Order-preserving Wasserstein Discriminant Analysis

Bing Su[1], Jiahuan Zhou[2], Ying Wu[2]
[1]Science & Technology on Integrated Information System Laboratory,
Institute of Software, Chinese Academy of Sciences, Beijing, 100190, China
[2]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, 60208
subingats@gmail.com, {jzt011, yingwu}@eecs.northwestern.edu

Abstract

Supervised dimensionality reduction for sequence data projects the observations in sequences onto a low-dimensional subspace to better separate different sequence classes. It is typically more challenging than conventional dimensionality reduction for static data, because measuring the separability of sequences involves non-linear procedures to manipulate the temporal structures. This paper presents a linear method, namely Order-preserving Wasserstein Discriminant Analysis (OWDA), which learns the projection by maximizing the inter-class distance and minimizing the intra-class scatter. For each class, OWDA extracts the order-preserving Wasserstein barycenter and constructs the intra-class scatter as the dispersion of the training sequences around the barycenter. The inter-class distance is measured as the order-preserving Wasserstein distance between the corresponding barycenters. OWDA is able to concentrate on the distinctive differences among classes by lifting the geometric relations with temporal constraints. Experiments show that OWDA achieves competitive results on three 3D action recognition datasets.

1. Introduction

The sequence classification problem arises in a wide range of real-world applications. A sequence is comprised of a series of ordered observations, where each individual observation is generally of no special interest, but the sequence as a whole represents the target object. The observations in the same sequence are not independent and their relationship reveals the temporal structure of the sequence.

The similarity between sequences, which plays a crucial role in classification, should take not only all the pairwise vector-level distances between observations but also such temporal dependencies into consideration. In most similarity measures for sequences, the temporal dependencies and alignments need to be inferred from the matrix of pairwise distances between observations under the temporal constraints. The complexity of constructing the pairwise distance matrix highly depends on the dimensionality of observations in sequences. Lower-dimensional representations have a significant effect on reducing the running time of calculating the similarity and building the subsequent models or classifiers. Discriminative representations that lead to small similarities for different patterns and large similarities for the sequences from the same class generally improve the classification performance.

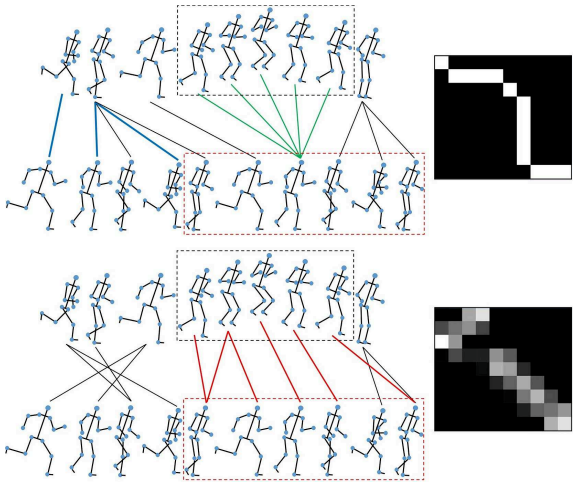


Figure 1. Top: the DTW [26] alignment. The alignment matrix is shown on the right. The white grid in row i and column j indicates that the i-th and j-th observations in the two sequences are aligned. Bottom: the OPW [34] alignment. The transport matrix is shown on the right. The grey value of a grid indicates the probability of aligning the corresponding observations.

Supervised dimensionality reduction for sequence data (DRS) attempts to learn such low-dimensional discriminative representations for observations in sequences by transforming the observations in the noisy high-dimensional space to a subspace. Generally, the transformation is learned by measuring the similarity or separability among sequences from different classes. However, unlike vector data, the representations of observations not directly act on the similarity between sequences but go through a nonlin-

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00998

ear warping or alignment inference. This makes it difficult to build the separability among sequence classes, formulate the discriminant objective, and develop efficient solutions.

Most existing DRS methods [29, 33, 30] employ dynamic time warping (DTW) [26] to perform the alignment. Due to the boundary condition and the strict order-preserving constraint, DTW cannot tackle local reorder distortions and may not fully capture the essential differences of different patterns. As shown in Fig. 1, the two action sequences "jump" and "run" differ in the boxed parts, where "jump" vacates after a run-up. In the beginning, the two sequences start from different poses, one takes the right leg first and the other takes the left leg first, resulting in reordered poses. Some different running poses are wrongly aligned by DTW (shown in blue bold). The vacated poses of "jump" are forced to align to a single pose of "run" (shown in green).

In this paper, we propose a linear supervised DRS method by employing the order-preserving Wasserstein (OPW) distance [34, 35] as the similarity measure between sequences. For each class, we extract the order-preserving Wasserstein barycenter and measure the dispersion of training sequences around the barycenter w.r.t. the OPW distance. We measure the inter-class separability between two classes as the OPW distance between the corresponding barycenters. In this way, the intra- and inter-class separabilities are uniformly measured with OPW. We learn the transformation by maximizing the overall separability.

OPW casts the temporal alignment as a transport problem. It encourages transport between temporally adjacent observations, but allows local reorders or distortions. In Fig. 1, the reordered running poses are correctly aligned by OPW. For the boxed parts, the vacated poses of "jump" are dispersedly aligned to different poses in a periodic cycle of "run" (shown in red). OPW is able to determine the true distinctive observation pairs that reflect the essential differences of two sequences, so that the DRS method can focus on discriminating these distinctions. In addition, different from the binary DTW alignment, the transport measures the probabilities of how different observation pairs contribute to the total difference. The probabilities among the boxed parts are scattered and more local relations among all observations are considered by the proposed DRS method.

The main contributions of this paper are three-fold. 1. We propose novel OPW-based separability measures among sequence classes which reflect their essential differences. 2. We provide mathematical derivations to compute the barycenter. 3. We construct new intra-class and inter-class scatters based on the learned optimal transports to employ more local pairwise differences.

2. Related Work

Supervised linear dimensionality reduction for static data has been extensively studied in the literature. The well-known linear discriminant analysis (LDA) learns the projection by maximizing the ratio of inter-class distance to the intra-class distance. Various methods are proposed to improve or extend LDA in specific situations. The null space LDA [7], generalized ULDA [44] and orthogonal LDA [43] deal with the small sample size problem. Heteroscedastic LDA [21] and subclass discriminant analysis [47] handle heteroscedastic data. Max-min distance analysis approaches [4, 46, 31, 32] tackle the class separation problem. Marginal Fisher analysis [41] only uses the neighboring samples and the samples distributed around the class boundaries to construct the intra-class and inter-class scatters. Wasserstein discriminant analysis [14] employs the regularized Wasserstein distance to measure the distance between the empirical probabilities of class populations.

