The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

A Probabilistic Hierarchical Model for Multi-View and Multi-Feature Classification

Jinxing Li,† Hongwei Yong,† Bob Zhang,‡ Mu Li,† Lei Zhang,† David Zhang†
(email: {csjxli, cshyong, csmuli, cslzhang, csdzhang}@comp.polyu.edu.hk)
†Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
‡ Department of Computer and Information Science, University of Macau, Macau, China
(email: bobzhang@umac.mo)

Abstract

Some recent works in classification show that the data obtained from various views with different sensors for an object contributes to achieving a remarkable performance. Actually, in many real-world applications, each view often contains multiple features, which means that this type of data has a hierarchical structure, while most of existing works do not take these features with multi-layer structure into consideration simultaneously. In this paper, a probabilistic hierarchical model is proposed to address this issue and applied for classification. In our model, a latent variable is first learned to fuse the multiple features obtained from a same view, sensor or modality. Particularly, mapping matrices corresponding to a certain view are estimated to project the latent variable from a shared space to the multiple observations. Since this method is designed for the supervised purpose, we assume that the latent variables associated with different views are influenced by their ground-truth label. In order to effectively solve the proposed method, the Expectation-Maximization (EM) algorithm is applied to estimate the parameters and latent variables. Experimental results on the extensive synthetic and two real-world datasets substantiate the effectiveness and superiority of our approach as compared with state-of-the-art.

Introduction

In many practical applications, the raw data is obtained from various domains or extracted from diverse modalities. For instance, a person can be verified by the fingerprint, palm print, or iris; a face image can be captured from different angles. These multiple types of data are called multi-view or multi-modal data, which have attracted much attention in recent years. Due to the comprehensive information provided by multiple views, multi-view methods have achieved better performances in many fields such as object recognition (Eleftheriadis, Rudovic, and Pantic 2015), (Li, Zhang, and Zhang 2017b), (Yang et al. 2012) (Jing et al. 2014), disease detection (Li et al. 2016) (Li, Zhang, and Zhang 2017a), cross modal learning (Song et al. 2015), and semi-supervised learning (Zhang and Zhang 2016) (Ceci et al. 2015) (Tao et al. 2017), regression (Zheng et al. 2015) etc., compared with those methods only using the information from single view.

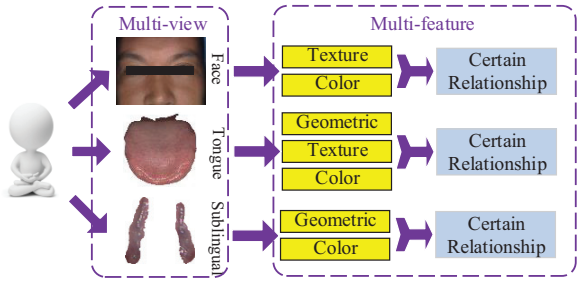

Figure 1: An example of the multi-view and multi-feature object. A person can be diagnosed through his or her tongue, face and sublingual vessel. Also, these modalities can be represented with different features.

Although many methods are proposed for multi-view data to integrate different views to get an outstanding performance, there are still some difficulties for us to tackle. One key problem is that apart from obtained multiple views, each view is also represented with various features that are fully valuable for classification. For instance, a person is verified by the combination of iris, fingerprint, face and palm print. Furthermore, an iris image is described with wavelet and gabor. So do other modalities. Similarly, Li et al. (Li et al. 2016) (Li, Zhang, and Zhang 2017a) also illustrated that the fatty liver and diabetes mellitus disease can be detected through a patient's tongue, face and sublingual vessel, and these three types of modalities are also represented with color, texture, and geometric features, as shown in Fig.1. To the best of our knowledge, many existing methods only established models for multiple modalities, which is limited for the data including multiple views and their corresponding multiple features. For this situation, a naive way is to only concatenate different features acquired from a same view as a large one, and the existing multi-view approach is then applied, as done in (Li et al. 2016) and (Li, Zhang, and Zhang 2017a). However, this kind of strategy ignores the correlation among multiple features, and the larger vector would result in over-fitting if the dimensionality of each feature is relatively large, encountering the performance degradation subsequently.

In order to address this problem, we propose a probabilis-

tic generative model by modeling the multi-view and multi-feature data under a hierarchical structure. With the observed features from a view or modality, a shared and latent variable is learned as the fused feature. Additionally, since our model is constructed for classification, the learned variables associated with different views are assumed to be independently influenced by their ground-truth label.

