Multi-View Time Series Classification: A Discriminative Bilinear Projection Approach

Sheng Li
Dept. of ECE
Northeastern University
Boston, MA, USA
shengli@ece.neu.edu

Yaliang Li
Dept. of CSE
SUNY Buffalo
Buffalo, NY, USA
yaliangl@buffalo.edu

Yun Fu
Dept. of ECE & College of CIS
Northeastern University
Boston, MA, USA
yunfu@ece.neu.edu

ABSTRACT

By virtue of the increasingly large amount of various sensors, information about the same object can be collected from multiple views. These mutually enriched information can help many real-world applications, such as daily activity recognition in which both video cameras and on-body sensors are continuously collecting information. Such multivariate time series (m.t.s.) data from multiple views can lead to a significant improvement of classification tasks. However, the existing methods for time series data classification only focus on single-view data, and the benefits of mutual-support multiple views are not taken into account. In light of this challenge, we propose a novel approach, named Multi-view Discriminative Bilinear Projections (MDBP), for extracting discriminative features from multi-view m.t.s. data. First, MDBP keeps the original temporal structure of m.t.s. data, and projects m.t.s. from different views onto a shared latent subspace. Second, MDBP incorporates discriminative information by minimizing the within-class separability and maximizing the between-class separability of m.t.s. in the shared latent subspace. Moreover, a Laplacian regularization term is designed to preserve the temporal smoothness within m.t.s.. Extensive experiments on two real-world datasets demonstrate the effectiveness of our approach. Compared to the state-of-the-art multi-view learning and m.t.s. classification methods, our approach greatly improves the classification accuracy due to the full exploration of multi-view streaming data. Moreover, by using a feature fusion strategy, our approach further improves the classification accuracy by at least 10%.

Keywords

Multi-view learning; bilinear projections; discriminative regularization; time series classification

1. INTRODUCTION

Nowadays information about one object can be continuously collected from multiple views in many domains such as health care and entertainment, due to the increasingly large amount of various sensors. The collected data can be represented as multi-view multivariate time series, which could lead to significant improvement of data mining tasks like classification. For instance, the daily activities of a subject can be captured by video cameras, depth cameras, and on-body sensors. These three sets of heterogeneous and dynamic measurements would provide mutually enriched information, which are helpful for improving the performance of activity recognition. In general, it has been well recognized that multi-view data usually enhance the overall model performance than single-view data, as long as the different views contain diverse information [3]. In this paper, we focus on the classification of multi-view multivariate time series, which plays a central role in extracting useful knowledge from the multi-view streaming data.

Although the classification of time-series data has been extensively studied during the past decade, they are only designed for single-view data. Traditional methods focus on the univariate time series (u.t.s.) classification [21, 44], by defining distance measures (e.g., dynamic temporal wrapping (DTW) [43], recurrence plot [39], edit distance [35] and elastic distance [32]), or extracting compact and effective features (e.g., time series shapelets [47] and segment based features [48]). Furthermore, as the streaming data might be characterized by multiple measurements simultaneously and represented as multivariate time series (m.t.s.), some recent works try to extract informative patterns from m.t.s., and have achieved promising results [36, 49, 23]. However, these methods cannot directly handle multi-view data, and the benefits of mutual-support multiple views are not taken into account.

On the other hand, multi-view learning has attracted increasing attention in recent years [45, 10], since it sophisticatedly models the consistency and diversity among multiple data views, and significantly boosts the learning performance than single-view methods. The popular multi-view learning algorithms are usually categorized as co-training, multiple kernel learning, and subspace learning [45]. However, existing multi-view algorithms are not customized for time series classification, as they simply ignore the unique properties of time series, such as the temporal smoothness.

To address the above challenges, we propose a novel approach, named Multi-view Discriminative Bilinear Projections (MDBP), for multi-view m.t.s. classification. Figure 1 illustrates the framework of our approach. MDBP aims to extract discriminative features from multi-view m.t.s. data, and it models the view consistency and temporal dynamics.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983780

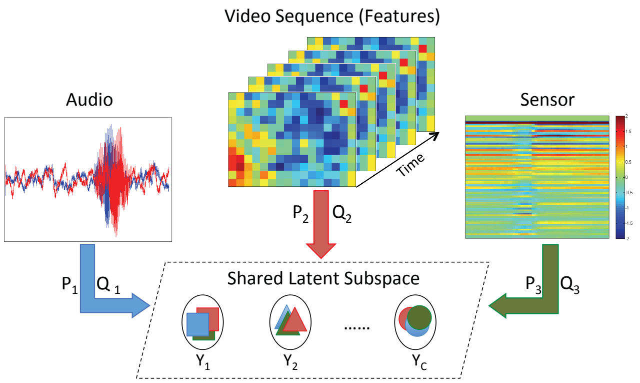


Figure 1: Framework of our MDBP approach. The Multimodal Spoken Word dataset contains three data views, including video, audio, and magnetic sensors. MDBP maps multi-view data onto a shared subspace through a pair of view-specific bilinear projections (i.e., P_v and Q_v), and incorporates discriminative information by enhancing the between-class separability. Y_i is the compact representation of i-th class in the latent space.

First, we assume that a m.t.s. sample and its counterparts observed in other views could share a compact representation in a low-dimensional latent subspace, as they indeed represent the same object. To preserve the original temporal structure in m.t.s., MDBP learns a pair of view-specific bilinear projections, which separately reduce the dimensions of measurements and timestamps. Second, MDBP enforces that samples belong to the same class share the same latent representation in the shared subspace, which reduces the within-class separability. Also, the latent representations of different classes are pushed away from each other, in order to enhance the between-class separability. Third, a Laplacian regularization term is designed to preserve the temporal smoothness of m.t.s. after projection. An efficient optimization algorithm based on gradient descent is designed to solve the problem. We evaluate the classification performance of our approach and baseline methods on two real-wold datasets including the UCI Daily and Sports Activity dataset, and the Multimodal Spoken Word dataset. Extensive results on both single-view and multi-view test scenarios demonstrate the superiority of our approach over the state-of-the-art methods. Our approach greatly improves the classification accuracy due to the full exploration of multi-view m.t.s. data.

The main contributions of this paper are summarized as follows.

- We propose a discriminative bilinear projection approach, MDBP, for multi-view multivariate time-series classification. To the best of our knowledge, this paper is the first attempt to apply multi-view dimensionality reduction for the m.t.s. classification problem.
- We model the view consistency by projecting multi-view m.t.s. onto a shared subspace, and incorporate the discriminative regularization and temporal smoothness regularization.
- We conduct extensive experiments on two real-world datasets, which demonstrate the effectiveness of our approach, compared to the state-of-the-art multi-view learning methods and m.t.s. classification methods.

