FISH-MML: Fisher-HSIC Multi-View Metric Learning

Changqing Zhang[1], **Yeqing Liu**[1], **Yue Liu**[1], **Qinghua Hu**[1*], **Xinwang Liu**[2] and **Pengfei Zhu**[1]
[1]School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]School of Computer, National University of Defense Technology, Changsha, China
{zhangchangqing, yeqing, liuyue76, huqinghua, zhupengfei}@tju.edu.cn, 1022xinwang.liu@gmail.com

Abstract

This work presents a simple yet effective model for multi-view metric learning, which aims to improve the classification of data with multiple views, e.g., multiple modalities or multiple types of features. The intrinsic correlation, different views describing same set of instances, makes it possible and necessary to jointly learn multiple metrics of different views, accordingly, we propose a multi-view metric learning method based on Fisher discriminant analysis (FDA) and Hilbert-Schmidt Independence Criteria (HSIC), termed as Fisher-HSIC Multi-View Metric Learning (FISH-MML). In our approach, the class separability is enforced in the spirit of FDA within each single view, while the consistence among different views is enhanced based on HSIC. Accordingly, both *intra-view* class separability and *inter-view* correlation are well addressed in a unified framework. The learned metrics can improve multi-view classification, and experimental results on real-world datasets demonstrate the effectiveness of the proposed method.

1 Introduction

With the rapid development of information acquirement technique, data are usually represented with different modalities or different types of features. In computer vision, images are often depicted with different types of descriptors based on color or texture cues; RGB-D images are considered as multimodal data consisting of depth and color information. For social network analysis (SNA), different relationships usually characterize the same set of users, and different types of attributes or textual information are often associated with those users, e.g., user-generated content or demographic details. Recently, there are intensive attentions on developing classification or clustering models for the data with multiple views, and the effectiveness has been empirically proven on diverse applications [Kumar *et al.*, 2011; Wang *et al.*, 2016; Gong, 2017; Zhao *et al.*, 2017; Cao *et al.*, 2015; Zhang *et al.*, 2015; 2017].

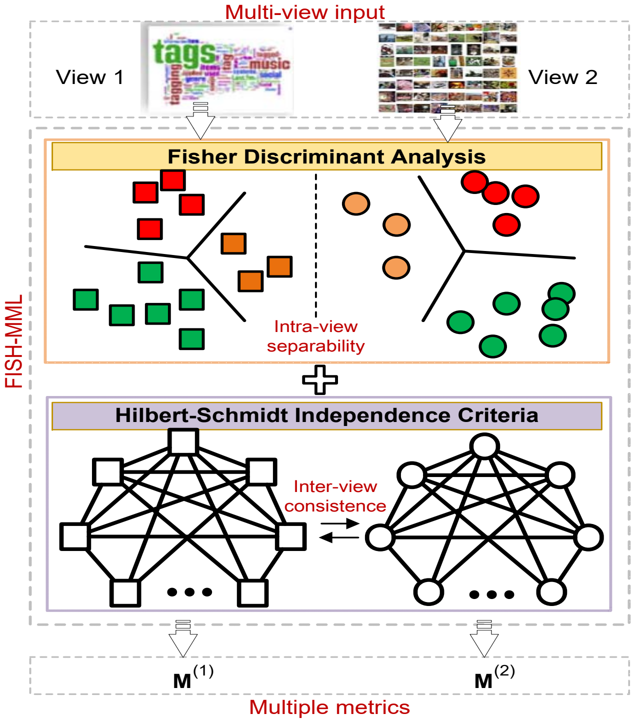


Figure 1: Illustration of Fisher-HSIC Multi-View Metric Learning. The proposed model enforces separability within each view using label information, and simultaneously respects consistence across different views.

As recognized by recent researches [Sindhwani and Rosenberg, 2008; Dhillon *et al.*, 2011; Liu *et al.*, 2016], the main challenge of exploiting multi-view data lies in how to effectively explore the underlying correlation among different views. It is nontrivial since although different views indeed depict same set of instances, the feature space of one view can be completely different from that in another view, which is known as heterogeneous features. For example, Euclidean distances are typically employed as the distance measure for HOG [Dalal and Triggs, 2005], while spatial matching kernels are widely used for the local descriptors SIFT [Lowe, 1999]. Moreover, features of one view can be high-dimensional while another one may be not and fea-

[*]Corresponding author: Qinghua Hu.

tures of one view may be much more noisy than another view. These challenges increase the difficulty of integrating different views. The representative and straightforward strategy is feature integration, i.e., directly concatenating all views into high-dimensional vectors or performing dimensionality reduction jointly (such as Canonical Correlation Analysis (CCA)). However, direct combination of feature vectors or simple linear combination of the outputs of different views can not guarantee promising performance, and CCA-based methods [Blaschko and Lampert, 2008; Chaudhuri *et al.*, 2009] only explore linear correlations thus neglect complex correlations in real applications.