The main contributions of this paper are shown as follows:
(1) To the best of our knowledge, this paper is the first one to investigate the multi-view and multi-feature classification problem. Many conventional multi-view methods, including (Li, Zhang, and Zhang 2017a), (Zhang and Zhang 2016), (Eleftheriadis, Rudovic, and Pantic 2015) and (Li et al. 2016), etc., only regard the multi-feature as a particular case of the multi-view, and apply the same algorithm to model the multi-view or multi-feature data. By contrast, this paper does not only exploit the relationship across different features, but also reveal the correlation among various views.
(2) A probabilistic hierarchical method is proposed for multi-view and multi-feature learning. This model hierarchically fuses multiple features through a latent variable. Different from many existing methods which only output the decided label, the corresponding probability belonging to a certain category is predicted by combining the fused variables from different views.
(3) The EM-based algorithm (Bishop 2006) is introduced to solve our model effectively. Particularly, a closed-form solution for each parameter or variable is obtained, and we alternatively update the parameters and variables until convergence.

Related Works

Recently, many works on multi-view learning have been proposed. The Canonical Correlation Analysis (CCA) (Hotelling 1936) aims to learn two types of mapping functions to project two views into a common space, maximizing their correlation. Considering the data corrupted by the noise and outliers, the robust CCA (Nicolaou et al. 2014) (Bach and Jordan 2005), sparse CCA (Archambeau and Bach 2009) were presented by introducing the Student-t density model and l_1 norm. Additionally, Andrew et al. and Wang et al. also described the deep canonical correlation analysis (DCCA) (Andrew et al. 2013) and deep canonically correlated autoencoders (DCCAE) (Wang et al. 2015) by extending the CCA to the deep structure. Due to the limitation of conventional CCA which is only adaptive for two views, the multi-view CCA (Chaudhuri et al. 2009) was presented to maximize the summarization of all pair correlation. Similarly, the supervised subspace learning-Linear Discriminative analysis (LDA) is also extended for multi-view data. For instance, the multi-view fisher discriminative analysis (MFDA) (Diethe, Hardoon, and Shawe-Taylor 2010) independently trains a classifier for each view to make the average distance of different categories as large as possible. Furthermore, multi-view discriminant analysis (MvDA) was proposed by Kan et al. (Kan et al. 2016) to extract the discriminative feature from a common space through a linear transformation, encouraging the extracted features belonging to the same class to be close and those belonging to the different classes to be far.

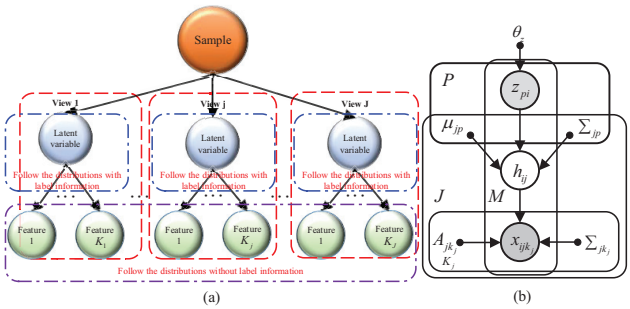

Figure 2: (a) The framework of the proposed method, where the number of the latent variables is equal to the number of observed views. (b) The probabilistic framework of the proposed method.

Due to the effectiveness of the sparse and collaborative representation, some joint representation based approaches were proposed. Unlike some approaches which only consider the common part, Li et al. (Li et al. 2016) separated the sparse representation coefficients into the similar and specific components (JSSL), which achieves a satisfied performance on the diabetes mellitus and impaired glucose regulation detection. The $l_{2,1}$ norm was introduced in (Yuan, Liu, and Yan 2012) to measure the consistence of the representation coefficients across different views (MTJSRC). Yang et al. (Yang et al. 2012) proposed a relaxed collaborative representation method (RCR) to make the coefficient belonging to different views be similar. Considering the label information, a discriminant collaborative representation method (JDCR) was proposed in (Li, Zhang, and Zhang 2017a) for the multi-view data. Besides, Guo (Guo 2013) also proposed a novel approach to get a convex subspace among various views (CSRL). Because of the power of dictionary learning, a sparse model was described in (Bahrampour et al. 2016) (UMDL and SMDL) to learn a multi-modal dictionary which greatly exploit the correlation among different modalities.

Although various methods have been proposed for the multi-view data, there are few works which take the aforementioned hierarchical data into account. Generally speaking, multiple features obtained from a single view would contain a certain relationship. So do multiple views captured from a same object. Thus, it is significant to present a novel approach to hierarchically exploit the valuable information across these types of features and views.

The Proposed Method

In this section, the hierarchical probabilistic model is analyzed for multi-view and multi-feature classification, followed by its efficient optimization and prediction inference.

The Hierarchical probabilistic Model

The framework and graphic model of the proposed method are shown in Fig.2. As we can see, an object is

observed from J views, and the j-th ($j \in \{1, \cdots, J\}$) view is represented by K_j types of features. Specifically, the data and the label of the i-th sample can be represented as $\{\mathbf{x}_{ijk_j} \in \mathbb{R}^{D_{jk_j}}\}_{j,k_j=1}^{J,K_j}$ and a one-hot categorical variable $\mathbf{z}_i \in \mathbb{R}^P$, respectively, where D_{jk_j} is the dimension of the k_j-th feature in the j-th view, P is the number of the classes, and $\mathbf{z}_i$ satisfies $z_{pi} \in \{0, 1\}$ and $\sum_{p=1}^{P} z_{pi} = 1$.

