebra and its applications*, 268:1–8, 1998. 6
[8] R. A. Brualdi, S. V. Parter, and H. Schneider. The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *Journal of Mathematical Analysis and Applications*, 16(1):31–50, 1966. 6
[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013. 2, 3, 4, 6
[10] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *ICML*, 2014. 3
[11] R. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *TPAMI*, 26(6):732–739, 2004. 2
[12] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein discriminant analysis. *arXiv preprint arXiv:1608.08063*, 2016. 6
[13] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. In *NIPS*, 2015. 3
[14] D. Garreau, R. Lajugie, S. Arlot, and F. Bach. Metric learning for temporal sequence alignment. In *NIPS*, 2014. 3
[15] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 2
[16] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR workshop on Human Communicative Behavior Analysis*, 2010. 6
[17] P.-F. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *TPAMI*, 31(2):306–318, 2009. 3
[18] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys*, 33(1):31–88, 2001. 3
[19] C. A. Ratanamahatana and E. Keogh. Making Time-Series Classification More Accurate Using Learned Constraints. In *SDM*, 2004. 3
[20] J. Rodriguez-Serrano and F. Perronnin. A Model-Based Sequence Similarity with Application to Handwritten Word-Spotting. *TPAMI*, 34(11):2108–2120, 2012. 3
[21] Y. Rubber, L. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668, 1997. 3
[22] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *TASSP*, 26(1):43–49, 1978. 2
[23] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007. 3
[24] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *NIPS*, 2015. 3
[25] D. F. Silva and G. E. Batista. Speeding up all-pairwise dynamic time warping matrix calculation. In *SDM*, 2016. 3
[26] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. 3, 6
[27] B. Su and X. Ding. Linear sequence discriminant analysis: a model-based dimensionality reduction method for vector sequences. In *ICCV*, 2013. 3
[28] B. Su, X. Ding, C. Liu, and Y. Wu. Heteroscedastic max-min distance analysis. In *CVPR*, 2015. 2
[29] B. Su, X. Ding, H. Wang, and Y. Wu. Discriminative dimensionality reduction for multi-dimensional sequences. *TPAMI*, 2017. 3, 6
[30] J. Su, S. Kurtek, E. Klassen, A. Srivastava, et al. Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552, 2014. 3
[31] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 2, 3
[32] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh. Indexing Multidimensional Time-Series. *VLDB*, 15(1):1–20, 2006. 3
[33] J. Wang, Z. Liu, and Y. Wu. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012. 6
[34] J. Wang and Y. Wu. Learning maximum margin temporal warping for action recognition. In *ICCV*, 2013. 3, 6
[35] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *CVPR*, 2012. 3
[36] F. Zhou and F. Torre. Canonical time warping for alignment of human behavior. In *NIPS*, 2009. 3