Metric learning can learn a distance function to well reflect the relationships between data points consistent with their semantic labels, which could benefit to subsequent tasks, e.g., classification or clustering. Generally, metric learning approaches seek a Mahalanobis distance with the paired samples or other side information encoding relationships of data. Compared with the Euclidean distance in original feature space, the learned distance metric could better reveal the relationships between data points. Due to its effectiveness, extensive metric learning methods [Weinberger and Saul, 2009; Davis *et al.*, 2007; Guillaumin *et al.*, 2009] have been proposed and widely applied in real-world applications.

In this paper, we propose to learn distance metric for data with multiple views, i.e., jointly learning multiple metrics to explore complementary information across multiple views. Towards this goal, we propose *Fisher-HSIC Multi-View Metric Learning* (FISH-MML) algorithm. On the one hand, we introduce Fisher discriminant analysis (FDA) to search for the optimal projections (corresponding to Mahalanobis distance metrics) to maximize class separability and preserve expressiveness within each view, which alleviates the difficulty of classification compared with original features (corresponding to Euclidean distance metrics). On the other hand, to explore the complementarity from different views, our model maximizes the dependence among different views with Hilbert-Schmidt Independence Criterion (HSIC), which ensures between-data relationships (under the learned metrics) of different views to be consistent in kernel space. The proposed approach is effectively optimized by using the Alternating Direction Minimization (ADM) strategy, and extensive experiments validate the effectiveness of our approach.

The highlights of this paper are summarized as follows:
(1) We propose a novel multi-view metric learning method, which is simple yet rather effective.
(2) Our model simultaneously enhances class separability within each view and explores complex correlation across multiple views in a unified framework.
(3) With FDA, class separability within each view is enhanced, while by using HSIC, our method can effectively explore correlations among different views.
(4) Based on Alternating Direction Method (ADM), our objective is efficiently optimized with guaranteed convergence.
(5) Experiments on real-world multi-view datasets validate the effectiveness of our method for classification.

2 Related Work

There have been quite a few distance metric learning approaches. The early work in [Xing *et al.*, 2003] learns distance metric with side information that indicates two data samples being similar or dissimilar. This method is formulated as a convex optimization problem. The work in [Davis *et al.*, 2007] introduces information theory and formulates the task as minimizing the differential relative entropy between two multivariate Gaussians under constraints on distance function. LMNN (Large Margin Nearest Neighbors) [Weinberger and Saul, 2009] learns a metric for k-Nearest Neighbor (kNN) classifier to enforce that k-nearest neighbors belong to the same class while examples from different classes are separated by a large margin. There are also some methods that learn distance metrics under sparsity [Ying *et al.*, 2009] or low-rank [Ding *et al.*, 2015] assumptions.

Due to the ubiquitousness of data with multiple modalities or descriptors, the literature of multi-view learning has spanned a very broad range. In metric learning domain, the method HMML (Heterogeneous Multi-Metric Learning) [Zhang *et al.*, 2011] proposes to jointly learn a set of heterogeneous metrics for multi-sensor data fusion, which generalizes LMNN [Weinberger and Saul, 2009] from single-view to multi-view learning. The work in [Xie and Xing, 2013] proposes a general framework of multi-modal distance metric learning based on multi-wing harmonium model, in which different modalities are embedded into a shared latent space. The researchers also propose a large margin multi-metric learning (LM3L) [Hu *et al.*, 2014] method for face and kinship verification. Recently, deep model-based metric learning methods [Hu *et al.*, 2017; Lu *et al.*, 2015] have been proposed. The sharable and individual multi-view deep metric learning (MvDML) approach [Hu *et al.*, 2017] jointly learns multiple distance metrics for multi-view data by seeking an individual distance metric for each view and a common representation for different views in a unified latent subspace.

3 Background

For notations used throughout this paper, boldface uppercase, boldface lowercase, and normal italic letters are utilized to denote matrix, vector, and scalar respectively. We denote feature matrix as $\mathbf{X} \in \mathbb{R}^{d \times n}$, where d and n are dimensionality of feature space and number of samples, respectively. $\mathbf{x}_i$ is the feature vector of the i^{th} samples. For data represented by V different views, we use $\mathbb{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{d_v \times n}\}_{v=1}^{V}$ to denote the set of feature matrices of multiple views with d_v being the dimensionality of the feature space for the v^{th} view.

Similar to traditional metric learning algorithms, in our model, we also focus on learning the Mahalanobis distance. In contrast, our model jointly learns these multiple Mahalanobis distances of multiple views. Generally, the Mahalanobis distance has the following definition:

Definition 3.1. (Mahalanobis distance). The Mahalanobis distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathbf{M}} = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j),$$

where the Mahalanobis matrix $\mathbf{M}$ is constrained to be symmetric positive-definite to assure the validity. □

4 Fisher-HSIC Multi-View Metric Learning

Firstly, we try to propose a general framework for multi-view metric learning (MML) that can jointly learn multiple metrics for multiple views. The general form of multi-view metric learning is given as follows:

$$\max_{\{\mathbf{M}^{(v)}\}_{v=1}^{V}} \underbrace{\mathcal{S}(\{\mathbf{M}^{(v)}\}_{v=1}^{V})}_{\text{class separability}} + \lambda \underbrace{\mathcal{C}(\{\mathbf{M}^{(v)}\}_{v=1}^{V})}_{\text{view consistence}}, \tag{1}$$

where $\mathbf{M}^{(v)}$ is the distance metric of the v^{th} view, and $\lambda > 0$ is a tradeoff hyperparameter. The above objective function searches the optimal metrics that can simultaneously maximize the class separability (with $\mathcal{S}(\cdot)$) and penalize the disagreement between different views (with $\mathcal{C}(\cdot)$).