Then some probability assumptions for these variables are made to construct our hierarchical model. Firstly, the categorical distribution is introduced for the categorical variable $\mathbf{z}_i$, which has the following form:

$$p(\mathbf{z}_i \mid \boldsymbol{\theta}_{\boldsymbol{Z}}) = \prod_{p=1}^{P} \pi_p^{z_{pi}} \tag{1}$$

where $\boldsymbol{\theta}_{\boldsymbol{Z}} = \{\pi_p\}_{p=1}^{P}$, $z_{pi} = 1$ if the i-th sample belongs to the p-th class, otherwise $z_{pi} = 0$ and the variable $\pi_p \in [0, 1]$ follows $\sum_{p=1}^{P} \pi_p = 1$, which means the probability of a sample belonging to its corresponding category.

In order to exploit the discriminative information, a latent variable $\mathbf{h}_{ij} \in \mathbb{R}^{D_j}$ corresponding to the j-th view of the i-th sample is then learned by imposing the ground-truth label on it. Specifically, for distinctive categories, the distributions of the latent variables belonging to different views are different, greatly exploiting the complementary information across multiple views. The most common and simple assumption for $\mathbf{h}_{ij}$ is that

$$p(\mathbf{h}_{ij} \mid \mathbf{z}_i, \boldsymbol{\theta}_{\boldsymbol{H}}) = \prod_{p=1}^{P} \mathcal{N}(\mathbf{h}_{ij} \mid \boldsymbol{\mu}_{jp}, \boldsymbol{\Sigma}_{jp})^{z_{pi}} \tag{2}$$

where $\boldsymbol{\theta}_{\boldsymbol{H}} = \{\boldsymbol{\mu}_{jp}, \boldsymbol{\Sigma}_{jp}\}_{j,p=1}^{J,P}$, meaning that its distribution for the p-th category is a Gaussian distribution with mean $\boldsymbol{\mu}_{jp}$ and covariance matrix $\boldsymbol{\Sigma}_{jp}$.

In general, it is reasonable to assume that multiple features are the projections from a shared variable through different mapping functions. Thus, in the proposed model, a mapping matrix for each feature in a same modality is learned to transform the the latent variable $\mathbf{h}_{ij}$ to the observed data $\mathbf{x}_{ijk_j}$ by a linear Gaussian model, which can be presented as following equation:

$$p(\mathbf{x}_{ijk_j} \mid \mathbf{h}_{ij}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \mathcal{N}(\mathbf{x}_{ijk_j} \mid \mathbf{A}_{jk_j}\mathbf{h}_{ij} + \mathbf{b}_{jk_j}, \boldsymbol{\Sigma}_{jk_j}) \tag{3}$$

where $\boldsymbol{\theta}_{\boldsymbol{X}} = \{\mathbf{A}_{jk_j}, \mathbf{b}_{jk_j}, \boldsymbol{\Sigma}_{jk_j}\}_{j,k_j=1}^{J,K_j}$, $\mathbf{A}_{jk_j}$ is the learned mapping matrix, $\mathbf{b}_{jk_j}$ is the bias and $\boldsymbol{\Sigma}_{jk_j}$ denotes the covariance matrix.

Moreover, we also make some reasonable conditional independence assumptions about different features and different views, including

$$\begin{aligned} &p(\{\mathbf{x}_{ijk_j}\}_{k_j=1}^{K_j} \mid \mathbf{h}_{ij}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \prod_{k_j=1}^{K_j} p(\mathbf{x}_{ijk_j} \mid \mathbf{h}_{ij}, \boldsymbol{\theta}_{\boldsymbol{X}}) \\ &p(\{\{\mathbf{x}_{ijk_j}\}_{k_j=1}^{K_j}, \mathbf{h}_{ij}\}_{j=1}^{J} \mid \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{H}}, \mathbf{z}_i) \\ &\quad = \prod_{j=1}^{J} p(\{\mathbf{x}_{ijk_j}\}_{k_j=1}^{K_j}, \mathbf{h}_{ij} \mid \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{H}}, \mathbf{z}_i) \end{aligned} \tag{4}$$