These advances cannot be applied to observations in sequences directly because the observations do not satisfy the basic i.i.d. assumption. Far less attention has been paid to DRS. In [28], a kernel-based sufficient dimensionality reduction approach is proposed to improve the performance of sequence labeling, where each observation in sequences has a label. In this paper, we learn the projection to improve the performance of sequence classification that each entire sequence is associated with a single label. In [18], a Mahalanobis distance for observations in sequences is learned to improve the performance of multivariate sequence alignment, where the ground-truth alignments between sequences are given. In this paper, we learn the projection without any alignment annotations. In [23], the embedding vectors of tree nodes are learned by minimizing a surrogate of the classification error using the nearest prototype classifier w.r.t. the tree edit distance, where the prototypes are selected from the training trees. In this paper, we minimize the distances between training sequences to the corresponding barycenters w.r.t. the OPW distance.

In [29, 33, 30], linear sequence discriminant analysis (LSDA) and max-min inter-sequence distance analysis (MMSDA) are proposed for DRS, respectively. LSDA and MMSDA extract a representative sequence and a intra-class variance matrix for each class based on the statistics of a trained HMM. The DTW distance between the representative sequences is used as the inter-class distance. The similarities for measuring the inter-class distance and intra-class scatter are inconsistent, because the HMM-based intra-class variance does not measure the dispersion of the DTW distances among the sequences. In this paper, we employ the OPW distance instead of the DTW distance as the similarity measure between sequences, and construct the intra-class scatter and the inter-class distance consistently w.r.t. the OPW distance. We extract the order-preserving Wasserstein barycenter as the representative sequence, which is non-parametric and has better scalability without the need of training HMMs with massive parameters. MMSDA opti-

mizes the max-min distance criterion, which is more suited to tackle the class separation problem. Note that the proposed method can also be extended by applying the max-min distance criterion to the constructed inter- and intra-class scatters. In this paper, we only compare with the DRS methods optimizing the same Fisher criterion.

3. Our Proposed Method

3.1. Background on OPW

We first briefly review the *order-preserving Wasserstein (OPW) distance* [34, 35]. For two sequences $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_x}]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{N_y}]$ with lengths N_x and N_y, respectively, where the dimension of features is q, i.e., $\boldsymbol{x}_i, \boldsymbol{y}_j \in \mathbb{R}^q$, the OPW distance is defined as:

$$\begin{aligned} & d_{OPW}(\boldsymbol{X}, \boldsymbol{Y}) := \langle \boldsymbol{T}^*, \boldsymbol{D} \rangle \\ s.t.\ & \boldsymbol{T}^* = \underset{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})}{arg\min} \langle \boldsymbol{T}, \boldsymbol{D} \rangle - \lambda_1 I(\boldsymbol{T}) + \lambda_2 KL(\boldsymbol{T}||\boldsymbol{P}) \end{aligned}, \tag{1}$$

where $\boldsymbol{D} := [d(\boldsymbol{x}_i, \boldsymbol{y}_j)]_{ij} \in \mathbb{R}^{N_x \times N_y}$ is the matrix of all the pairwise distances between supporting points, $d(\cdot, \cdot)$ is set to the squared Euclidean distance in this paper. $\boldsymbol{T} := [t_{ij}]_{ij} \in \mathbb{R}^{N_x \times N_y}$ is the transport matrix, $\langle \cdot, \cdot \rangle$ is the Frobenius dot product, and $U(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \{\boldsymbol{T} \in \mathbb{R}_+^{N_x \times N_y} | \boldsymbol{T}\mathbf{1}_{N_y} = \boldsymbol{\alpha}, \boldsymbol{T}^T\mathbf{1}_{N_x} = \boldsymbol{\beta}\}$ is the feasible set of the transport $\boldsymbol{T}$. $I(\boldsymbol{T}) = \sum_{i,j} \frac{t_{ij}}{(\frac{i}{N_x} - \frac{j}{N_y})^2 + 1}$ is the inverse difference moment of the transport matrix $\boldsymbol{T}$ to encourage the local homogeneity that large values appear near the diagonal, and $KL(\boldsymbol{T}||\boldsymbol{P})$ is the Kullback-Leibler divergence between $\boldsymbol{T}$ and a prior distribution $\boldsymbol{P}$. $\boldsymbol{P}$ is a two-dimensional distribution whose values decrease gradually from the diagonal to both sides following a Gaussian distribution in any transverse sections. $\lambda_1 > 0$ and $\lambda_2 > 0$ are two hyper-parameters. It is assumed that the weights of instances in the same sequence are the same, *i.e.*, $\boldsymbol{\alpha} = (\frac{1}{N_x}, \cdots, \frac{1}{N_x})$ and $\boldsymbol{\beta} = (\frac{1}{N_y}, \cdots, \frac{1}{N_y})$, respectively. In [34], OPW is solved by the Sinkhorn's fixed point algorithm with a complexity of $N_x N_y q$.

3.2. Order-preserving Wasserstein barycenter

For a sequence class with a set of training sequences, we want to extract a single representative sequence that reveals the average temporal structures and general evolution trends, which can serve as the mean sequence of a set of sequences similar to the mean vector of a set of vectors. Extending the averaging operation to sequences is challenging. As the lengths of different sequences are different, it is not plausible to perform directly averaging to the observations at the same time step.

Recall that the mean of a set of vectors can also be viewed as the barycenter of the vectors with regard to the Euclidean distance. Similarly, for sequence data, the barycenter of a set of sequences with regard to a sort of sequence distance can also act as the mean sequence in some sense. We extract the barycenter with regard to the OPW distance, which we call the *order-preserving Wasserstein barycenter*.

The barycenter $\boldsymbol{U} = (\boldsymbol{\mu}, \boldsymbol{\gamma})$ consists of a sequence of ordered supporting points and a weight sequence associating each supporting point with a probability value. $\boldsymbol{\mu} = [\boldsymbol{\mu}_i, i = 1, \cdots, L]$ is the sequence of supporting points and $\boldsymbol{\gamma} = [\gamma_i, i = 1, \cdots, L]$ is the sequence of associated weights. $\boldsymbol{\gamma}$ lies in the simplex Θ_L. L is a pre-set value, which indicates the maximum allowed number of supporting points of the barycenter.