Table 1: Notations

Notations	Descriptions
X_{vij}	The j-th m.t.s. sample in i-class, v-th view
P_v	Bilinear projection for the v-th view
Q_v	Bilinear projection for the v-th view
Y_i	Representation of i-th class in latent space
L_p	Temporal Laplacian matrix
V	Number of views
C	Number of classes
N_i	Number of samples in the i-th class
d_v	Dimensionality of m.t.s. sample in v-th view
m_v	Length of m.t.s. sample in v-th view

The rest of the paper is organized as follows. In Section 2, we formally define several basic concepts, and present the problem of multi-view m.t.s. classification. In Section 3, we propose the MDBP approach and describe the optimization algorithm. The experimental results and discussions are reported in Section 4. In Section 5, we review the related works and discuss how they differ from our approach. Section 6 is the conclusion.

2. PROBLEM DEFINITION

In this section, we first present the definitions of *multivariate time series (m.t.s.)* and *multi-view m.t.s.*, and then formally define the problems of *m.t.s. classification* and *multi-view m.t.s. classification.* Table 1 summarizes the notations used throughout this paper.

Definition 1. **Multivariate Time Series (m.t.s.).** *A multivariate time series (m.t.s.) $X = [x_1, \cdots, x_m] \in \mathbb{R}^{d\times m}$ is an ordered sequence of d-dimensional vectors, in which x_i is the observation at the i-th timestamp, and m is the length of time series.*

Nowadays time-series data are usually collected from multiple views or multiple modalities. We define the multi-view multivariate time series as follows.

Definition 2. **Multi-View Multivariate Time Series (m.t.s.).** *A multi-view m.t.s. $\hat{X} = \{X_{(v)}\}$, $v = 1, \cdots, V$ is a set of time series data collected from multiple views, where $X_{(v)} \in \mathbb{R}^{d_v \times m_v}$ denotes the time series observed in the v-th view, d_v is the number of measurements of time series, m_v is the length of time series, and V is the total number of views.*

For simplicity, we assume that the time series within the same view have been synchronized and preprocessed, and therefore they have the same length m_v.

We focus on the classification task in this paper. The formal definition of multivariate time series classification is as follows.

Definition 3. **Multivariate Time Series (m.t.s.) Classification.** *Let $\mathcal{C}$ denote a set of class labels, and $C = |\mathcal{C}|$ is the total number of classes. The task of m.t.s. classification is to learn a classier, which is a function $\mathcal{F} : \mathbf{X} \rightarrow \mathcal{C}$, where $\mathbf{X}$ is a set of m.t.s..*

The traditional m.t.s. classification algorithms are mainly designed for single-view data, and they cannot directly handle the multi-view data. Although some practical tricks might be adopted, such as vectorizing multi-view data to single-view ones, we argue that considerable information might be discarded during this process. In this paper, we extend the single-view m.t.s. classification problem to the multi-view setting, and present the formal definition as follows.

Definition 4. **Multi-View Multivariate Time Series Classification.** *Let $\hat{\mathbf{X}} = \{X_{i,v} | i = 1, \cdots, N, v = 1, \cdots, V\}$ denote a set of multi-view m.t.s., where N is the number of m.t.s. in each view, and $\mathcal{C} = \{C_1, \cdots, C_c\}$ denote a set of class labels shared by V views. The task of multi-view m.t.s. classification is to learn a classier from $\hat{\mathbf{X}}$, and therefore to infer the class label for test m.t.s. X_{test} which might be observed in any view.*

We notice that the basic m.t.s. classification can be considered as a special case of multi-view m.t.s. classification when there is only one view available. In the multi-view case, the m.t.s. classification problem becomes more challenging. For example, the consistency between multiple views should be modeled.

3. MULTI-VIEW DISCRIMINATIVE BILINEAR PROJECTION (MDBP)

In this section, we propose the multi-view discriminative bilinear projections (MDBP) approach for m.t.s. classification. We first introduce our motivation, and then present the model details. Finally, the optimization algorithm with discussions is provided.

3.1 Motivation

For multi-view m.t.s. classification, several key problems should be taken into account, including:

(1) *How to build the consistency and interactions of m.t.s. from multiple views?*

(2) *How to extract discriminative features from multi-view m.t.s.?*

(3) As the data size has to be increased in the multi-view case, *how to improve the computational efficiency of training and test?*

We aim to address all of the above challenges by designing a multi-view dimensionality reduction approach. The basic idea is to project multi-view data onto a common subspace, and then perform the classification of m.t.s. using the low-dimensional representations. Our motivations of using dimensionality reduction are three-folds. First, as multi-view data are usually drawn from diverse data spaces, seeking common projections would allow us to bridge the gap between multiple data views. Learning a shared subspace is also considered as a popular strategy in multi-view learning [13, 8]. Second, specifically designed regularization functions, such as discriminative regularization, can be naturally incorporated into the multi-view dimensionality reduction framework. Third, we aim to learn linear projections for each view, and therefore, the training and test would be efficient.

3.2 Formulation of MDBP

We assume that a set of m.t.s. $\hat{\mathbf{X}}$ observed from V views belong to C different classes, $\hat{\mathbf{X}} = \{X_{vij} | v = 1, \cdots, V,\ i = 1, \cdots, C,\ j = 1, \cdots, N_i\}$, where N_i is the number of samples in the i-class for each view.

The general formulation of MDBP is

$$\min_{P,Q,Y} \Phi(\hat{\mathbf{X}}, P, Q, Y) + \lambda_1 \Theta(Y) + \lambda_2 \Omega(P, X), \tag{1}$$

where P and Q are bilinear projections, and Y is the low-dimensional representation of $\hat{\mathbf{X}}$. λ_1 and λ_2 are two trade-off parameters that balance the effects of different terms. Eq (1) contains three components. The first term $\Phi(\hat{\mathbf{X}}, P, Q, Y)$ represents the multi-view bilinear dimensionality reduction, which characterizes the connections between different views. The second term $\Theta(Y)$ carries discriminative regularization, and the last term $\Omega(P, X)$ models the temporal smoothness. We will detail the three components in the following.

Learning Shared Representations Across Views

We propose to learn bilinear projections for reducing the dimensionality of m.t.s.. Let $P_v \in \mathbb{R}^{d_v \times p}$ and $Q_v \in \mathbb{R}^{m_v \times q}$ denote a pair of linear projections for the v-th view, and then the m.t.s. X_{vij} can be transformed by

$$Y_{vij} = P_v^\top X_{vij} Q_v, \tag{2}$$

where $Y_{vij} \in \mathbb{R}^{p \times q}$ is the low-dimensional representation of X_{vij}.

The major benefits of employing bilinear projections are two-folds. First, bilinear projections allow us to preserve the original structure of m.t.s., especially the temporal structures, which makes it easier to incorporate temporal smoothness regularizations along the time dimension. Second, compared to other dimensionality reduction methods, bilinear projections have less computational cost for both training and test, which is suitable for dealing with long-duration time series data.