In order to acquire a simple representation derivation, let $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{M}$, $\mathbf{H} = \{\mathbf{h}_{ij}\}_{i,j=1}^{M,J}$ and $\mathbf{X} = \{\mathbf{x}_{ijk_j}\}_{i,j,k_j=1}^{M,J,K_j}$. Taking the aforementioned independent and identically distributed (i.i.d.) assumption into account, the join distribution w.r.t. all variables is obtained:

$$\begin{aligned} &P(\mathbf{X}, \mathbf{Z}, \mathbf{H} \mid \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) = \\ &\prod_{i=1}^{M}\{p(\mathbf{z}_i \mid \boldsymbol{\theta}_Z) \prod_{j=1}^{J}\{p(\mathbf{h}_{ij} \mid \mathbf{z}_i, \boldsymbol{\theta}_H) \prod_{k_j=1}^{K_j} p(\mathbf{x}_{ijk_j} \mid \mathbf{h}_{ij}, \boldsymbol{\theta}_X)\}\} \end{aligned} \tag{5}$$

which is a probabilistic hierarchical model. Generally speaking, it is infeasible to estimate the covariance matrix $\boldsymbol{\Sigma}_{jk_j}$ and $\boldsymbol{\Sigma}_{jp}$, when the dimensions of the features and the latent variables are large. To avoid overfitting, $\boldsymbol{\Sigma}_{jp}$ and $\boldsymbol{\Sigma}_{jk_j}$ can be set to be $\sigma_{jp}^2\mathbf{I}$ and $\sigma_{jk_j}^2\mathbf{I}$ in this case, where σ_{jp} and σ_{jk_j} are two 1-D variables to control their variances of all dimensions, and $\mathbf{I}$ is identical matrix.

To estimate the parameters of this probabilistic method, the log-likelihood function w.r.t. all variables should be optimized. Since it is difficult to directly observe the latent variable $\mathbf{H}$, the log-likelihood function only related to the multi-view and multi-feature data $\mathbf{X}$ and its label variable $\mathbf{Z}$ is considered. Therefore, the objective function is

$$\log P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) \tag{6}$$

However, it is quite different between marginalizing $\mathbf{H}$ in Eq.(5) and optimizing the objective function (6). Fortunately, the Expectation-Maximization(EM) (Bishop 2006) algorithm can be readily utilized for efficiently solving this kind of problem with latent variables.

Optimization

As analyzed above, the EM algorithm, which is a two-stage iterative optimization technique for finding maximum likelihood solutions, is employed to estimate the model parameters. Specifically, the posterior probability of latent variable $\mathbf{H}$ will be calculated in E-step, followed by the estimation of the value of parameters $\boldsymbol{\theta}_Z$, $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_X$ in M-step.

E Step: Primarily, we use the current values of all parameters to evaluate the posterior probabilities of $\mathbf{H}$. The log-posterior function

$$\begin{aligned} &\log P(\mathbf{H} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) \propto \log P(\mathbf{X}, \mathbf{H}, \mathbf{Z} \mid \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) \\ &\propto \sum_{i=1}^{M}\{\sum_{j=1}^{J}\{-\frac{1}{2}\mathbf{h}_{ij}^T(\sum_{p=1}^{P} z_{pi}\boldsymbol{\Sigma}_{jp}^{-1} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^T\boldsymbol{\Sigma}_{jk_j}^{-1}\mathbf{A}_{jk_j})\mathbf{h}_{ij} \\ &+ \mathbf{h}_{ij}^T(\sum_{p=1}^{P} z_{pi}\boldsymbol{\Sigma}_{jp}^{-1}\boldsymbol{\mu}_{jp} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^T\boldsymbol{\Sigma}_{jk_j}^{-1}(\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j}))\}\} \end{aligned} \tag{7}$$

is an a quadratic form function w.r.t. $\mathbf{h}_{ij}$. Thus the posterior probability of $\mathbf{h}_{ij}$ follows a Gaussian distribution, which can be rewritten as follows:

$$p(\mathbf{h}_{ij} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) = \mathcal{N}(\mathbf{h}_{ij} \mid \boldsymbol{\mu}_{ij}^H, \boldsymbol{\Sigma}_{ij}^H) \tag{8}$$

where

$$\boldsymbol{\Sigma}_{ij}^{H} = (\sum_{p=1}^{P} z_{pi}\boldsymbol{\Sigma}_{jp}^{-1} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^{T}\boldsymbol{\Sigma}_{jk_j}^{-1}\mathbf{A}_{jk_j})^{-1}$$

$$\boldsymbol{\mu}_{ij}^{H} = \boldsymbol{\Sigma}_{ij}^{H}\{\sum_{p=1}^{P} z_{pi}\boldsymbol{\Sigma}_{jp}^{-1}\boldsymbol{\mu}_{jp} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^{T}\boldsymbol{\Sigma}_{jk_j}^{-1}(\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j})\} \tag{9}$$