Given a set of sequences $\boldsymbol{X}_k, k = 1, \cdots, N$, let $\boldsymbol{D}_k$ denote the matrix of all pairwise ground distances between any $\boldsymbol{\mu}_i$ and observations in $\boldsymbol{X}_k$, and $\boldsymbol{T}_k$ denote the transport between $\boldsymbol{U}$ and $\boldsymbol{X}_k$. The optimal transport determined by OPW is given by $arg\, W(\boldsymbol{U}, \boldsymbol{X}_k)$, where

$$W(\boldsymbol{U}, \boldsymbol{X}_k) = \min_{\boldsymbol{T}_k \in U(\boldsymbol{\gamma}, \boldsymbol{\beta}_k)} \langle \boldsymbol{T}_k, \boldsymbol{D}_k \rangle - \lambda_1 I(\boldsymbol{T}_k) + \lambda_2 KL(\boldsymbol{T}_k||\boldsymbol{P}). \tag{2}$$

By assuming that these sequences are equally weighted, the order-preserving Wasserstein barycenter is such that

$$\boldsymbol{U} = \underset{\boldsymbol{U}}{arg\min} \sum_{k=1}^{N} \frac{1}{N} W(\boldsymbol{U}, \boldsymbol{X}_k). \tag{3}$$

Both the supporting points and their weights need to be learned. However, the objective function (3) is not convex w.r.t. them simultaneously. We employ the alternating updating strategy to minimize (3). We first update the weight sequence $\boldsymbol{\gamma}$ and the optimal transports $\boldsymbol{T}_k, k = 1, \cdots, N$ by fixing the supporting points. We reformulate the objective of OPW for optimizing $\boldsymbol{T}_k$ as follows, where the deduction is presented in the supplementary material.

$$\langle \boldsymbol{T}_k, \boldsymbol{D}_k \rangle - \lambda_1 I(\boldsymbol{T}_k) + \lambda_2 KL(\boldsymbol{T}_k||\boldsymbol{P}) = \lambda_2 KL(\boldsymbol{T}_k||\boldsymbol{K}_k), \tag{4}$$

where $d_{ij}^k = d_k(\boldsymbol{\mu}_i, \boldsymbol{x}_j^k)$, $s_{ij}^{\lambda_1} = \frac{\lambda_1}{(\frac{i}{N} - \frac{j}{M})^2 + 1}$, and $\boldsymbol{K}_k = [p_{ij} e^{\frac{1}{\lambda_2}(s_{ij}^{\lambda_1} - d_{ij}^k)}]_{ij}$.

$\boldsymbol{D}_k, k = 1, \cdots, N$ are fixed since $\boldsymbol{\mu}$ is fixed, hence $\boldsymbol{K}_k$ are also fixed. Problem (3) is thereby reformulated as

$$\begin{aligned} & \min_{\boldsymbol{\gamma}, \boldsymbol{T}_k, k=1,\cdots,N} \sum_{k=1}^{N} \tfrac{1}{N} KL(\boldsymbol{T}_k||\boldsymbol{K}_k) \\ & s.t.\ \exists \boldsymbol{\gamma} \in \Theta_L, \boldsymbol{T}_k \mathbf{1}_{N_k} = \boldsymbol{\gamma}, \forall k = 1, \cdots, N \\ & \quad \boldsymbol{T}_k{}^T \mathbf{1}_L = [\tfrac{1}{N_k}, \cdots, \tfrac{1}{N_k}]^T, k = 1, \cdots, N \end{aligned}. \tag{5}$$

By defining $\boldsymbol{T} = (\boldsymbol{T}_k)_{k=1}^N \in (\mathbb{R}_+^{L \times N_k})^N$ and $\boldsymbol{K} = (\boldsymbol{K}_k)_{k=1}^N \in (\mathbb{R}_+^{L \times N_k})^N$, Problem (5) is rewritten as

$$\begin{aligned} & \min_{\boldsymbol{\gamma}, \boldsymbol{T}} KL_N(\boldsymbol{T}||\boldsymbol{K}), \boldsymbol{\gamma} \in \Theta_L \\ & s.t.\ \boldsymbol{T} \in \boldsymbol{\Phi}_1 \cap \boldsymbol{\Phi}_2 \end{aligned}, \tag{6}$$

where $KL_N(\boldsymbol{T}||\boldsymbol{K}) := \sum_{k=1}^{N} \frac{1}{N} KL(\boldsymbol{T}_k||\boldsymbol{K}_k)$,

$$\boldsymbol{\Phi}_1 := \left\{ \boldsymbol{T} \in (\mathbb{R}_+^{L\times N_k})^N : {\boldsymbol{T}_k}^T \mathbf{1}_L = [\tfrac{1}{N_k}, \cdots, \tfrac{1}{N_k}]^T, \forall k \right\},$$

$$\boldsymbol{\Phi}_2 := \left\{ \boldsymbol{T} \in (\mathbb{R}_+^{L\times N_k})^N : \exists \boldsymbol{\gamma} \in \Theta_L, \boldsymbol{T}_k \mathbf{1}_{N_k} = \boldsymbol{\gamma}, \forall k \right\}.$$

In [3], it is shown that the iterative Bregman projection [5, 2] can solve Problem (6) efficiently. Specifically, as proved in [34], each $\boldsymbol{T}_k$ is a rescaled version of $\boldsymbol{K}_k$ with the form of $diag(\boldsymbol{\kappa}_{k1})\boldsymbol{K}_k diag(\boldsymbol{\kappa}_{k2})$, and the scaling vectors can be updated using the Sinkhorn's iterations:

$$\boldsymbol{\kappa}_{k1}^{(n)} \leftarrow \boldsymbol{\gamma}^{(n)} ./ \boldsymbol{K}_k \boldsymbol{\kappa}_{k2}^{(n)}, \tag{7}$$

$$\boldsymbol{\kappa}_{k2}^{(n+1)} \leftarrow [\tfrac{1}{N_k}, \cdots, \tfrac{1}{N_k}]^T ./ (\boldsymbol{K}_k)^T \boldsymbol{\kappa}_{k1}^{(n)}. \tag{8}$$

As given in [3], $\boldsymbol{\gamma}^{(n)}$ is the update of the weights:

$$\boldsymbol{\gamma}^{(n)} \leftarrow \prod_{k=1}^{N} \left(\boldsymbol{\kappa}_{k1}^{(n)} \odot ((\boldsymbol{K}_k)^T \boldsymbol{\kappa}_{k2}^{(n)}) \right)^{\frac{1}{N}}. \tag{9}$$

where $\odot$ is the element-wise product. The iterations continue until convergence. Given the learned weights and the fixed supporting points, we perform OPW to obtain the updates of the optimal transports $\boldsymbol{T}_k$, for $k = 1, \cdots, N$.