Eq. (2) assumes that each view shares a pair of linear projections. However, it doesn't take view correlation into account. A more reasonable assumption is that, a sample and its counterparts collected from other views could have the same low-dimensional representation in a common subspace. Moreover, as we focus on classification tasks, we further assume that samples from the same class, no matter which views they belong to, would share approximately the same representations in the common subspace. Therefore, we rewrite Eq. (2) as: $Y_i \approx P_v^\top X_{vij} Q_v$, which encourages samples of the same class from all the views to be as close as possible in the common subspace.

Then we formulate the multi-view dimensionality reduction term $\Phi(\hat{\mathbf{X}}, P, Q, Y)$ as

$$\Phi(\hat{\mathbf{X}}, P, Q, Y) = \sum_{i=1}^{C} \sum_{v=1}^{V} \sum_{j=1}^{N_i} \left\| X_{vij} - P_v Y_i Q_v^\top \right\|_{\mathrm{F}}^2, \tag{3}$$

where $\| \cdot \|_{\mathrm{F}}$ is the matrix Frobenius norm. Here we assume that the projections P_v and Q_v are semi-orthogonal matrices, i.e., $P_v^\top P_v = \mathbf{I_p}$ and $Q_v^\top Q_v = \mathbf{I_q}$.

Incorporating Discriminative Regularization

For classification tasks, the learned low-dimensional representations via dimensionality reduction should be discriminative. Actually, $\Phi(\hat{\mathbf{X}}, P, Q, Y)$ in Eq. (3) already makes use of the label information, as it maps the same-class samples onto a stationary point in the low-dimensional common space. It implicitly incorporates discriminative information, however, the separability among classes hasn't been included, which is a key for classification problems as suggested by the Fisher criterion [11].

Therefore, to explicitly incorporate the discriminative information, we push the low-dimensional representations of different classes, Y_i and Y_k $(i \neq k)$, far away from each other. The discriminative regularization term $\Theta(Y)$ is defined as

$$\Theta(Y) = -\sum_{i=1}^{C} \sum_{k=1,k\neq i}^{C} \|Y_i - Y_k\|_{\mathrm{F}}^2 . \tag{4}$$

As we need to maximize the summation of pairwise distances between Y_i and Y_k, a negative sign is added in order to use $\Theta(Y)$ in the minimization problem Eq. (1).

The discriminative regularization shown in Eq.(4) is view-independent, as it is implemented in the shared subspace. This strategy not only simplifies the model complexity, but also closely relates to the final classification task that is usually performed in the low-dimensional subspace.

Modeling Temporal Smoothness

In reality, many types of time series data, such as human activities, slightly change in successive timestamps, such as the smooth transitions of human activities [26]. In other words, time series data own the property of locally smoothness, which brings informative prior knowledge for learning models. By using bilinear projections, our model does not break the temporal structures of input time series X, in which the temporal smoothness is usually observed. However, after projecting X to a low-dimensional subspace via P_v, the temporal smoothness might be undermined in the projected data $P_v X$.

To address this problem, we aim to design a smoothness regularization term on $P_v X_{vk}$, where X_{vk} is the k-th sample in the v-th view. In light of the Laplacian regularization [16], we propose a multi-view temporal Laplacian regularization $\Omega(P_v, X_{vk})$ to enforce the smoothness as follows

$$\begin{aligned}
\Omega(P_v, X_{vk}) &= \tfrac{1}{2} \textstyle\sum_{i,j=1}^{N} W_{ij} \left\| P_v^\top X_{vk(,i)} - P_v^\top X_{vk(,j)} \right\|_2^2 \\
&= \textstyle\sum_{i=1}^{N} P_v^\top X_{vk(,i)} D_{ii} X_{vk(,i)}^\top P_v - \sum_{i,j=1}^{N} P_v^\top X_{vk(,i)} W_{ij} X_{vk(,j)}^\top P_v \\
&= \mathrm{tr}(P_v^\top X_{vk} D X_{vk}^\top P_v - P_v^\top X_{vk} W X_{vk}^\top P_v) \\
&= \mathrm{tr}(P_v^\top X_{vk} (D - W) X_{vk}^\top P_v) \\
&= \mathrm{tr}(P_v^\top X_{vk} (L_P) X_{vk}^\top P_v),
\end{aligned} \tag{5}$$

where $X_{vk(,i)} \in \mathbb{R}^{d \times 1}$ is the i-th column in X_{vk}, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, W is a pre-defined weight matrix that carries the smoothness prior, D is a diagonal matrix whose entries are $D_{ii} = \sum_j W_{ij}$, and $L_P (= D - W)$ is the Laplacian matrix.

Let Z_{vk} denote the projected feature of X_{vk}, $Z_{vk} = P_v^\top X_{vk}$. It is clear that each column in X_{vk} or Z_{vk} corresponds to a timestamp. In reality, successive neighbors in X_{vk} usually slightly change over time, which can be considered as prior information of temporal smoothness. By setting a proper weighting matrix W, we can transfer such temporal smoothness from X_{vk} to Z_{vk} using the Laplacian regularization $\Omega(P_v, X_{vk})$. Let s denote the number of successive neighbors, the entry in W is computed as

$$W_{ij} = \begin{cases} 1, \text{ if } |i - j| \leq \frac{s}{2} \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

In this way, the successive columns in Z_{vk} are encouraged to be similar to each other. Note that we only adopt binary weights in Eq. (6). Other sophisticated graph weighting schemes could also be employed to construct W.

Then, the regularization term $\Omega(P, X)$ used in Eq. (1) is defined as a summation of $\Omega(P_v, X_{vk})$ over all of the views and samples

$$\Omega(P, X) = \sum_{v=1}^{V} \sum_{k=1}^{N} \Omega(P_v, X_{vk}). \tag{7}$$

Objective Function

To sum up, the objective function of our MDBP approach is:

$$\begin{aligned}
\min_{P_v, Q_v, Y_i} \quad & f(P_v, Q_v, Y_i) = \sum_{i=1}^{C} \sum_{v=1}^{V} \sum_{j=1}^{N_i} \left\| X_{vij} - P_v Y_i Q_v^\top \right\|_{\mathrm{F}}^2 \\
& -\lambda_1 \sum_{i=1}^{C} \sum_{k=1,k\neq i}^{C} \|Y_i - Y_k\|_{\mathrm{F}}^2 \\
& +\lambda_2 \sum_{v=1}^{V} \sum_{k=1}^{N} \mathrm{tr}(P_v^\top X_{vk} (L_P) X_{vk}^\top P_v) \\
s.t. \quad & P_v^\top P_v = \mathbf{I_p},\ Q_v^\top Q_v = \mathbf{I_q}, v = 1, \cdots, V.
\end{aligned} \tag{8}$$

In Eq. (8), orthogonal constraints $P_v^\top P_v = \mathbf{I_p}$ and $Q_v^\top Q_v = \mathbf{I_q}$ are incorporated, where $\mathbf{I_p} \in \mathbb{R}^{p \times p}$ and $\mathbf{I_q} \in \mathbb{R}^{q \times q}$ are two identity matrices. Orthogonality in a projection matrix means that any two basis vectors in this projection are orthogonal to each other, which has the advantages of compactness and reducing redundancy.