Separability and Expressiveness

To ensure the class separability within each view, FDA is introduced into our model, which is based on the definitions of between-class and total scatter matrices:

$$\begin{aligned} \mathbf{S}_b &= \frac{1}{n}\sum_{j=1}^{g} n_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T, \\ \mathbf{S}_t &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T, \\ \boldsymbol{\mu}_j &= \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_i^j,\ \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \end{aligned} \tag{2}$$

where $\mathbf{x}_i^j$ denotes the feature vector of the i^{th} sample in class $\mathcal{C}_j$, $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}$ are sample means for class $\mathcal{C}_j$ and the whole data set, respectively. g and n_j are the number of classes and the number of samples belonging to class $\mathcal{C}_j$, respectively.

On the one hand, we note that Euclidean distance is involved in Eq.(2) which is used in FDA. This could be improved with the Mahalanobis distance, i.e., $(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu}) \rightarrow (\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{M}(\mathbf{x}_i - \boldsymbol{\mu})$. On the other hand, when $\mathbf{M}$ is a symmetric positive-definite matrix, $d_{\mathbf{M}}$ is a metric. Specifically, as any symmetric positive semi-definite matrix $\mathbf{M} \in S_+^d$ can be decomposed as $\mathbf{M} = \mathbf{P}^T\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{k\times d}$ and $k \geq rank(\mathbf{M})$. Accordingly, the distance metric function can be rewritten as $d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{P}^T\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j) = ||\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j)||_2^2$, which corresponds to the Euclidean distance between projected feature vectors. Therefore, with respect to the new space (induced by $\mathbf{P}^{(v)}$), the between-class and total scatter matrices in the v^{th} view, i.e., $\mathbf{S}_b^{(v)}$ and $\mathbf{S}_t^{(v)}$, are induced as follows:

$$\begin{aligned} \mathbf{S}_b^{(v)} &= \frac{1}{n}\sum_{j=1}^{g} n_j(\boldsymbol{\mu}_j^{(v)} - \boldsymbol{\mu}^{(v)})(\boldsymbol{\mu}_j^{(v)} - \boldsymbol{\mu}^{(v)})^T, \\ \mathbf{S}_t^{(v)} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i^{(v)} - \boldsymbol{\mu}^{(v)})(\mathbf{z}_i^{(v)} - \boldsymbol{\mu}^{(v)})^T, \\ \boldsymbol{\mu}_j^{(v)} &= \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{z}_i^{j(v)}, \boldsymbol{\mu}^{(v)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i^{(v)}, \end{aligned} \tag{3}$$

where $\mathbf{z}^{(v)} = \mathbf{P}^{(v)}\mathbf{x}^{(v)}$ with $\mathbf{z}^{(v)}$ being the projected feature vector corresponding to $\mathbf{x}^{(v)}$. $\boldsymbol{\mu}_j^{(v)}$ and $\boldsymbol{\mu}^{(v)}$ are sample means of the v^{th} view for class $\mathcal{C}_j$ and the whole data set, respectively.

Then we aim to maximize the class separability within each view for the most discriminative capability parameterized by the projections $\{\mathbf{P}^{(v)}\}_{v=1}^V$. In the spirit of FDA, we need to optimize the following unconstrained optimization objective function:

$$\max_{\{\mathbf{P}^{(v)}\}_{v=1}^{V}} \sum_{v=1}^{V} Tr(\mathbf{S}_b^{(v)}; \mathbf{P}^{(v)}) - \gamma Tr(\mathbf{S}_t^{(v)}; \mathbf{P}^{(v)}), \tag{4}$$

where $Tr(\cdot)$ is the matrix trace operator and γ is a tunable parameter to balance the two terms involved. $Tr(\mathbf{S}_b^{(v)}; \mathbf{P}^{(v)})$ denotes the trace operator for $\mathbf{S}_b^{(v)}$ conditioned on $\mathbf{P}^{(v)}$.

Note that, we do not constrain γ in Eq.(4) to be positive and there are different meanings for positive and negative cases. With $\gamma > 0$, the objective function maximizes the interclass scattering while simultaneously minimizes the total scattering and thus, the intra-class scattering is automatically minimized. However, as recognized by [Cheng *et al.*, 2011], it may be sensitive to spurious features of the data in high-dimensional case. This inspired us to set $\gamma < 0$ to pursuit the expressiveness. Specifically, in this manner, it actually maximizes not only discriminativeness but also expressiveness jointly. Recall the objective function of PCA (Principal Component Analysis), i.e.,

$$\begin{aligned} &\max_{\mathbf{v}} \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{v}^T\mathbf{x}_i - \mathbf{v}^T\boldsymbol{\mu}\}^2 = \mathbf{v}^T\mathbf{S}\mathbf{v} \\ &\text{with } \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T, \end{aligned} \tag{5}$$

where we take the first component as example for simplicity, and the constraint $\mathbf{v}^T\mathbf{v} = 1$ indicates that we are only interested in the direction instead of its magnitude. Similar to PCA which maximizes the above objective function, we maximize $Tr(\mathbf{S}_t^{(v)}; \mathbf{P}^{(v)})$ with respect to $\mathbf{P}^{(v)}$ to account for expressiveness.