Then, we update the supporting point sequence $\boldsymbol{\mu}$ with fixed weight sequence $\boldsymbol{\gamma}$ and optimal transports $\boldsymbol{T}_k^*, k = 1, \cdots, N$. In Eq. (2), only the first term evolves $\boldsymbol{\mu}$. By viewing the sequences $\boldsymbol{\mu}$ and $\boldsymbol{X}_k$ as matrices, we have

$$\begin{aligned} \langle \boldsymbol{T}_k^*, \boldsymbol{D}_k \rangle = & \, diag(\boldsymbol{\mu}^T \boldsymbol{\mu})^T \boldsymbol{\gamma} - 2 \left\langle \boldsymbol{T}_k^*, \boldsymbol{\mu}^T \boldsymbol{X}_k \right\rangle \\ & + diag(\boldsymbol{X}_k^T \boldsymbol{X}_k)^T [\tfrac{1}{N_k}, \cdots, \tfrac{1}{N_k}]^T \end{aligned}.$$

We follow [9] to optimize the local quadratic approximation of the following function: $diag(\boldsymbol{\mu}^T\boldsymbol{\mu})^T\boldsymbol{\gamma} - 2\left\langle \boldsymbol{T}_k^*, \boldsymbol{\mu}^T\boldsymbol{X}_k \right\rangle = \left\| \boldsymbol{\mu} diag(\boldsymbol{\gamma}^{\frac{1}{2}}) - \boldsymbol{X}_k {\boldsymbol{T}_k^*}^T diag(\boldsymbol{\gamma}^{-\frac{1}{2}}) \right\|^2 - \left\| \boldsymbol{X}_k {\boldsymbol{T}_k^*}^T diag(\boldsymbol{\gamma}^{-\frac{1}{2}}) \right\|^2$. This leads to the Newton update:

$$\boldsymbol{\mu} \leftarrow \boldsymbol{X}_k {\boldsymbol{T}_k^*}^T diag(\boldsymbol{\gamma}^{-1}). \tag{10}$$

For all N training sequences, $\boldsymbol{\mu}$ is updated by

$$\boldsymbol{\mu} \leftarrow (1-\xi)\boldsymbol{\mu} + \xi(\sum_{k=1}^{N} \boldsymbol{X}_k {\boldsymbol{T}_k^*}^T) diag(\boldsymbol{\gamma}^{-1}), \tag{11}$$

where $\xi \in [0, 1]$ is a pre-set value.

We cycle the two alternative procedures until convergence or a maximum number of steps is reached. It was shown in [2, 3] that the iterative Bregman projection for updating $\boldsymbol{\gamma}$ converges linearly. The convergence rate of the Newton's method for updating $\boldsymbol{\mu}$ is quadratic. It can be difficult to obtain the global convergence rate of the overall alternating optimization. In our experiments, it converges in about 10 iterations. The complexity per iteration is $O(NTLq)$, where T is the average length of sequences.

3.3. Covariance

For a set of sequences, the barycenter reflects the average evolution. The dispersion of the sequences around the barycenter can be straightforwardly measured by accumulating the OPW distances:

$$d_w = \sum_{k=1}^{N} d_{OPW}(\boldsymbol{U}, \boldsymbol{X}_k) = \sum_{k=1}^{N} \langle \boldsymbol{T}_k^*, \boldsymbol{D}_k \rangle,$$

where the optimal transports $\boldsymbol{T}_k^*$ between $\boldsymbol{U}$ and $\boldsymbol{X}_k$, for $k = 1, \cdots, N$, are the by-products when determining the barycenter, so no extra calculations are needed.

To measure the covariance over different dimensions, we define a covariance matrix $\boldsymbol{\Gamma}$ so that $tr(\boldsymbol{\Gamma}) = d_w$. $\boldsymbol{\Gamma}$ can be constructed by accumulating the weighted outer products between any $\boldsymbol{\mu}_i$ and observations in $\boldsymbol{X}_k$ as follows:

$$\boldsymbol{\Gamma} = \sum_{k=1}^{N} \sum_{i=1}^{L} \sum_{j=1}^{N_k} {t_{ij}^k}^* (\boldsymbol{\mu}_i - \boldsymbol{x}_j^k)(\boldsymbol{\mu}_i - \boldsymbol{x}_j^k)^T. \tag{12}$$

We can find that $\boldsymbol{\Gamma}$ captures all local relations between elements of the barycenter and the observations in all sequences. All element-observation pairs contribute to the total covariance with different weights. The weight of a pair $(\boldsymbol{\mu}_i, \boldsymbol{x}_j^k)$ is actually the corresponding element ${t_{ij}^k}^*$ of the learned transport $\boldsymbol{T}_k^*$, so it reflects the probability of matching the pair. In this way, the local pairwise relations or joint probabilities are encoded. The weights are larger for the pairs that have high joint probabilities, since the matched pairs probably correspond to the same temporal structure. The differences between pairs with low joint probabilities are also incorporated, but with smaller weights, to consider soft alignments and compensate possible missing matches. As a result, the constructed $\boldsymbol{\Gamma}$ better reflects the spatial-temporal variances in different dimensions.

3.4. Learning the projection

Our goal is to learn a transformation that projects the observations in sequences onto a low-dimensional subspace, in which the sequences from different classes get better separated. We employ the Fisher criterion to maximize the separability, *i.e.*, we maximize the ratio of the inter-sequence-class distance to the intra-sequence-class dispersion.

For each sequence class $\omega_c, c = 1, \cdots, C$, we extract the order-preserving Wasserstein barycenter $\boldsymbol{U}^c$ and the covariance matrix $\boldsymbol{\Gamma}^c$ from the training sequences of the class. C is the total number of classes. We define the intra-sequence-class scatter as the weighted sum of covariances:

$$\boldsymbol{\Gamma}_w = \sum_{c=1}^{C} p^c \boldsymbol{\Gamma}^c, \tag{13}$$

where p^c is the estimated prior probability of class ω_c.