3.3 Optimization Algorithm

We develop an efficient optimization algorithm based on gradient descent to solve the problem in Eq. (8).

Although Eq. (8) is not jointly convex to all the variables P_v, Q_v and Y_i, it is convex to each of them when the other variables are fixed. We use gradient descent to alternately update each variable. Given $P_v^{(t)}, Q_v^{(t)}, Y_i^{(t)}$ obtained in the t-th step, the update rules at the $t+1$ step are

$$P_v^{(t+1)} = P_v^{(t)} - \gamma \tfrac{\partial}{\partial P_v} f(P_v, Q_v, Y_i),\ v = 1, \cdots, V, \tag{9}$$

$$Q_v^{(t+1)} = Q_v^{(t)} - \gamma \tfrac{\partial}{\partial Q_v} f(P_v, Q_v, Y_i),\ v = 1, \cdots, V, \tag{10}$$

$$Y_i^{(t+1)} = Y_i^{(t)} - \gamma \tfrac{\partial}{\partial Y_i} f(P_v, Q_v, Y_i),\ i = 1, \cdots, C, \tag{11}$$

where γ is the learning rate.

The detailed derivatives are shown below

$$\begin{aligned}
\tfrac{\partial}{\partial P_v} = \quad & -\sum_{i=1}^{C} \sum_{j=1}^{N_i} 2(X_{vij} - P_v Y_i Q_v^\top) Q_v Y_i^\top \\
& +\lambda_2 \sum_{k=1}^{N} 2 P_v^\top X_{vk} (L_P) X_{vk}^\top.
\end{aligned} \tag{12}$$

$$\tfrac{\partial}{\partial Q_v} = \quad -\sum_{i=1}^{C} \sum_{j=1}^{N_i} 2 Y_i^\top P_v^\top (X_{vij} - P_v Y_i Q_v^\top). \tag{13}$$

Algorithm 1 Solving Problem in Eq. (8)

```
Input: Multi-view m.t.s. sample set $\hat{\mathbf{X}}$,
       parameters $\lambda_1$, $\lambda_2$, $s$, $\gamma$, $maxIter$.
Output: Bilinear projections $P_v$, $Q_v$
        class-specific shared representation $Y_i$.
 1: Normalize each time series sample;
 2: Compute the Laplacian matrix $L_p$ according to Eq. (5)
    and Eq. (6);
 3: Initialize $P_v$, $Q_v$ and $Y_i$ with random matrices;
 4: for loop $t$ from 1 to $maxIter$ do
 5:     for view $v$ from 1 to $V$ do
 6:         Update projection $P_v$ using Eq. (9);
 7:         Orthogonalize $P_v$;
 8:         Update projection $Q_v$ using Eq. (10);
 9:         Orthogonalize $Q_v$;
10:     end for
11:     for class $i$ from 1 to $C$ do
12:         Update latent presentation $Y_i$ using Eq. (11);
13:     end for
14:     if the objective converges then
15:         Return $P_v$, $Q_v$ and $Y_i$.
16:     end if
17: end for
```

$$\begin{aligned}\frac{\partial}{\partial Y_i} = \quad & -\sum_{i=1}^{C}\sum_{j=1}^{N_i} 2P_v^\top (X_{vij} - P_v Y_i Q_v^\top) Q_v \\ & -\lambda_1 \sum_{k=1, k\neq i}^{C} 2(Y_i - Y_k).\end{aligned} \tag{14}$$

Note that the orthogonal constraints shown in Eq. (8) are implemented by a post-processing step during the update. The complete optimization algorithm is summarized in Algorithm 1. We will show the convergence property of our algorithm in Section 4.3.

After obtaining the subspaces P_v and Q_v, the nearest neighbor classifier can be employed to classify a test m.t.s. T_v. The complete procedures of MDBP are provided in Algorithm 2.

Time Complexity Analysis

The computational cost of Algorithm 1 mainly depends on the Step 6, Step 8, and Step 12, which cost $\mathbf{O}(N(dpq + dqm + pm^2))$, $\mathbf{O}(N(dpq + dqm))$, and $\mathbf{O}(N(pdm + pmq))$, respectively. Indeed, our algorithm reduces the dimensionality of time series, which means $p \ll d$ and $q \ll m$. Thus, the overall time complexity of the three steps is simplified to $\mathbf{O}(N(dm + m^2))$.

In addition, our algorithm converges after several iterations, and there are usually a few views in reality. It indicates that our approach is approximately linear to the sample size N when $N \gg \max(d, m)$, and therefore, our approach can be easily deployed for large-scale applications.

3.4 Comparison with Existing Methods

The first term in Eq. (1), $\sum_{i=1}^{C}\sum_{v=1}^{V}\sum_{j=1}^{N_i} \left\|X_{vij} - P_v Y_i Q_v^\top\right\|_{\mathrm{F}}^2$, looks similar to the formulation of matrix tri-factorization [29], which also factorizes a data matrix into three unknown components. However, our approach is motivated from the multi-view learning scenario, and the factorized components carry consistency constraints across views or across classes. For instance, the view-specific projection P_v is shared by every sample in the v-th view.

Algorithm 2 MDBP Approach

```
Input: Multi-view m.t.s. training sample set $\hat{\mathbf{X}}$,
       single-view m.t.s. test sample $T_v$.
Output: Predicted class label $c_t$ for $T_v$
Normalize each time series sample;
Calculate the projections $P_v$ and $Q_v$ using Algorithm 1;
Project $X_{vi},\ i = 1, \cdots, N$, to the shared subspace by
  $Z_{vi} = P_v^\top X_{vi} Q_v$;
Project $T_v$ to the shared subspace by $\hat{Z}_v = P_v^\top T_v Q_v$;
Predict the class label of $T_v$ using NN classifier, by com-
  paring $\hat{Z}_v$ with $Z_{vi}$.
```

Although some existing multi-view learning algorithms also project multi-view data to a common subspace [13, 8, 20], our approach differs from them in that: (1) we employ the bilinear projections to map high-dimensional m.t.s. to a shared low-dimensional subspace; (2) we design a novel discriminative regularization term for multi-view dimensionality reduction. (3) we focus on the time series data classification, and design a Laplacian regularization term to enforce the temporal smoothness.

4. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the classification performance of our approach and baseline methods on two datasets, and perform quantitative analysis on parameter sensitivity.