M Step: In M-step, all parameters are re-estimated by optimizing a concave low-bound function for (6), being

$$L(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H) = E_H[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{H} \mid \boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H)] \tag{10}$$

with a unique maximum point. In Eq.(10), the format is similar to log joint density $\log p(\mathbf{X}, \mathbf{Z}, \mathbf{H} \mid \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X)$, except to replace $\mathbf{h}_{ij}$ and $\mathbf{h}_{ij}\mathbf{h}_{ij}^T$ with $E[\mathbf{h}_{ij}]$ and $E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T]$, respectively. Since the mean and covariance of the posterior probability for $\mathbf{h}_{ij}$ are $\boldsymbol{\mu}_{ij}^H$ and $\boldsymbol{\Sigma}_{ij}^H$, which have been calculated in E-step, $E[\mathbf{h}_{ij}]$ and $E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T]$ can be obtained through the following equation:

$$\begin{aligned} E[\mathbf{h}_{ij}] &= \boldsymbol{\mu}_{ij}^H \\ E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T] &= \boldsymbol{\Sigma}_{ij}^H + \boldsymbol{\mu}_{ij}^H(\boldsymbol{\mu}_{ij}^H)^T \end{aligned} \tag{11}$$

By calculating the derivative of the low-bound function $L(\boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X)$ w.r.t. $\boldsymbol{\theta}_Z$, $\boldsymbol{\theta}_H$, and $\boldsymbol{\theta}_X$, and setting it to be zero, the parameters can be estimated with closed-form solutions. The results are listed as follows.

For the parameters corresponding to $\boldsymbol{\theta}_X$, the solutions are

$$\begin{aligned} \mathbf{A}_{jk_j} &= \{\sum_{i=1}^{M}(\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j})E[\mathbf{h}_{ij}^T]\}\{\sum_{i=1}^{M} E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T]\}^{-1} \\ \mathbf{b}_{jk_j} &= \frac{1}{M}\sum_{i=1}^{M}\{\mathbf{x}_{ijk_j} - \mathbf{A}_{jk_j}E[\mathbf{h}_{ij}]\} \\ \boldsymbol{\Sigma}_{jk_j} &= \frac{1}{M}\sum_{i=1}^{M}\{\mathbf{A}_{jk_j}E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T]\mathbf{A}_{jk_j}^T - 2\mathbf{A}_{jk_j}E[\mathbf{h}_{ij}] \\ &\quad (\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j})^T + (\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j})(\mathbf{x}_{ijk_j} - \mathbf{b}_{jk_j})^T\} \end{aligned} \tag{12}$$

For the parameters corresponding to $\boldsymbol{\theta}_H$, the solutions are

$$\begin{aligned} \boldsymbol{\mu}_{jp} &= \frac{1}{\sum_{i=1}^{M} z_{pi}}\sum_{i=1}^{M} z_{pi}E[\mathbf{h}_{ij}] \\ \boldsymbol{\Sigma}_{jp} &= \frac{1}{\sum_{i=1}^{M} z_{pi}}\sum_{i=1}^{M} z_{pi}\{E[\mathbf{h}_{ij}\mathbf{h}_{ij}^T] - 2E[\mathbf{h}_{ij}]\boldsymbol{\mu}_{jp}^T + \boldsymbol{\mu}_{jp}\boldsymbol{\mu}_{jp}^T\} \end{aligned} \tag{13}$$

To estimate the parameter $\boldsymbol{\theta}_Z = \{\pi_p\}_{p=1}^P$, the Lagrange Multiplier term is introduced to meet $\sum_{p=1}^P \pi_p = 1$. By calculating the derivative of the Lagrange function w.r.t. π_p and setting it to 0, the solution of π_p is then obtained according to the following equation:

$$\pi_p = \frac{\sum_{i=1}^{M} z_{pi}}{\sum_{p=1}^{P}\sum_{i=1}^{M} z_{pi}} \tag{14}$$

From Eq. (12), Eq. (13) and Eq. (14), we can see that each step has a closed-form solution which would greatly facilitate the parameter estimation process. According to the convergence theory of EM algorithm, each update of the parameters acquired from an E-step followed by an M-step can guarantee the increase of the log likelihood function. Hence, to obtain a local maximin point, we alternatively execute E-step and M-step until convergence. The Algorithm 1 illustrates the details of the optimization. In this paper, our proposed method is named as Hierarchical Multi-view Multi-feature Fusion (HMMF).

Algorithm 1 [HMMF]Hierarchical Multi-view Multi-feature Fusion

Input: Observed data: $\mathbf{X}$; label: $\mathbf{Z}$;
Initialization: $\boldsymbol{\theta}_Z$;$\boldsymbol{\theta}_H$; $\boldsymbol{\theta}_X$

```
(Calculate θ_Z)
for p = 1, ..., P do
    Calculate π_p by Eq (14)
end for
while not converged do
    E-step:
    for i = 1, ..., M do
        for j = 1, ..., J do
            Evaluate Σ_ij^H and μ_ij^H by Eq.(9),
            Calculate E[h_ij] and E[h_ij h_ij^T] by Eq.(11).
        end for
    end for
    M-step: (re-estimate θ_H and θ_X)
    for j = 1, ..., J do
        for p = 1, ..., P do
            calculate μ_jp and Σ_jp through Eq.(13)
        end for
        for k_j = 1, ..., K_j do
            calculate Σ_{jk_j}, b_{jk_j} and A_{jk_j} through Eq.(12)
        end for
    end for
end while
```