Consistence across Multiple Views

The above objective function focuses on seeking metrics that jointly maximize discriminativeness and expressiveness. Our model is devoted to handle data with multiple views where complementarity is critical, hence now we try to explore complementary information from multiple views by using Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005]. To determine complex associations of two signals, HSIC has been theoretically [Gretton *et al.*, 2005] and empirically [Xiao and Guo, 2015; Song *et al.*, 2007] justified to be a proper measure of (in)dependence when associated with a universal kernel.

Letting the observations $\mathbf{Z}^{(v)}$ and $\mathbf{Z}^{(w)}$ (corresponding to two different views) contain n data points $\{(\mathbf{z}_i^{(v)}, \mathbf{z}_i^{(w)}) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$ that are jointly drawn from a probability distribution $P_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}$, the consistence between two views is measured by the dependence between $\mathbf{z}^{(v)}$ and $\mathbf{z}^{(w)}$. The dependence measured by HSIC is computed according to the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$

in Hilbert space. A large HSIC value indicates strong dependence with respect to the choice of kernels. The HSIC is defined as HSIC$(P_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}, \mathcal{F}, \mathcal{G}) := ||C_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}||_{\text{HS}}^2$, where $||\mathbf{A}||_{\text{HS}} = \sqrt{\sum_{i,j} a_{ij}^2}$. $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert Space (RKHS) on $\mathcal{X}$ and $\mathcal{Y}$, respectively. The cross-covariance is a function that gives the covariance of two random variables and defined as $C_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}} = \mathbb{E}_{\mathbf{z}^{(v)}\mathbf{z}^{(w)}}[(\phi(\mathbf{z}^{(v)}) - \mu_{\mathbf{z}^{(v)}}) \otimes (\varphi(\mathbf{z}^{(w)}) - \mu_{\mathbf{z}^{(w)}})]$, where $\mu_{\mathbf{z}^{(v)}} = \mathbb{E}(\phi(\mathbf{z}^{(v)}))$, $\mu_{\mathbf{z}^{(w)}} = \mathbb{E}(\varphi(\mathbf{z}^{(w)}))$, and $\otimes$ is the tensor product. $\phi(\mathbf{z}^{(v)})$ and $\varphi(\mathbf{z}^{(w)})$ are functions that map $\mathbf{z}^{(v)} \in \mathcal{X}$ and $\mathbf{z}^{(w)} \in \mathcal{Y}$ to kernel space $\mathcal{F}$ and $\mathcal{G}$ with respect to the kernel functions $k_v(\mathbf{z}_i^{(v)}, \mathbf{z}_j^{(v)}) =< \phi(\mathbf{z}_i^{(v)}), \phi(\mathbf{z}_j^{(v)}) >$ and $k_w(\mathbf{z}_i^{(w)}, \mathbf{z}_j^{(w)}) =< \varphi(\mathbf{z}_i^{(w)}), \varphi(\mathbf{z}_j^{(w)}) >$. Accordingly, we have the empirical HSIC defined as:

$$\text{HSIC}(\mathbf{Z}^{(v)}, \mathbf{Z}^{(w)}) = (n-1)^{-2} tr(\mathbf{K}_v \mathbf{H} \mathbf{K}_w \mathbf{H}), \quad (6)$$

where $\mathbf{K}_v$ and $\mathbf{K}_w$ are the Gram matrices with $k_{v,ij} = k_v(\mathbf{z}_i^{(v)}, \mathbf{z}_j^{(v)})$, $k_{w,ij} = k_w(\mathbf{z}_i^{(w)}, \mathbf{z}_j^{(w)})$. $h_{ij} = \delta_{ij} - 1/n$ centers the Gram matrix to have zero mean in the feature space. In our implementation, we use the inner product kernel function, i.e., $\mathbf{K}^{(v)} = \mathbf{Z}^{(v)^T}\mathbf{Z}^{(v)} = \mathbf{X}^{(v)^T}\mathbf{P}^{(v)^T}\mathbf{P}^{(v)}\mathbf{X}^{(v)}$, and promising performance are achieved. Note that maximizing HSIC$(\mathbf{Z}^{(v)}, \mathbf{Z}^{(w)})$ enhances the dependency between $\mathbf{K}^{(v)}$ and $\mathbf{K}^{(w)}$, which penalties the disagreement between kernel matrices from different views parameterized by the projections $\mathbf{P}^{(v)}$ and $\mathbf{P}^{(w)}$.