4.1 UCI Daily and Sports Activity Dataset

The UCI Daily and Sports Activity Dataset [2, 31] contains motion sensor data of 19 daily and sports activities, such as sitting, standing, walking, running, jumping, etc.. Each activity is performed by 8 subjects (4 female and 4 male, between the ages 20 and 30) for 5 minutes. In particular, the subjects are asked to perform these activities in there own styles without any restrictions. As a result, the time series samples for each activity have considerable inter-subject variations in terms of speed and amplitude, which makes it difficult for accurate classification. During the data collection, nine sensors are put on each of the following five units: torso, right arm, left arm, right leg, and left leg. Thus, there are 45 sensors in total, and each sensor is calibrated to acquire data at 25 Hz sampling frequency. The 5-minute time series collected from each subject is divided into 5-second segments. For each activity, the total number of segments is 480, and each segment is considered as a m.t.s. sample of size 45 × 125, corresponding to 45 sensors and 125 timestamps.

Two-View Setting

We design a two-view experimental setting on the UCI Daily and Sports Activity dataset. Specifically, the first 27 sensors on torso, right arm and left arm are treated as View-1, while the rest 18 sensors on right leg and left leg as View-2. The activities are observed from two distinct views (i.e., two groups of sensors) simultaneously. Also, the m.t.s. samples in two views have the same number of timestamps.

Baselines

Our MDBP approach is a multi-view dimensionality reduction method for time series classification. We mainly compare it with single-view and multi-view dimensionality reduction methods. The single-view methods include principal component analysis (PCA) [19], linear discriminant analysis (LDA) [11], locality preserving projections (LPP) [41], and two-dimensional LDA (2DLDA) [46]. The multi-view methods include canonical correlation analysis (CCA) [17] and multi-view discriminant analysis (MvDA) [20]. In addition, we also compare our approach with a popular classification method, support vector machine (SVM) [7], and the state-of-the-art time series classification method, one-nearest-neighbor dynamic time warping (1NN-DTW) [43]. For all the baselines except 2DLDA, we have to vectorize each m.t.s. sample into a single vector. Our approach and 2DLDA learn linear projections without vectorizing m.t.s..

Results

There are 480 samples for each activity per view. We randomly choose $N_{tr} \in \{10, 20, 30, 40, 50\}$ samples from each activity (per view) to construct the training set, and the remaining samples are used to construct the test set. For singe-view baselines, we separately train two models on the training sets of two views, and report the classification accuracy on each view. For multi-view methods, we train the model by jointly using samples from two views, and also report the accuracy on each view. The parameters in our approach and baselines are tuned using 5-fold cross validation on the training set. The learning rate γ in our approach is empirically set to 0.01. We will analyze the parameter sensitivity of our approach in Section 4.3.

We randomly choose N_{tr} training samples from each activity 10 times, and report the average classification accuracy of our approach and baselines in Table 2. Our observations are:

- For smaller training sets (e.g., $N_{tr} = 10$), the supervised dimensionality reduction methods like LDA usually achieve higher accuracies than unsupervised methods such as PCA and LPP. The reason is that unsupervised methods cannot accurately estimate the data distribution without sufficient sampling, while supervised information used in LDA play a critical role in this scenario. When the training set grows, PCA achieves comparable results than LDA, and LPP outperforms LDA significantly.
- By preserving the original temporal structure of m.t.s. data, 2DLDA obtains the best results among all the single-view methods, but it cannot make use of the complementary information from multiple views.
- The multi-view methods usually perform better than single-view methods. For instance, the unsupervised multi-view method CCA always obtains higher accuracies than PCA and LPP in the case of View-1; the supervised multi-view method MvDA performs best among all the baseline methods, which demonstrates the effectiveness of multi-view learning and supervised regularization.
- Our approach achieves the highest classification accuracy in every case. Compared to MvDA, the accuracy is improved by at least 6% on average. It demonstrates the superiority of incorporating discriminative information and temporal smoothness to multi-view dimensionality reduction.

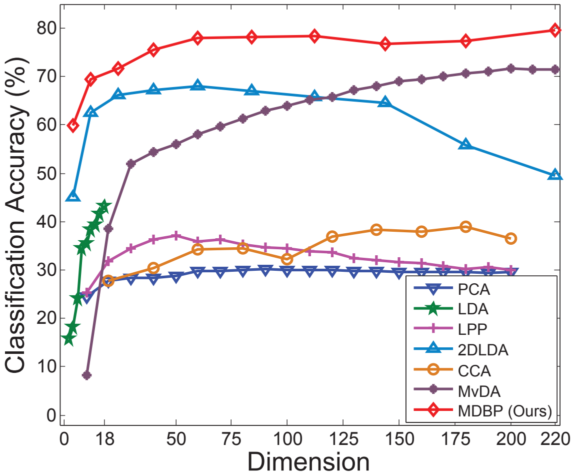


Figure 2: Classification accuracy (%) with different dimensions on View-1 of UCI Daily and Sports Activity dataset.

Choosing a proper dimension plays a key role in various dimensionality reduction methods, which is still an open problem to date. Figure 2 shows the classification performance of our approach and baselines with different dimensions, when N_{tr} is set to 20, on the View-1 of UCI Daily and Sports Activity dataset. It shows that PCA achieves quite stable results when the dimension is higher than 60. LPP achieves its best performance when the dimension is around 50, and CCA favors a higher dimension. Although LDA increases the accuracy significantly with more dimensions, it is limited by the maximum number of dimension which is less than the number of classes. MvDA requires about 200 dimensions to achieve its best performance. Our approach obtains good performance with only 60 dimensions, and it consistently outperforms other baselines in each case.

4.2 Multimodal Spoken Word Dataset

The Multimodal Spoken Word dataset is collected to study the speaker-dependent speech recognition problem, which helps us understand the speech translation for assistive communication. One subject is asked to speak 73 words, and each word is repeated for eight times. The speech of every word is recorded by three types of signals, including audio, video, and magnetic sensors. For audio signals, we extract 20 different features from them, such as Linear Predictive Codings (LPC) [14] and Mel-Frequency Cepstral Coefficients (MFCC) [34]. The videos capture the face of speaker during speech. We crop the mouth regions in the video, and extract the Local Binary Pattern (LBP) [1] features from each frame. Twenty-four magnetic sensors are placed on the tongue of the subject, which track the positions and movement trajectories of the tongue during speech. Clearly, all of the three modalities can be represented as multivariate time series.

Three-View Setting

A three-view experimental setting is designed on the Multimodal Spoken Word dataset. The m.t.s. of sensors, video,

Table 2: Classification Accuracy (%) on UCI Daily Activity Dataset. N_{tr} is the number of training samples randomly chosen from each activity.