Output: $\boldsymbol{\theta}_Z$;$\boldsymbol{\theta}_H$; $\boldsymbol{\theta}_X$

Complexity: To simplify the description of the computational complexity per iteration, here we firstly give several definitions: $D_1 = \max(D_j)$, $D_2 = \max(D_{jk_j})$, and $K = \max(K_j)$. Thus the complexity of our algorithm is $O(MC + J(CD_1^3 + K(MD_2D_1 + D_2^2D_1 + D_2^3)))$ for the general covariance matrix and $O(MC + JD_1^3 + JKMD_1D_2)$ for the diagonal covariance matrix, where M, C and J are the number of samples, categories and views, respectively. To be honest, the algorithm converges in 10 iterations in most cases. However, to make a full convergence, we set the number of iterations to 100 in our experiments.

Prediction

According to the Bayesian principle, the posterior probability of a given test sample $\mathbf{x} = \{\mathbf{x}_{jk_j}\}_{j,k_j=1}^{J,K_j}$ belonging to the p-th class is calculated through

$$\begin{aligned} &p(z_p = 1 \mid \mathbf{x}, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X) \\ &= \frac{p(\mathbf{x} \mid z_p = 1, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X)p(z_p = 1)}{\sum_{p=1}^{P} p(\mathbf{x} \mid z_p = 1, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X)p(z_p = 1)}. \end{aligned} \tag{15}$$

Since the second term of the numerator $p(z_p = 1) = \pi_p$, the key problem is how to calculate the first term of the numer-

ator. Actually, we get its value by the following process:

$$\begin{aligned}
&\log p(\mathbf{x} \mid z_p = 1, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X) \\
&= \log \int p(\mathbf{x}, \mathbf{h} \mid z_p = 1, \boldsymbol{\theta}_Z, \boldsymbol{\theta}_H, \boldsymbol{\theta}_X) d\mathbf{h} \\
&= \sum_{j=1}^{J} \{ (-\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_{jp}|}{|\boldsymbol{\Sigma}_{jp}^H|} - \frac{1}{2} \boldsymbol{\mu}_{jp}^T \boldsymbol{\Sigma}_{jp}^{-1} \boldsymbol{\mu}_{jp}) + \\
&\quad \frac{1}{2} (\boldsymbol{\mu}_{jp}^H)^T (\boldsymbol{\Sigma}_{jp}^H)^{-1} \boldsymbol{\mu}_{jp}^H + \sum_{k_j}^{K_j} (-\frac{D_{jk_j}}{2} \log(2\pi) - \\
&\quad \frac{1}{2} \log |\boldsymbol{\Sigma}_{jk_j}| - \frac{1}{2} (\mathbf{x}_{jk_j} - \mathbf{b}_{jk_j})^T \boldsymbol{\Sigma}_{jk_j}^{-1} (\mathbf{x}_{jk_j} - \mathbf{b}_{jk_j})) \}
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{jp}^H &= (\boldsymbol{\Sigma}_{jp}^{-1} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^T \boldsymbol{\Sigma}_{jk_j}^{-1} \mathbf{A}_{jk_j})^{-1} \\
\boldsymbol{\mu}_{jp}^H &= \boldsymbol{\Sigma}_{jp}^H \{ \boldsymbol{\Sigma}_{jp}^{-1} \boldsymbol{\mu}_{jp} + \sum_{k_j=1}^{K_j} \mathbf{A}_{jk_j}^T \boldsymbol{\Sigma}_{jk_j}^{-1} (\mathbf{x}_{jk_j} - \mathbf{b}_{jk_j}) \}
\end{aligned} \tag{17}$$

Every parameter or variable is well-defined in Eq.(16). Thus first term of the numerator in Eq(15) as well as the posterior probability of this new sample belonging to the p-th class are easily gained. If only requiring the predicted label, the logarithm of the numerator in Eq(15) for all classes can be calculated, and predicted category is the one with the max value.

Experimental Results

In this section, we conduct experiments on both synthetic and real-world datasets to demonstrate the superiority of the proposed method. The datasets used in this paper are first described, followed by the experimental setting. The comparison among different approaches is then analyzed.

Datasets and Experimental Setting

The synthetic data is generated according to the assumption of the proposed method. Particularly, given the values of D_j and D_{k_j}, the parameters $\mathbf{A_{jk_j}}$, $\mathbf{\Sigma_{jk_j}}$, $\mathbf{\Sigma_{jp}}$ and $\mu_{\mathbf{jp}}$ are randomly generated. The latent variable $\mathbf{h}_{ij}$ is then obtained by following $\mathcal{N}(\mathbf{h}_{ij} \mid \sum_{p=1}^{P} z_{pi} \boldsymbol{\mu}_{jp}, \sum_{p=1}^{P} z_{pi} \boldsymbol{\Sigma}_{jp})$. Consequently, the observations are acquired according to $\mathcal{N}(\mathbf{x}_{ijk_j} \mid \mathbf{A}_{jk_j} \mathbf{h}_{ij}, \boldsymbol{\Sigma}_{jk_j})$. Without loss generality, we set the dimensionality D_j to be the same for each view. So does D_{jk_j}. In this experiment, we set D_j and D_{jk_j} to be 10 and 20, respectively.