We measure the distance between two classes ω_c and $\omega_{c'}$ by the OPW distance between the corresponding order-preserving Wasserstein barycenters.

$$d_b(\omega_c, \omega_{c'}) = d_{OPW}(\boldsymbol{U}^c, \boldsymbol{U}^{c'}) = \langle \boldsymbol{T}^*_{cc'}, \boldsymbol{D}_{cc'} \rangle, \quad (14)$$

where $\boldsymbol{D}_{cc'}$ is the matrix of all pairwise distance between $\boldsymbol{\mu}_i^c$ and $\boldsymbol{\mu}_j^{c'}$, and $\boldsymbol{T}^*_{cc'}$ is the optimal OPW transport between the two barycenters. The corresponding between-class scatter $\boldsymbol{\Gamma}_{b(cc')}$ is the weighted sum of outer products between elements of the two barycenters, so that $d_b(\omega_c, \omega_{c'}) = tr(\boldsymbol{\Gamma}_{b(cc')})$:

$$\boldsymbol{\Gamma}_{b(cc')} = \sum_{i=1}^{L} \sum_{j=1}^{L} t_{ij}^{cc'*} (\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_j^{c'})(\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_j^{c'})^T. \quad (15)$$

We define the overall inter-sequence-class scatter as the weighted sum of all pairwise between-class scatters:

$$\boldsymbol{\Gamma}_b = \sum_{c=1}^{C-1} \sum_{c'=c+1}^{C} p^c p^{c'} \boldsymbol{\Gamma}_{b(cc')}. \quad (16)$$

We can observe again that all the differences between elements in all barycenters contribute to the overall inter-class scatter according to different weights. The weight $t_{ij}^{cc'*}$ of a pair $(\boldsymbol{\mu}_i^c, \boldsymbol{\mu}_j^{c'})$ encodes the local relations of the two elements and indicates their joint probability. $\boldsymbol{\Gamma}_b$ concentrates more on the differences between the pairs with large joint probabilities. Such differences reflect the essential distinctions of two classes, because the matched pairs represent the homologous temporal structures and thus are distinctive for discriminating the two classes. Different from the alignments by DTW, where the weights are 1 for a small portion of aligned pairs and 0 for other pairs, the weights by OPW are soft probability values and hence $\boldsymbol{\Gamma}_b$ also incorporates the differences between the pairs with smaller weights. This compromises more information and is more robust to incorrect or ambiguous alignments caused by noises.

When both the features in sequences and their dimensions are not linearly related, the ranks of $\boldsymbol{\Gamma}_w$ and $\boldsymbol{\Gamma}_b$ are $min(N^t, q)$ and $min(CL, q)$, respectively, where N^t is the number of all features in all training sequences. When $N^t \geq q$ ($CL \geq q$), $\boldsymbol{\Gamma}_w$ ($\boldsymbol{\Gamma}_b$) is full-rank. In extreme cases when there are too few training sequences so that $N^t < q$, we can use PCA to remove the null space of $\boldsymbol{\Gamma}_w$ or add a identity matrix multiplied by a small scalar to $\boldsymbol{\Gamma}_w$.

The objective of learning the projection $\boldsymbol{W}$ using the Fisher criterion is formulated as the ratio-trace problem:

$$\max_{\boldsymbol{W}} tr((\boldsymbol{W}^T \boldsymbol{\Gamma}_w \boldsymbol{W})^{-1} \boldsymbol{W}^T \boldsymbol{\Gamma}_b \boldsymbol{W}). \quad (17)$$

The optimal $\mathbf{W}^*$ of Problem (17) is the matrix whose columns are the eigenvectors of $\boldsymbol{\Gamma}_w^{-1} \boldsymbol{\Gamma}_b$ w.r.t. the d' largest eigenvalues, where d' is the reduced dimensionality. The proposed DRS method is called *Order-preserving Wasserstein Discriminant Analysis (OWDA)*.

3.5. Complexity

Let N_a and T denote the average number of sequences per class and the average length of sequences, respectively. The complexities for calculating the barycenters for all C classes, calculating the inter-class and intra-class scatters, and solving (17) are $O(CN_aTLq)$, $O(C^2L^2q^2)$, $O(CN_aLTq^2)$, and $O(q^3)$, respectively. The overall complexity is $O(C^2L^2q^2) + O(CN_aLTq^2) + O(q^3)$. It scales linearly with the number of samples, but cubically with the dimension of features q due to the eigen-decomposition (17). We simultaneously diagonalize the intra-class and inter-class scatters [43] to solve (17). Any advanced methods for large-scale eigen-decomposition can be applied to accelerate our method.

4. Experiments

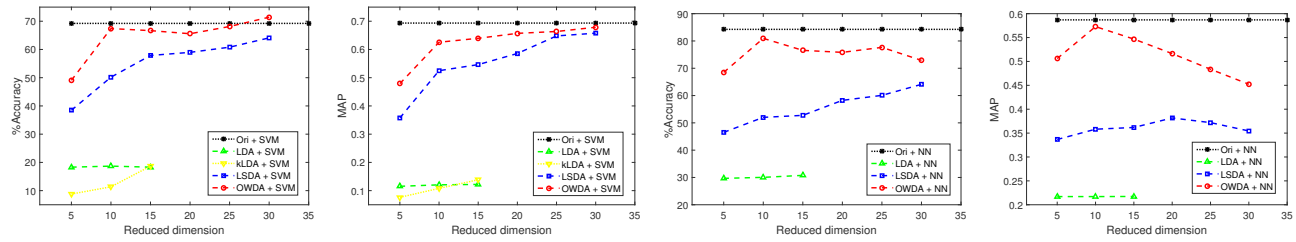
We evaluate the proposed OWDA method on three 3D-action datasets. Evaluations on the influence of hyper-parameters are presented in the supplementary file.

4.1. Results on the Action3D dataset

The **MSR Sports Action3D dataset [19, 37]** contains 557 depth sequences captured by Kinect camera from 20 sports actions. Ten persons performed each action for two or three times. The skeleton joint positions of humans are also available in this dataset.

We extract a feature vector from each frame as the observation of the frame. In this way, we represent each video by a sequence of observations. We employ the pairwise-joint-angle-based features provided by the authors of [37] as the frame-wise observation, which are the relative angles of all the 3D joints w.r.t. other joints. The dimensionality of the observation is 192. In [37, 38], the authors split the dataset into a training set and a test set, where the training set includes the sequences performed by about half of the persons and the test set includes the rest. We follow this experimental setup and report our results on the test set.