Method		$N_{tr}=10$		$N_{tr}=20$		$N_{tr}=30$		$N_{tr}=40$		$N_{tr}=50$	
		View-1	View-2	View-1	View-2	View-1	View-2	View-1	View-2	View-1	View-2
Single-View	PCA [19]	27.63	21.73	31.17	23.34	32.01	24.03	33.04	24.56	33.97	24.86
	LDA [11]	31.35	14.29	38.51	13.33	42.27	14.73	42.92	15.75	44.19	16.32
	SVM [7]	22.32	20.80	21.45	18.05	21.47	17.58	21.51	17.75	21.18	18.09
	LPP [41]	27.60	21.18	39.96	30.39	48.79	36.91	55.14	42.22	59.31	46.05
	2DLDA [46]	53.37	55.24	67.59	64.70	73.15	70.55	76.09	72.13	78.93	75.79
	1NN-DTW [43]	41.05	38.53	43.67	40.33	48.92	45.26	61.55	50.18	63.91	52.70
Multi-View	CCA [17]	28.36	18.05	43.10	20.02	51.61	22.28	56.92	24.07	60.14	26.84
	MvDA [20]	56.43	57.98	75.03	74.24	81.20	74.77	80.37	76.95	85.88	81.08
	MDBP (Ours)	**70.29**	**67.93**	**82.58**	**77.31**	**87.55**	**81.81**	**89.83**	**83.42**	**91.35**	**84.45**

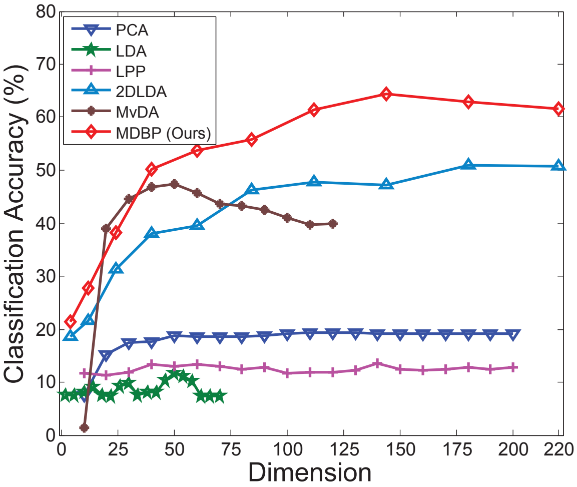


Figure 3: Classification accuracy (%) with different dimensions on View-1 of Multimodal Spoken Word dataset.

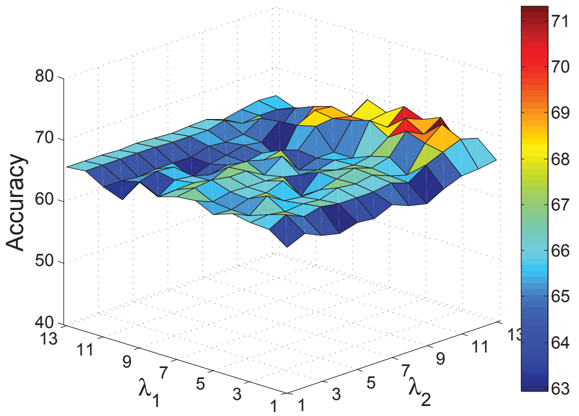


Figure 4: Parameter sensitivity of λ_1 and λ_2 used in our approach on UCI Daily and Sports Activity dataset (View-1). The indexes from 1 to 13 on x and y axis correspond to a set of parameters $\{0,\ 10^{-4},\ 5\times10^{-4},\ 10^{-3},\ 5\times10^{-3},\ 10^{-2},\ 5\times10^{-2},\ 0.1,\ 0.5,\ 1,\ 5,\ 10,\ 20\}$.

and audio are separately denoted as View-1, View-2 and View-3. We compare our approach with the baselines described in Section 4.1. The m.t.s. within each view are preprocessed to have the same length. In addition, MvDA requires that samples in different views should have the same dimension, while our approach does not have such a constraint. For MvDA, we have to perform preprocessing to make sure that samples in three views share the same dimension.

Results

We randomly choose $N_{tr} \in \{3, 4, 5\}$ samples from each word (per view) to construct the training set, and the remaining samples are used for the test. This process is repeated for 10 times. Table 3 shows the average classification accuracy of our approach and baselines in different settings. We observe that traditional subspace learning methods, such as PCA, LDA and LPP, obtain very poor performance on this dataset, due to the small-sample-size problem. Moreover, the classification task on this dataset is more challenging than that on the UCI Daily and Sports Activity dataset, as there are more classes. 2DLDA keeps the temporal structure of raw m.t.s., and therefore it outperforms other single-view methods. MvDA obtains poor performance on View-2 and View-3, due to the following reasons: (1) MvDA constructs joint scatter matrices across different views, which works well on multi-view data with similar types of features in each view, such as the UCI dataset used in Section 4.1. However, the Multimodal Spoken Word dataset contains three different types of signals, which can hardly be characterized by a joint scatter matrix. (2) MvDA requires that samples in different views should have the same dimension, which results in certain information loss. (3) MvDA breaks the temporal structure by vectorizing the m.t.s. samples. Table 3 shows that our approach achieves consistently better results than baselines on all the three views.

Figure 3 shows the accuracy of our approach and baselines with different dimensions when N_{tr} is set to 3. Our approach obtains higher accuracy than other baselines in most cases.

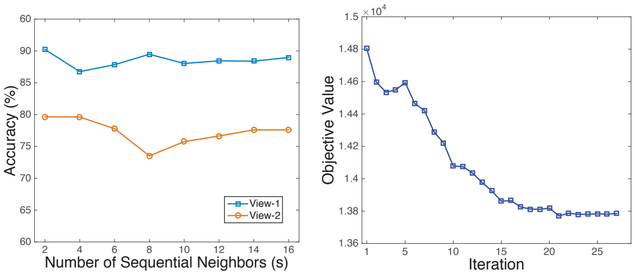
4.3 Discussions

Parameter Sensitivity and Convergence

There are three major parameters in our approach, including λ_1, λ_2 and s. The first two balance the effects of discriminative regularization and temporal smoothness regularization, and parameter s denotes the number of sequential neighbors used to construct the Laplacian matrix. Fig-

Table 3: Classification accuracy (%) on Multimodal Spoken Word dataset. N_{tr} is the number of training samples randomly chosen from each word.