We also select the biomedical dataset (Li et al. 2016) to evaluate the performance of the proposed method. This biomedical dataset was collected by the Hong Kong Polytechnic University at the Guangdong Provincial TCM Hospital, Guangdong, China, from the early 2014 to the late 2015, which aims to detect the Diabetes Mellitus (DM) disease from the healthy samples. Each instance can be represented by three modalities: facial image, tongue image and sublingual image. Concretely, the face image can be represented by the 24-dimensional color feature (4 block×6 dimension) and another 5-dimensional texture feature; the tongue image can be represented with 12-dimensional color feature, 9-dimensional texture feature and 13-dimensional geometry feature; and the sublingual image can be represented with 6-dimensional color feature and 6-geometrical feature. This dataset consists of 192 healthy and 198 DM samples. Additionally, 40, 50, 60, and 70 instances in each category are randomly selected for training with five independent times, and the rest of sample are exploited for testing.

The third one is the Wiki Text-Image dataset (Rasiwasia et al. 2010) collected from the WikiPedia's featured articles. In this dataset, each sample can be represented by two modalities including an image and a text. According to (Rasiwasia et al. 2010), 10 most populated categories (at least 150 instances per category) containing art, biology, geography, history, literature, media, music, royalty, sport and warfare are used for training and testing. Particularly, (Rasiwasia et al. 2010) separates the database into 2173 training samples and 693 test samples. The image view is described with the 128-D SIFT histogram image feature and the text view is presented with 10-D latent Dirichlet features. In order to make this dataset be multi-view and multi-feature style, we also apply the Alexnet (Krizhevsky, Sutskever, and Hinton 2012) to extract a CNN feature from the provided images. Note that, the dimensionality of the output of the Alexnet is reduced from 4096 to 30 in this paper to decrease the training time.

In order to illuminate the superiority of our method, we also make it compare with some single- and multi-view based strategies including DPL (Gu et al. 2014), MDL (UMDL and SMDL) (Bahrampour et al. 2016), JDCR (Li, Zhang, and Zhang 2017a), and CSRL (Guo 2013) on the real-world datasets. For DPL, we concatenate all features in all views as a single one. For other approaches, we concatenate all features in each view as a vector, thus vectors in different views are regarded as their inputs. Since CSRL aim to learn a latent variable, we apply SVM to it to do the classification.

For the parameter tuning on synthetic and Wiki Text-Image datasets, we tune the dimension D_j of the latent variable through 5-fold cross-validation using training data. In fact, we find that D_j being close to $\min(D_{jk_j})$ is fine for both datasets. For the Biomedical dataset, since the dimension of several features is around 5 and according to results of the first and third datasets, we set D_j to be 5 empirically.

Experimental Results on Three Datasets

Synthetic Dataset: In this experiment, we randomly generate four types of synthetic datasets, which are ($J = 2$, K_j=2), ($J = 3$, K_j=3), ($J = 4$, K_j=4) and ($J = 5$, K_j=5). As mentioned above, without loss generality, we set the number K_j of types of features in each view to be same. Note that D_j and D_{jk_j} are set to be 10 and 20, respectively. Additionally, we randomly generate 5 categories whose number of training and test samples is 20 and 100 in each class, respectively. In order to demonstrate the superiority of the hierarchical fusion, we reconstruct the inputs in another three types. For instance, as shown in Tab.1, when the number of views and their corresponding features are

Table 1: The classification accuracies on the synthetic dataset obtained by HMMF.

	(J, K_j)			
Model	(2,2)	(2,1)	(1,4)	(4,1)
Accuracy	**82.2%**	80.6%	63.6%	80.4%
Model	(3,3)	(3,1)	(1,9)	(9,1)
Accuracy	**86.8%**	85.0%	68.6%	84.8%
Model	(4,4)	(4,1)	(1,16)	(16,1)
Accuracy	**93.6%**	92.4%	82.0%	92.0%
Model	(5,5)	(5,1)	(1,25)	(25,1)
Accuracy	**94.4%**	93.0%	90.2%	91.0%

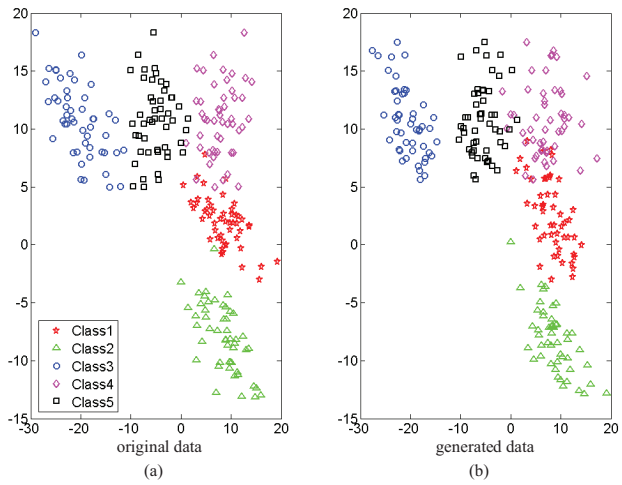

Figure 3: The comparison of the distributions between the original data and generated data.