Objective Function

For multi-view metric learning, we jointly enhance intra-view separability, expressiveness, and inter-view correlations with respect to the learned metrics in a unified objective function:

$$\begin{aligned} \max_{\{\mathbf{P}^{(v)}\}_{v=1}^V} & \sum_{v=1}^V tr(\mathbf{S}_b{}^{(v)}; \mathbf{P}^{(v)}) + \lambda_1 \sum_{v=1}^V tr(\mathbf{S}_t{}^{(v)}; \mathbf{P}^{(v)}) \\ & + \lambda_2 \textstyle\sum_{v \neq w} \text{HSIC}(\mathbf{P}^{(v)}\mathbf{X}^{(v)}, \mathbf{P}^{(w)}\mathbf{X}^{(w)}) \\ s.t.\ & \mathbf{P}^{(v)}\mathbf{P}^{(v)^T} = \mathbf{I},\ v = 1, ..., V, \end{aligned} \quad (7)$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters encoding the belief degrees for expressiveness and inter-view consistence, respectively. Note that we impose orthogonal constraints on $\mathbf{P}^{(v)}$ (i.e., $\mathbf{P}^{(v)}\mathbf{P}^{(v)^T} = \mathbf{I}$) for the following reasons: first, it can address the scale issue, since without this constraint, the values of $\mathbf{P}^{(v)}$ will be arbitrarily large to maximize the objective; second, it is consistent with the requirement of expressiveness in PCA (see Eq.(5)); last but not the least, it also provides convenience for optimization which will be discussed later. Our objective function can be rewritten as follows:

$$\begin{aligned} & \max_{\{\mathbf{P}^{(v)}\}_{v=1}^V} \sum_{v=1}^V tr\big(\mathbf{P}^{(v)}(\mathbf{A} + \lambda_1\mathbf{B} + \lambda_2\mathbf{C})\mathbf{P}^{(v)^T}\big) \\ = & \max_{\{\mathbf{P}^{(v)}\}_{v=1}^V} \sum_{v=1}^V tr\big(\mathbf{P}^{(v)}\mathbf{D}\mathbf{P}^{(v)^T}\big) \end{aligned} \quad (8)$$

with

$$\begin{aligned} \mathbf{A} &= \frac{1}{n}\sum_{j=1}^{g} n_j (\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_i^{j(v)} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{(v)}) \\ &\qquad (\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_i^{j(v)} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{(v)})^T, \\ \mathbf{B} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^{(v)} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{(v)})(\mathbf{x}_i^{(v)} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{(v)})^T, \\ \mathbf{C} &= \sum_{w=1;w\neq v}^{V} \mathbf{X}^{(v)}\mathbf{H}\mathbf{K}^{(w)}\mathbf{H}\mathbf{X}^{(v)^T}, \\ \mathbf{D} &= \mathbf{A} + \lambda_1\mathbf{B} + \lambda_2\mathbf{C}. \end{aligned} \quad (9)$$

The discriminativeness, expressiveness, and consistence are accounted by $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively. Given the condition $\mathbf{P}^{(v)}\mathbf{P}^{(v)^T} = \mathbf{I}$ and with variables of the other views fixed, updating $\mathbf{P}^{(v)}$ is an eigenvalue decomposition task which could be efficiently solved. Once $\mathbf{P}^{(v)}$s are learned, we can get $\mathbf{M}^{(v)}$ by $\mathbf{M}^{(v)} = \mathbf{P}^{(v)^T}\mathbf{P}^{(v)}$. Then the multi-view data $X_i = \{\mathbf{x}_i{}^{(1)}, \ldots, \mathbf{x}_i{}^{(V)}\}$ can be projected by $\mathbf{P}^{(v)}$s and transformed into $\hat{X}_i = \{\mathbf{P}^{(1)}\mathbf{x}_i{}^{(1)}, \ldots, \mathbf{P}^{(V)}\mathbf{x}_i{}^{(V)}\}$. With concatenation of all the projected feature vectors, existing classification methods (e.g., kNN) could be employed.

To summarize, our approach has the following merits: (1) our model is simple yet effective for multi-view metric learning; (2) our model can jointly learn multiple metrics by simultaneously enforcing separability, expressiveness, and exploring complex correlations among different views; (3) both intra-view relationships of data points and inter-view correlations of different views are addressed seamlessly in a unified framework; (4) our approach is solved efficiently with the alternating direction method (ADM), and since the value of our objective function is non-decreasing with iterations, the algorithm is guaranteed to converge.

5 Experiments

We conduct experiments on four real-world datasets and compare our FISH-MML with existing state-of-the-art methods in terms of diverse evaluation measures.

5.1 Setting

The datasets employed are as follows:

•**handwritten**[1] contains 2000 images of 10 classes from number 0 to 9. There are 6 types of descriptors exacted: Pix (view1), Fou (view2), Fac (view3), ZER (view4), KAR (view5) and MOR (view6).