Method		$N_{tr} = 3$			$N_{tr} = 4$			$N_{tr} = 5$		
		View-1	View-2	View-3	View-1	View-2	View-3	View-1	View-2	View-3
Single-View	PCA [19]	17.73	17.10	12.47	18.49	17.67	13.42	19.36	17.58	13.70
	LDA [11]	11.86	13.78	13.97	11.61	14.66	12.88	11.74	15.34	14.06
	SVM [7]	14.38	21.26	11.37	14.04	22.29	11.88	12.60	22.28	11.74
	LPP [41]	14.71	13.01	12.93	16.20	13.80	13.01	16.55	13.79	12.51
	2DLDA [46]	50.08	64.66	21.15	55.27	69.04	37.36	62.83	71.69	50.55
	1NN-DTW [43]	53.71	65.29	25.45	59.59	58.90	38.47	65.20	72.05	52.33
Multi-View	MvDA [20]	49.73	39.97	18.75	49.93	38.15	23.20	44.02	32.33	21.25
	MDBP (Ours)	**66.44**	**69.01**	**39.51**	**70.24**	**76.10**	**41.08**	**73.01**	**78.36**	**61.14**



(a) Sensitivity of s

(b) Convergence curve

Figure 5: (a) Parameter sensitivity of s and (b) convergence curve of our approach on UCI Daily and Sports Activity dataset.

ure 4 shows the sensitivity of λ_1 and λ_2 on the UCI Daily and Sports Activity dataset. We have the following observations: (1) By setting either λ_1 or λ_2 to 0 (i.e., removing the regularization terms in Eq. (8)), the accuracy of our approach drops significantly. It validates the effectiveness of incorporating discriminative and temporal information into our approach. (2) Our approach obtains relatively stable performance with the settings $\lambda_1 \in [5 \times 10^{-4},\ 1]$ and $\lambda_2 \in [1,\ 20]$. Figure 5(a) shows the sensitivity of parameter s. It shows that our approach is not very sensitive to the setting of s, and $s = 2$ usually leads to a better performance.

Figure 5(b) shows the converge curve of our approach on the UCI Daily and Sports Activity dataset. Our approach quickly converges with only 25 iterations, which makes it efficient for large-scale applications.

Experiments with Data Fusion and Feature Fusion

In the above experiments, we assume that the test m.t.s. is only available in one view, as shown in Table 2 and Table 3. In practice, however, test m.t.s. might be available in multiple views. For single-view methods, strategies like data fusion and feature fusion can be applied to generate a final prediction of class label. Multi-view methods can adopt the feature fusion strategy. In data fusion, a m.t.s. observed from multiple views are first vectorized, and then concatenated to a long vector. In feature fusion, the compact features are extracted from each view first, and then those feature vectors can be combined.

Table 4 shows the accuracy of our approach and baselines using one or two available fusion strategies on the UCI Daily and Sports Activity dataset. Comparing Table 4 and Table 2, we observe that the accuracies of PCA, SVM, 2DLDA,

Table 4: Fusion results on UCI Daily and Sports Activity dataset when $N_{tr} = 10$.

Method		Data Fusion	Feature Fusion
Single-View	PCA [19]	31.91	30.76
	LDA [11]	18.01	24.40
	SVM [7]	31.31	31.41
	LPP [41]	9.90	23.15
	2DLDA [46]	57.62	58.30
Multi-View	CCA [17]	-	25.90
	MvDA [20]	-	67.24
	Ours	-	**78.96**

and MvDA can be improved by fusing data or features. LPP obtains better performance with the feature fusion strategy. However, LDA cannot take advantages of the fusion strategies, due to the performance gap between View-1 and View-2. Our approach improves the classification accuracy by at least 10% with the feature fusion strategy. It indicates that the features extracted from two views have complementary information that are useful for m.t.s. classification.

5. RELATED WORK

In this section, we briefly introduce three research topics that are very related to our approach, including subspace learning, multivariate time series classification, and multiview learning.

5.1 Subspace Learning

Subspace learning is an effective technique in extracting informative features from data, which reduces the dimensionality of data through linear or nonlinear projections. According to the availability of class labels, subspace learning methods can be mainly divided into three groups: unsupervised methods, semi-supervised methods, and supervised methods. The unsupervised methods only utilize unlabeled data [19], semi-supervised methods make use of the partial labeled data [4], and supervised methods learn subspaces using the fully labeled data [11, 18, 25]. The representative unsupervised methods include principal component analysis (PCA) [19] and locality preserving projections (LPP) [16]. PCA projects data into a low-dimensional subspace by maximizing the variance of data. LPP is a non-parametric method, which preserves the neighborhood structure of samples on manifold. Linear discriminant analysis (LDA) [11] is a classical supervised subspace learning method, which maximizes the between-class scatter and minimize the within-class scatter of projected samples. In addition, traditional methods require the vectorized data as input, while some advanced methods learn bilinear projections that directly process high-order data (e.g., images or EEG signals) without vectorization [9, 6].

The major differences between our approach and existing subspace learning methods are: (1) our approach deals with the multi-view time series data by modeling view consistency and temporal regularization; (2) our approach incorporates a novel supervised regularization for classification tasks.

5.2 Multivariate Time Series Classification

The classification of multivariate time series (m.t.s.) has become a popular research topic in recent years [24, 36, 50]. Many effective algorithms have been presented to solve this problem, which can be roughly categorized into three groups: (1) distance metric; (2) classifier design; (3) dimensionality reduction. The first group of methods focus on designing distance metrics for m.t.s., by considering the temporal dynamics and the possible misalignment problem in m.t.s. [36]. The classifier design methods usually adapt the effective classifiers from other domains to m.t.s. classification, such as SVM [50], recurrent probabilistic neural network [15], convolutional nonlinear component analysis [49], etc.. The dimensionality reduction methods project high-dimensional m.t.s. to a low-dimensional subspace by satisfying certain criteria. Weng *et al.* employed two-dimensional singular value decomposition [42] and locality preserving projections [41] for m.t.s. classification. Li *et al.* designed a common principal component analysis method for m.t.s. classification. Other interesting explorations on m.t.s. classification include feature selection [33], temporal abstraction [37], and tensor factorization [5].

Different from all of the existing multivariate time series classification methods, our approach focus on the classification of multi-view multivariate time series data, which hasn't been explored before.

5.3 Multi-View Learning

Multi-view learning has been receiving increasing attention in recent years. One implicit assumption is that either view alone has sufficient information about the samples, but the learning complexity can be reduced by eliminating hypotheses from each view if different views contribute diverse information [45]. Multi-view learning has been widely applied to many problems, such as clustering [13, 30], classification [40, 20], semi-supervised learning [12, 22], person re-identification [27], and outlier detection [28].

Projecting data collected from multiple views onto a shared subspace is considered as an effective strategy in multi-view learning. The classical method, canonical correlation analysis(CCA) [17], projects two sets of observations onto a subspace by maximizing their correlations, which has been extended to multiple views [38]. Most recently, Ding *et al.* incorporated low-rank constraints in learning common subspace for multi-view data. Kan *et al.* extended the linear discriminant analysis method to multi-view setting [20] and obtained impressive performance on image classification.