We employ the proposed OWDA method to project the observations in sequences onto subspaces with different dimensions. In the learned subspaces, we employ two sequence classifiers to classify the transformed sequences: the SVM classifier and the nearest neighbor (NN) classifier. For the SVM classifier, we first encode each sequence of observations into a fixed-dimensional vector by the unsupervised rank pooling [12]. Rank pooling learns two linear functions to rank the forward and reverse timing orders of the observations by the support vector regression, respectively. The parameters of the two linear functions are concatenated to form the pooling vector. Then, we train linear SVMs by taking these resulting vectors as input. We determine the hyper-parameter C of the linear SVMs by cross-validation. At the testing phase, we encode the test sequence of obser-



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

Figure 2. (a) Accuracies with the SVM classifier (b) MAPs with the SVM classifier (c) Accuracies with the NN classifier and (d) MAPs with the NN classifier as functions of the dimensionality of the subspace on the MSR Action3D dataset.

vations into a vector by rank pooling, and then employ the leaned SVMs to classify the encoded vector.

For the NN classifier, we employ the OPW distance as the dissimilarity measure between two sequences. Specifically, for a test sequence, we calculate its OPW distance to all training sequences. We predict its class label as the label of the training sequence which has the smallest OPW distance with it among all training sequences.

We compare the proposed OWDA with other dimensionality reduction methods for sequences. OWDA employs the Fisher criterion. As discussed in Section 2, different criteria are generally suited for different cases. In addition, OWDA can also be extended by employing other criteria. Therefore, to obtain a fair comparison, we only compare with those methods based on Fisher criterion, including LDA, kernel LDA (k-LDA), and LSDA. For LSDA, we use the same hyper-parameters as in [29, 33]. For OWDA, the hyper-parameter L is fixed to 8 in all our experiments.

LDA and kLDA are based on the *i.i.d.* assumption. To apply them to sequence data, we view the observations in sequences as independent samples with the same class label. Generally, a single sequence contains a number of observations, using all observations in all sequences as training samples results in a large-scale kernel matrix. It is impracticable to perform kLDA with such kernel due to the huge space and computational overhead. Therefore, we only sample less than 5 observations per sequence for training. We employ the drtoolbox [36] to implement LDA and kLDA. In addition, we also evaluate the performances using both classifiers in the original space.

We adopt the accuracy and MAP (mean average precision) as performance measures. For the SVM classifier, we train a multi-class SVM to evaluate the classification accuracy. We train a binary SVM for each class and use the scores to rank all training encoded vectors to evaluate the MAP. Additional evaluations by using the multi-class precision and recall as performance measures with this classifier are presented in the supplementary file. For the NN classifier, to evaluate the MAP, we view each test sequence as a query to rank all training sequences with the OPW distance.

The results of different DRS methods with different reduced dimensions are shown in Fig. 2. We can observe that the proposed OWDA outperforms other DRS methods by a significant margin with both classifiers. Especially when a few dimensions (*e.g.*, less than 15) are preserved, OWDA outperforms the second LSDA by a margin of about 10% for both accuracy and MAP. Compared with the original 192-dimensional observations, OWDA achieves better accuracy and comparable MAP using only 30 dimensions for the SVM classifier, and achieves comparable accuracy and MAP using only 10 dimensions for the NN classifier.

4.2. Results on the Activity3D dataset

The **MSR Daily Activity3D dataset [37]** contains 320 daily activity sequences from 16 activity classes. The sequences were captured by a Kinect device. Ten subjects performed each activity in two poses.

On this dataset, we employ the pairwise-joint-position-based features provided by the authors of [37, 38] as the frame-wise observations, whose dimensionality is 390. We follow the split of the dataset as in [37, 38] again and report our results on the test set. We compare OWDA to LDA and LSDA. Other experimental settings remain the same as those on the MSR Action3D dataset.

Fig. 3 depicts the performances of different DRS methods as functions of the reduced dimension by both classifiers. For the SVM classifier, classifying the original sequences directly without any DRS methods performs best, but the proposed OWDA performs better than other DRS methods. Especially, the proposed OWDA using only 25 dimensions achieves comparable MAP with the original sequences. Since the activities in this dataset show larger variations than the actions in the Action3D dataset, the sequences in the same class may be spread in different clusters. *E.g.*, different persons may perform the same action "call cellphone" using different hands or poses. A single barycenter for each class is unable to distinguish such situations. Therefore, OWDA performs inferior to the original features. Adding the number of barycenters per class may further increase the performances of OWDA.

For the NN classifier, other DRS methods obtain better accuracies than OWDA, but OWDA achieves much better MAP than other methods. For a test sequence, the NN classifier only employs its nearest training sequence when

calculating the accuracy, but ranks all training sequences according to the OPW distances w.r.t. it when calculating the MAP. OWDA minimizes the overall dispersion for sequence classes and maximizes the overall separability among classes. This makes most sequences from different classes more different, but does not pay special attention to the margins among classes. For a test sequence, the nearest training sequence may not belong to the same class due to noises or variances, but generally, most training sequences from the same class will be ranked in front of those from different classes.

4.3. Results on the ChaLearn dataset

The **ChaLearn Gesture Recognition dataset [11, 10]** contains 955 Italian gesture sequences captured by Kinect camera from 20 different Italian gestures. Because we focus on individual sequence classification rather than sequence detection or segmentation, we follow [42, 24, 12] to perform experiments on the segmented sequences given by the ground-truth segments. Each segmented sequence contains only one gesture instance. 27 persons performed these gestures. Other annotations of this dataset include the foreground segmentation and joint skeletons.

On this dataset, we employ the histogram-of-joint-positions-based frame-wise features provided by the authors of [12]. Specifically, for each frame, the relative locations of body joints are quantized w.r.t. a pre-clustered codebook, and the histogram of the quantized codewords serves as the feature with a dimensionality of 100. This dataset includes training set, validation set, and test set. Following [42, 24, 12], we learn the projections and train the classifiers on the training set, and report the results on the validation set. Other experimental settings remain the same as those on the MSR Action3D dataset.

Fig. 4 presents the results of different DRS methods as functions of the reduced dimension by both classifiers. For the SVM classifier, OWDA outperforms other methods by a margin of about 5% on most reduced dimensions. OWDA is the only DRS method that is able to improve the original features. Moreover, OWDA achieves this by preserving only 25 dimensions. This indicates that OWDA enhances the temporal separability and discards noises successfully. The performances of LDA and kLDA are far below those of other methods. The reason is that the observations in sequences are not independent. Performing LDA and kLDA forcibly by viewing them as independent samples not only aggravates the within-class ambiguity, but also may break their temporal relations. Moreover, LDA and kLDA can preserve $C - 1 = 19$ dimensions at most. It is difficult to separate sequences from different classes with such few dimensions. In contrast, since the barycenter of each class has $L = 8$ supporting points, OWDA is able to preserve $LC - 1 = 159$ dimensions, if $d > 159$.