•**Caltech101-7**[2] contains a subset of images from Caltech101. There are 7 categories selected with 1474 images: faces, motorbikes, dollar-bill, garfield, snoopy, stop-sign, and windsor-chair. 6 types of features are used: Gabor (view1), WM (view2), CENTRIST (view3), HOG (view4), GIST (view5) and LBP (view6).

[1] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

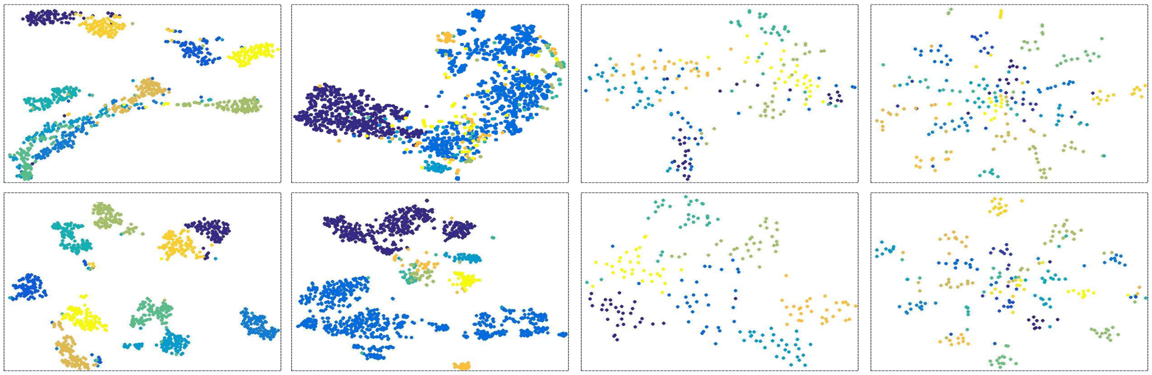
[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Method	Metrics	kNN	ITML	LMNN	LDML	GMML	Ours
handwritten	Accuracy	.972±.008	.969±.013	.928±.012	.976±.005	.977±.006	**.979±.006**
	F1-score	.945±.016	.939±.026	.864±.022	.953±.011	.955±.011	**.959±.011**
	Precision	.944±.017	.939±.026	.863±.023	.953±.011	.955±.011	**.958±.013**
	Recall	.946±.015	.938±.026	.865±.020	.953±.010	.955±.011	**.959±.010**
Caltech101	Accuracy	.879±.014	.901±.014	.838±.024	.879±.014	.879±.014	**.977±.007**
	F1-score	.863±.024	.900±.023	.881±.019	.863±.024	.867±.024	**.982±.008**
	Precision	.815±.029	.841±.030	.862±.033	.815±.029	.821±.027	**.967±.015**
	Recall	.961±.027	.969±.020	.903±.034	.955±.026	.964±.025	**.997±.001**
MSRA	Accuracy	.719±.069	.795±.068	.752±.097	.793±.079	.721±.068	**.874±.042**
	F1-score	.583±.093	.660±.103	.605±.143	.656±.124	.583±.093	**.766±.078**
	Precision	.572±.099	.657±.099	.604±.138	.660±.126	.574±.099	**.779±.081**
	Recall	.596±.096	.671±.090	.608±.152	.654±.128	.596±.096	**.755±.085**
football	Accuracy	.727±.069	**.847±.047**	.608±.076	.812±.051	.749±.055	.824±.038
	F1-score	.498±.124	**.743±.091**	.474±.115	.698±.085	.556±.107	.680±.116
	Precision	.428±.151	**.718±.109**	.455±.123	.677±.105	.497±.132	.645±.138
	Recall	.635±.057	**.776±.090**	.502±.117	.725±.081	.651±.057	.729±.108

Table 1: Comparison to metric learning methods with best single view.

Method	Metrics	kNN	ITML	LMNN	LDML	GMML	HMML	EMGMML	Ours
handwritten	Accuracy	.941±.015	.948±.013	.922±.020	.944±.012	.939±.011	.927±.013	.839±.012	**.979±.006**
	F1-score	.886±.027	.901±.022	.854±.037	.892±.020	.884±.018	.861±.023	.770±.017	**.959±.011**
	Precision	.884±.028	.901±.022	.853±.038	.892±.019	.884±.018	.860±.023	.760±.019	**.958±.013**
	Recall	.888±.026	.901±.023	.854±.036	.892±.021	.885±.019	.861±.023	.780±.016	**.959±.010**
Caltech101	Accuracy	.882±.012	.915±.016	.830±.061	.882±.012	.881±.013	.921±.013	.919±.007	**.977±.007**
	F1-score	.862±.021	.921±.018	.810±.099	.862±.062	.861±.023	.913±.016	.903±.010	**.982±.008**
	Precision	.787±.034	.873±.028	.778±.113	.787±.034	.786±.033	.861±.025	.830±.015	**.967±.015**
	Recall	.954±.024	.974±.016	.847±.084	.954±.024	.952±.030	.970±.016	.991±.004	**.997±.001**
MSRA	Accuracy	.700±.092	.769±.057	.767±.098	.700±.092	.702±.095	.798±.044	.802±.030	**.874±.042**
	F1-score	.555±.125	.616±.089	.627±.140	.555±.125	.556±.140	.641±.071	.651±.032	**.766±.078**
	Precision	.540±.125	.595±.086	.622±.135	.540±.125	.543±.148	.620±.062	.628±.036	**.779±.081**
	Recall	.575±.137	.641±.103	.632±.148	.575±.137	.574±.142	.667±.096	.675±.033	**.755±.085**
football	Accuracy	.631±.077	.580±.078	.416±.053	.627±.078	.651±.070	.522±.093	.702±.088	**.824±.038**
	F1-score	.432±.106	.413±.071	.230±.075	.426±.105	.469±.095	.305±.078	.349±.174	**.680±.116**
	Precision	.351±.103	.377±.071	.209±.070	.345±.101	.402±.087	.261±.077	.287±.179	**.645±.138**
	Recall	.576±.119	.460±.080	.264±.096	.571±.125	.570±.121	.383±.107	.483±.126	**.729±.108**