Existing multi-view learning algorithms do not take the temporal information into account, which are not suitable for m.t.s. classification. Our MDBP approach incorporates a temporal smoothness regularization, and its effectiveness has been validated by extensive experiments.

6. CONCLUSIONS

In this paper, we propose a multi-view bilinear projection approach named MDBP for classifying m.t.s. that are collected from multiple views. MDBP projects multi-view data to a shared subspace through view-specific bilinear projections that preserve the temporal structure of m.t.s., and learns discriminative features by incorporating a novel supervised regularization. The temporal smoothness is also modeled in MDBP, with the help of Laplacian regularization. An efficient optimization algorithm based on gradient descent is designed to solve the problem. We conduct extensive experiments on a daily activity benchmark dataset and a recently collected multimodal spoken word dataset. Experimental results show that our approach obtains remarkable improvements over the state-of-the-art multi-view learning and multivariate time-series classification methods. The parameter sensitivity, convergence property and multi-view fusion are also evaluated and discussed. In our future work, we will develop an online version of MDBP to deal with multi-view m.t.s. in a real-time fashion.

7. ACKNOWLEDGMENTS

This work is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

8. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100. ACM, 1998.

[4] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–7. IEEE, 2007.

[5] Y. Cai, H. Tong, W. Fan, P. Ji, and Q. He. Facets: Fast comprehensive mining of coevolving high-order time series. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 79–88. ACM, 2015.

[6] C. Christoforou, R. Haralick, P. Sajda, and L. C. Parra. Second-order bilinear discriminant analysis. *The Journal of Machine Learning Research*, 11:665–685, 2010.

[7] C. Cortes and V. Vapnik. Support vector machine. *Machine Learning*, 20(3):273–297, 1995.

[8] Z. Ding and Y. Fu. Low-rank common subspace for multi-view learning. In *IEEE International Conference on Data Mining*, pages 110–119. IEEE, 2014.

[9] M. Dyrholm, C. Christoforou, and L. C. Parra. Bilinear discriminant component analysis. *The Journal of Machine Learning Research*, 8:1097–1111, 2007.

[10] Z. Fang and Z. Zhang. Simultaneously combining multi-view multi-label learning with maximum margin classification. In *Proceedings of IEEE International Conference on Data Mining*, pages 864–869. IEEE, 2012.

[11] R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8(4):376–386, 1938.

[12] S. Günnemann, I. Färber, M. Rüdiger, and T. Seidl. Smvc: semi-supervised multi-view clustering in subspace

projections. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 253–262. ACM, 2014.

[13] Y. Guo. Convex subspace representation learning from multi-view data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, volume 1, page 2, 2013.

[14] A. Harma and U. K. Laine. A comparison of warped and conventional linear predictive coding. *IEEE Transactions on Speech and Audio Processing*, 9(5):579–588, 2001.

[15] H. Hayashi, T. Shibanoki, K. Shima, Y. Kurita, and T. Tsuji. A recurrent probabilistic neural network with dimensionality reduction based on time-series discriminant component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(12):3021–3033, 2015.

[16] X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, pages 153–160, 2004.

[17] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[18] X.-Y. Jing, S. Li, D. Zhang, J. Yang, and J.-Y. Yang. Supervised and unsupervised parallel subspace learning for large-scale image recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10):1497–1511, 2012.

[19] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.

[21] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[22] C. Lan and J. Huan. Reducing the unlabeled sample complexity of semi-supervised multi-view learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 627–634. ACM, 2015.

[23] H. Li. Accurate and efficient classification based on common principal components analysis for multivariate time series. *Neurocomputing*, 171:744–753, 2016.

[24] K. Li, S. Li, and Y. Fu. Early classification of ongoing observation. In *IEEE International Conference on Data Mining*, pages 310–319. IEEE, 2014.

[25] S. Li and Y. Fu. Learning robust and discriminative subspace with low-rank constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2015.

[26] S. Li, K. Li, and Y. Fu. Temporal subspace clustering for human motion segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4461, 2015.

[27] S. Li, M. Shao, and Y. Fu. Cross-view projective dictionary learning for person re-identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2155–2161, 2015.

[28] S. Li, M. Shao, and Y. Fu. Multi-view low-rank analysis for outlier detection. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2015.

[29] T. Li, V. Sindhwani, C. H. Ding, and Y. Zhang. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *Proceedings of the SIAM International Conference on Data Mining*, pages 293–302. SIAM, 2010.

[30] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2750–2756, 2015.

[31] M. Lichman. UCI machine learning repository, 2013.

[32] J. Lines and A. Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015.

[33] R. Liu, S. Xu, C. Fang, Y.-w. Liu, Y. L. Murphey, and D. S. Kochhar. Statistical modeling and signal selection in multivariate time series pattern classification. In *The 21st International Conference on Pattern Recognition*, pages 2853–2856. IEEE, 2012.

[34] B. Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[35] P.-F. Marteau and S. Gibet. On recursive edit distance kernels with application to time series classification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6):1121–1133, 2015.

[36] J. Mei, M. Liu, Y. Wang, and H. Gao. Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. *IEEE transactions on Cybernetics*, 2015.

[37] R. Moskovitch and Y. Shahar. Classification of multivariate time series via temporal abstraction and time intervals mining. *Knowledge and Information Systems*, 45(1):35–74, 2015.

[38] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses*, pages 1–4, 2010.

[39] D. F. Silva, V. De Souza, and G. E. Batista. Time series classification using compression distance of recurrence plots. In *IEEE 13th International Conference on Data Mining*, pages 687–696. IEEE, 2013.

[40] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1083–1092, 2015.

[41] X. Weng and J. Shen. Classification of multivariate time series using locality preserving projections. *Knowledge-Based Systems*, 21(7):581–587, 2008.

[42] X. Weng and J. Shen. Classification of multivariate time series using two-dimensional singular value decomposition. *Knowledge-Based Systems*, 21(7):535–539, 2008.

[43] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1033–1040, 2006.

[44] Z. Xing, J. Pei, and S. Y. Philip. Early classification on time series. *Knowledge and information systems*, 31(1):105–127, 2012.

[45] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[46] J. Ye, R. Janardan, and Q. Li. Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 1569–1576, 2004.

[47] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 947–956. ACM, 2009.

[48] Z. Zhang, J. Cheng, J. Li, W. Bian, and D. Tao. Segment-based features for time series classification. *The Computer Journal*, 55(9):1088–1102, 2012.

[49] Y. Zheng, Q. Liu, E. Chen, J. L. Zhao, L. He, and G. Lv. Convolutional nonlinear neighbourhood components analysis for time series classification. In *Advances in Knowledge Discovery and Data Mining*, pages 534–546. Springer, 2015.

[50] P.-Y. Zhou and K. C. Chan. A feature extraction method for multivariate time series classification using temporal patterns. In *Advances in Knowledge Discovery and Data Mining*, pages 409–421. Springer, 2015.