Method	Precision	Recall	F-score
Wu et al. [40]	0.599	0.593	0.596
Pfister et al. [24]	0.612	0.623	0.617
Fernando et al. [13]	0.753	0.751	0.752
Cherian et al. [8]	0.753	0.752	0.751
LSDA+SVM [33]	0.768	0.767	0.767
OWDA+SVM	**0.773**	**0.773**	**0.772**

Table 1. Comparison with other methods on the ChaLearn dataset.

For the NN classifier, both OWDA and LSDA improve the original features greatly. Compared with LSDA, OWDA achieves comparable accuracy and much higher MAP. Specifically, OWDA outperforms the original features by a margin of 20%. The MAPs of OWDA are 5% higher than those of LSDA on almost all dimensions. Comparisons using other performance measures on the three datasets are presented in the supplementary file, where similar trends can be observed.

4.4. Training time

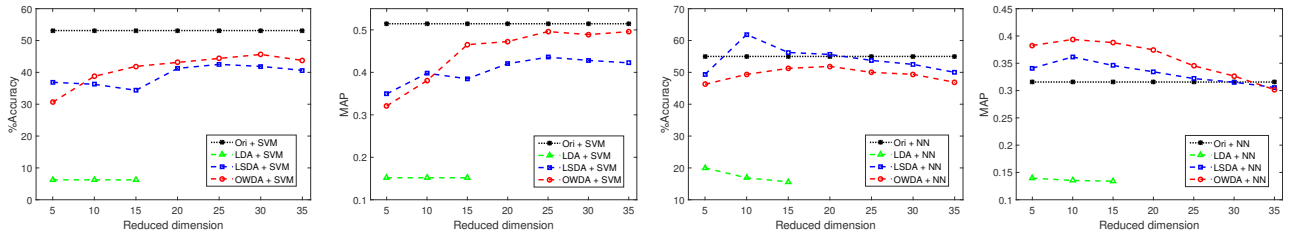
For OWDA, in most cases, the calculation of the barycenter converges in about 10 iterations. The procedures after learning the barycenters are closed-form calculations. Therefore, the practical training time is not too long. On the MSR Action3D dataset, the MSR Activity3D dataset, and the Chalearn dataset, the training times of OWDA are 43.1753, 265.7691, 385.8162 (sec), respectively.

4.5. Comparison with state-of-the-art methods

Our goal is not to design an end-to-end sequence classification method, but to develop a DRS method that produces low-dimensional discriminative temporal representations. Our method can serve as a ubiquitous component in different classification pipelines to improve the original representations and benefit the subsequent classifiers. For example, recurrent neural networks (RNNs) are seldom used for feature learning, but often as classifiers by taking handcrafted or CNN-learned frame-wide features as input. Our method can be applied to these features before they are fed into RNNs. In this way, RNNs can estimate fewer parameters and better capture the temporal dependencies.

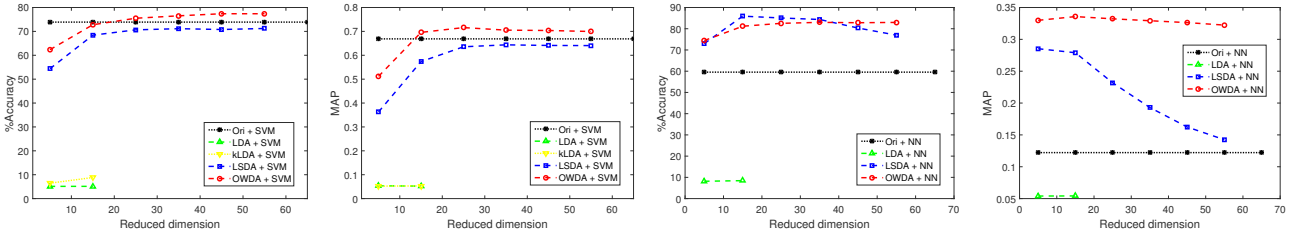
On the ChaLearn dataset, we have shown that our method outperforms other DRS methods and improves different sequence classification methods. We compare our results by using the frame-wide features in [12] and the SVM-based classifier with some other methods. Multi-class precision, recall, and F-score are used as performance measures as in [40, 24, 13, 8, 33]. Comparisons are shown in Tab. 1. Our method followed by a relatively simple SVM classifier with rank pooling outperforms other methods.

On the MSR Activity3D dataset, covariance representations and kernel-SVM based methods such as Ker-RP-



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

Figure 3. (a) Accuracies with the SVM classifier (b) MAPs with the SVM classifier (c) Accuracies with the NN classifier and (d) MAPs with the NN classifier as functions of the dimensionality of the subspace on the MSR Daily Activity3D dataset.



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

Figure 4. (a) Accuracies with the SVM classifier (b) MAPs with the SVM classifier (c) Accuracies with the NN classifier and (d) MAPs with the NN classifier as functions of the dimensionality of the subspace on the Chalearn Gesture dataset.

Method	Accuracy
Actionlet Ensemble [37]	85.8%
Moving Pose [45]	73.8%
COV-$J_{\mathcal{H}}$-SVM [15]	75.5%
Ker-RP-POL [39]	96.9%
Ker-RP-RBF [39]	96.3%
Kernelized-COV [6]	96.3%
LRTS [17]	80.6%
Qiao et al. [25]	75.0%
Baradel et al. [1]	90.0%
Luo et al. [22]	86.9%
Ji et al. [16]	81.3%
DSSCA SSLM [27]	97.5%
MDMTL [20]	93.8%
OWDA+Kernelized-COV	**98.1%**

Table 2. Comparison with state-of-the-art methods on the MSR Activity3D dataset.

POL [39] and Kernelized-COV [6] achieve superior results. Kernelized-COV employs the Kernelized covariance of all frame-wide features of a sequence as the representation of the sequence. Our proposed OWDA can be applied before Kernelized-COV to enhance the temporal representations. Specifically, we employ the frame-wide features provided in [39], which are based on the velocity and acceleration of the joint positions [45] and have a dimensionality of 120. We perform the proposed OWDA to reduce the dimension to 80 and then employ Kernelized-COV for classification. As shown in Tab. 2, the result obtained in this way outperforms the state-of-the-art results on this dataset.