Table 2: Comparison to metric learning methods with multiple views.



(a) handwritten (b) Caltech101 (c) MSRA (d) football

Figure 2: Visualization of features with t-SNE. The top row corresponds to direct concatenation of the original feature vectors of multiple views (i.e., $[\mathbf{x}^{(1)}; ...; \mathbf{x}^{(V)}]$), while the bottom row is the visualization result of our approach (i.e., $[\mathbf{P}^{(1)}\mathbf{x}^{(1)}; ...; \mathbf{P}^{(V)}\mathbf{x}^{(V)}]$).

•**MSRA** [Liu *et al.*, 2010] contains 210 images labeled with 7 classes: tree, building, airplane, cow, face, car, and bicycle. 6 types of features are extracted: CENT (view1), CMT (view2), GIST (view3), HOG (view4), LBP (view5), and SIFT (view6).

•**football**[3] consists of 248 English Premier League football players on Twitter labeled with 20 communities. There are 6 views describing relationships between two users: follows (view1), followed by (view2), mentions (view3), mentioned by (view4), retweets (view5) and retweed by (view6).

We compared our method with the following baselines:

○ kNN. We conduct kNN based on Euclidean distance for each single view of features and feature concatenation.

○ ITML (Information-Theoretic Metric Learning) [Davis *et al.*, 2007]. The method characterizes the metric using a Mahalanobis distance by formulating the problem as minimizing the differential relative entropy between two multivariate Gaussians under constraints on the distance function.

○ LMNN ((Large Margin Nearest Neighbors) [Weinberger and Saul, 2009]. The method learns a Mahanalobis distance metric to improve kNN classification.

○ LDML (Logistic Discriminant Metric Learning) [Guillaumin *et al.*, 2009]. This method employs logistic discriminant to learn a metric such that positive pairs have smaller distances than negative pairs.

○ HMML (Heterogeneous Multi-Metric Learning) [Zhang *et al.*, 2011]. The method proposes to jointly learn multiple optimal homogenous/heterogeneous metrics in order to fuse the data collected from multiple sensors for classification by generalizing the LMNN framework.

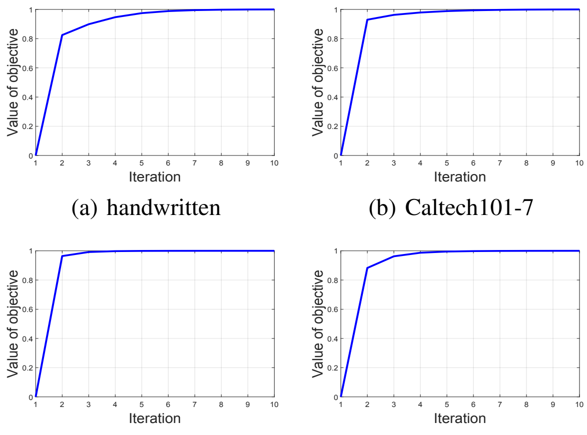
○ GMML (Geometric Mean Metric Learning) [Zadeh *et al.*, 2016]. The method is built on geometric intuition, and learns a symmetric positive definite matrix by formulating it as a smooth, strictly convex optimization problem.

○ EMGMML (Efficient Multi-modal Geometric Mean Metric Learning) [Liang *et al.*, 2017]. The method proposes to learn a set of optimal homogenous/heterogeneous metrics by generalizing the GMML framework.

Each dataset is randomly partitioned into 80% for training and 20% for testing. Then 20% samples are randomly selected from the training set as validation set for parameter tuning. We select the value from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ for λ_1 and λ_2. Uniformly, we set the number of nearest neighborhoods to 5 for all methods on each dataset. For the randomness involved in data partition, we run 10 times and report the averaged performance with deviation.

5.2 Results

Since the objective is non-decreasing with the iterations, the algorithm is guaranteed to converge. We conduct convergence experiments on four datasets and show as in Fig.3. As shown in Table 1, we first compared ours with existing metric learning methods with the best single view. It is observed that our FISH-MML achieves the best performance on 3 out of 4 datasets in terms of all evaluation metrics. As a strong competitor, ITML performs as the best on football. However, on handwritten, Caltech101-7 and MSRA, ITML does not perform very well. As shown in Table 2, we also compared our method with multi-view metric learning approaches. For traditional single-view metric learning methods, we concatenate feature vectors from multiple views as input. There are also two comparisons which are specially designed for multi-view data, i.e., HMML and EGMML. Our method outperforms all the comparisons on these four datasets, which further demonstrates the advantage of FISH-MML in exploring complementarity from multiple views. Fig.2 intuitively demonstrates the advantage of our approach by using t-distributed stochastic neighbor embedding (t-SNE) [Maaten and Hinton, 2008], since clusters in terms of ground-truth labels with our model are more compact and separable than those of directly combining different views.



(a) handwritten (b) Caltech101-7

(c) MSRA (d) football

Figure 3: Convergence experiment.

6 Conclusions

This paper has proposed a metric learning model for multi-view data which aims at jointly learning multiple metrics for multiple views. Our proposal has the advantage of simultaneously exploring the intra-view relationships and inter-view correlations in a unified framework. Specifically, we introduce Fisher discriminant analysis to enhance separability and expressiveness, and utilize Hilbert-Schmidt Independence Criteria to ensure consistence across different views. Our method is relatively simple to implement and easy to optimize with guaranteed convergence to local minimal. Experiments on benchmark datasets have verified the advantages of our approach over state-of-the-art metric learning methods.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grand No:61602337, 61732011, 61702358).

References

[Blaschko and Lampert, 2008] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. In *CVPR*, pages 1–8, 2008.

[3] http://mlg.ucd.ie/aggregation/index.html

[Cao *et al.*, 2015] X. Cao, C. Zhang, H. Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.

[Chaudhuri *et al.*, 2009] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136, 2009.

[Cheng *et al.*, 2011] Qiang Cheng, Hongbo Zhou, and Jie Cheng. The fisher-markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. *IEEE T-PAMI*, 33(6):1217–1233, 2011.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[Dhillon *et al.*, 2011] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via c-ca. In *NIPS*, pages 199–207, 2011.

[Ding *et al.*, 2015] Zhengming Ding, Sungjoo Suh, Jae-Joon Han, Changkyu Choi, and Yun Fu. Discriminative low-rank metric learning for face recognition. In *FG*, volume 1, pages 1–6, 2015.

[Gong, 2017] Chen Gong. Exploring commonality and individuality for multi-modal curriculum learning. In *AAAI*, pages 1926–1933, 2017.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, volume 16, pages 63–78. Springer, 2005.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009.

[Hu *et al.*, 2014] Junlin Hu, Jiwen Lu, Junsong Yuan, and Yap-Peng Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *ACCV*, pages 252–267. Springer, 2014.

[Hu *et al.*, 2017] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Sharable and individual multi-view metric learning. *IEEE T-PAMI*, 2017.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[Liang *et al.*, 2017] Jianqing Liang, Qinghua Hu, Pengfei Zhu, and Wenwu Wang. Efficient multi-modal geometric mean metric learning. *Pattern Recognition*, 2017.

[Liu *et al.*, 2010] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung Yeung Shum. Learning to detect a salient object. *IEEE T-PAMI*, 33(2):353–367, 2010.

[Liu *et al.*, 2016] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In *AAAI*, pages 1888–1894, 2016.

[Lowe, 1999] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[Lu *et al.*, 2015] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[Sindhwani and Rosenberg, 2008] Vikas Sindhwani and David S Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *ICML*, pages 976–983, 2008.

[Song *et al.*, 2007] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *ICML*, pages 823–830, 2007.

[Wang *et al.*, 2016] Shuyang Wang, Zhengming Ding, and Yun Fu. Coupled marginalized auto-encoders for cross-domain multi-view learning. In *IJCAI*, pages 2125–2131, 2016.

[Weinberger and Saul, 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[Xiao and Guo, 2015] Min Xiao and Yuhong Guo. Feature s-pace independent semi-supervised domain adaptation via kernel matching. *IEEE T-PAMI*, 37(1):54–66, 2015.

[Xie and Xing, 2013] Pengtao Xie and Eric P Xing. Multi-modal distance metric learning. 2013.

[Xing *et al.*, 2003] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *NIPS*, pages 521–528, 2003.

[Ying *et al.*, 2009] Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse metric learning via smooth optimization. In *NIPS*, pages 2214–2222, 2009.

[Zadeh *et al.*, 2016] Pourya Zadeh, Reshad Hosseini, and Suvrit S-ra. Geometric mean metric learning. In *ICML*, pages 2464–2471, 2016.

[Zhang *et al.*, 2011] Haichao Zhang, Thomas S Huang, Nasser M Nasrabadi, and Yanning Zhang. Heterogeneous multi-metric learning for multi-sensor fusion. In *Information Fusion (FUSION)*, pages 1–8, 2011.

[Zhang *et al.*, 2015] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, pages 1582–1590, 2015.

[Zhang *et al.*, 2017] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao. Latent multi-view subspace clustering. In *CVPR*, pages 4333–4341, 2017.

[Zhao *et al.*, 2017] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017.