5. Conclusion

In this paper, we have presented a linear DRS method, *i.e.*, OWDA, to map the non-independent observations in sequences onto a low-dimensional subspace, so that the entire sequences from different classes are better discriminated with the OPW distance. To manipulate the structured sequences with various lengths, we learn the OPW barycenter of the sequence samples from a class to represent the average temporal structures and evolutions. We construct the covariance of the class in such a way that the trace of the covariance measures the variability of the OPW distances between the sequence samples and the barycenter. Similarly, we construct the pair-wise inter-class scatter so that the trance of the scatter measures the OPW distance between the corresponding barycenters of the two classes. We show that the intra- and inter-class scatters are actually the weighted sums of all the pairwise out-products between observations in sequences or elements of barycenters. Therefore, all the local relationships are learned and incorporated. Experimental results on the three 3D action datasets demonstrate the effectiveness of the proposed OWDA.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61603373, No. 61976206, Youth Innovation Promotion Association CAS No. 2019110, National Science Foundation grant IIS-1619078, IIS-1815561, and the Army Research Office ARO W911NF-16-1-0138.

References

[1] Fabien Baradel, Christian Wolf, and Julien Mille. Pose-conditioned spatio-temporal attention for human action recognition. *arXiv preprint arXiv:1703.10106*, 2017.

[2] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[3] Jean David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *Siam Journal on Scientific Computing*, 37(2), 2014.

[4] W. Bian and D. Tao. Max-min distance analysis by using sequential sdp relaxation for dimension reduction. *TPAMI*, 33(5):1037–1050, 2011.

[5] L. M Brègman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Ussr Computational Mathematics & Mathematical Physics*, 7(3):200–217, 1967.

[6] Jacopo Cavazza, Andrea Zunino, Marco San Biagio, and Vittorio Murino. Kernelized covariance for action recognition. In *ICPR*. IEEE, 2016.

[7] Li Fen Chen, Hong Yuan Mark Liao, Ming Tat Ko, Ja Chen Lin, and Gwo Jong Yu. A new lda-based face recognition system which can solve the small sample size problem. *PR*, 33(10):1713–1726, 2000.

[8] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. Generalized rank pooling for activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3222–3231, 2017.

[9] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, 2014.

[10] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Isabelle Guyon, Vassilis Athitsos, Hugo Escalante, Leonid Sigal, Antonis Argyros, Cristian Sminchisescu, et al. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *ICMI*, 2013.

[11] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ICMI*, 2013.

[12] Basura Fernando, Efstratios Gavves, Jose Oramas M., Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015.

[13] Basura Fernando, Efstratios Gavves, José Oramas M., Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *TPAMI*, 39(4):773–787, 2017.

[14] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *arXiv preprint arXiv:1608.08063*, 2016.

[15] Mehrtash Harandi, Mathieu Salzmann, and Fatih Porikli. Bregman divergences for infinite dimensional covariance matrices. In *CVPR*, 2014.

[16] Xiaopeng Ji, Jun Cheng, Wei Feng, and Dapeng Tao. Skeleton embedded motion body partition for human action recognition using depth sequences. *Signal Processing*, 143:56–68, 2018.

[17] Chengcheng Jia and Yun Fu. Low-rank tensor subspace learning for rgb-d action recognition. *TIP*, 25(10):4641–4652, 2016.

[18] Rémi Lajugie, Damien Garreau, Francis Bach, and Sylvain Arlot. Metric learning for temporal sequence alignment. In *NIPS*, 2014.

[19] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *IEEE Int'l Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[20] An-An Liu, Ning Xu, Wei-Zhi Nie, Yu-Ting Su, and Yong-Dong Zhang. Multi-domain & multi-task learning for human action recognition. *TIP*, 2019.

[21] M. Loog and R.P.W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *TPAMI*, 26(6):732–739, 2004.

[22] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *ICCV*, 2017.

[23] Benjamin Paaßen, Claudio Gallicchio, Alessio Micheli, and Barbara Hammer. Tree edit distance learning via adaptive symbol embeddings. In *ICML*, 2018.

[24] Tomas Pfister, James Charles, and Andrew Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *ECCV*, 2014.

[25] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. *PR*, 66:202–212, 2017.

[26] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[27] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *TPAMI*, 40(5):1045–1058, 2018.

[28] Alex Shyr, Raquel Urtasun, and Michael I Jordan. Sufficient dimension reduction for visual sequence classification. In *CVPR*, 2010.

[29] Bing Su and Xiaoqing Ding. Linear sequence discriminant analysis: a model-based dimensionality reduction method for vector sequences. In *ICCV*, 2013.

[30] Bing Su, Xiaoqing Ding, Changsong Liu, Hao Wang, and Ying Wu. Discriminative transformation for multi-dimensional temporal sequences. *TIP*, 26(7):3579–3593, 2017.

[31] Bing Su, Xiaoqing Ding, Changsong Liu, and Ying Wu. Heteroscedastic max-min distance analysis. In *CVPR*, 2015.

[32] Bing Su, Xiaoqing Ding, Changsong Liu, and Ying Wu. Heteroscedastic maxmin distance analysis for dimensionality reduction. *TIP*, 27:4052–4065, 2018.

[33] Bing Su, Xiaoqing Ding, Hao Wang, and Ying Wu. Discriminative dimensionality reduction for multi-dimensional sequences. *TPAMI*, 2018.

[34] Bing Su and Gang Hua. Order-preserving wasserstein distance for sequence matching. In *CVPR*, 2017.

[35] Bing Su and Gang Hua. Order-preserving optimal transport for distances between sequences. *TPAMI*, 2018.

[36] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008.

[37] Jiang Wang, Zicheng Liu, and Ying Wu. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[38] Jiang Wang and Ying Wu. Learning maximum margin temporal warping for action recognition. In *ICCV*, 2013.

[39] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li. Beyond covariance: Feature representation with nonlinear kernel matrices. In *ICCV*, 2015.

[40] Jiaxiang Wu, Jian Cheng, Chaoyang Zhao, and Hanqing Lu. Fusing multi-modal features for gesture recognition. In *ACM on International conference on multimodal interaction*, 2013.

[41] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *TPAMI*, 29(1):40–51, 2007.

[42] Angela Yao, Luc Van Gool, and Pushmeet Kohli. Gesture recognition portfolios for personalization. In *CVPR*, 2014.

[43] Jieping Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *JMLR*, 6:483–502, 2005.

[44] Jieping Ye, Ravi Janardan, Qi Li, and Haesun Park. Feature extraction via generalized uncorrelated linear discriminant analysis. In *ICML*, 2004.

[45] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.

[46] Yu Zhang and Dit-Yan Yeung. Worst-case linear discriminant analysis. In *NIPS*, 2010.

[47] M. Zhu and A.M. Martinez. Subclass discriminant analysis. *TPAMI*, 28(8):1274–1286, 2006.