($J = 3$, K_j=3), we concatenate the features in each view as a single one. Thus a novel input, whose J is 3 and K_j is 1 (3,1) is obtained. Additionally, we also follow the input as many existing multi-view methods do. We regard the sample with 9 types of features as the input and two cases including ($J = 1$, K_j=9 (1,9)) and ($J = 9$, K_j=1 (9,1)) are consequently acquired. From Tab.1 we can see that our proposed hierarchical fusion strategy always achieves a noticeable accuracy compared with other cases. The performance obtained by HMMF is particularly better than that in the third column, indicating the significance of hierarchical fusion structure. Furthermore, with the increasing of the number of views and features, there is also a remarkable improvement in the classification accuracy. The result is only 82.2% in the case of ($J = 2$, $K_j = 2$), while it has a great achievement when (J, K_j) rise to (4,4).

Due to application of the EM algorithm in our optimization, a local solution can be obtained. Thus it is necessary to have a discussion on the initialization. To be honest, we can initialize parameters in two ways: be consistent and be random. First: in our synthetic experiment, we generate the mapping matrices A through PCA (X_{jk_j} is the input), and $\boldsymbol{\Sigma}$ matrices, $\boldsymbol{\mu}$ and $\boldsymbol{b}$ vectors are set to be identical matrices and

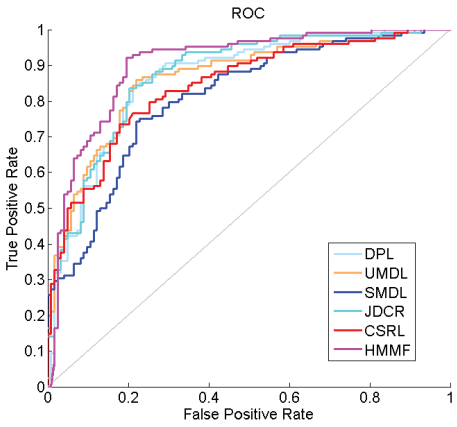

Figure 4: The ROC curves obtained by different methods in DM detection.

zeros vectors, respectively. In this way, the experimental result is consistent. Second: another one is to initialize parameters randomly. Although this method will have a influence on results, the fluctuation of results is acceptable. For instance, we randomly initialize all parameters with 10 times in the synthetic experiment when ($J = 3, K = 3$). The corresponding accuracy is (86.4%, 85.2%, 86.4%, 85.4%, 85.6%, 86.4%, 86.0%, 86.6%, 85.8%, 86.0%), which indicates that our method is little sensitive to the initialization.

Additionally, we also visualize the data generated through the estimated parameters. In order to display distributions of different categories more clearly, we randomly re-generate the synthetic data by enlarging the mean of distributions belonging to different classes. The Fig.3(a) shows the locations of the first two dimensional points of the synthetic data in the first type of features in the first view when $J = 3$ and $K_j = 3$. Inputting this data into our model, we can subsequently obtain the parameters $\boldsymbol{\theta}_Z$, $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_X$. Then these estimated parameters are exploited in our model to regenerate the five-class data, as shown in Fig.3(b). It is easy to see that the distributions of different categories generated according to the estimated parameters are quite similar to that in original data, which relatively substantiates the effectiveness and superiority of our hierarchical fusion model.

Biomedical Dataset: The accuracy as well as the sensitivity and specificity calculated by various strategies are tabulated in Tab.2. Note that, Sensitivity = TruePos./(TurePos.+FalsePos.) and Specificity = TrueNeg./(TrueNeg.+FalseNeg.). It is easy to observe that the proposed method always gets the better performance in classification compared with other approaches. In contrast to UMDL, SMDL and CSRL, HMMF is obviously superior. Particularly, HMMF achieves 82.0%, 82.7%, 83.2% and 84.9% in accuracy when the training number is 40, 50, 60 and 70, respectively, while the best results obtained by the aforementioned methods are only 77.4%, 74.7%, 79.1% and 79.6%. In comparison to DPL, RCR, MTJSRC and JDCR, HMMF is also competitive, gaining about 2% improvement in classification accuracy. For instance, when the training number is 40 or 60, the classifi-

Table 2: The accuracy, sensitivity and specificity values obtained by different methods on the Biomedical dataset when the number of training samples is 40, 50, 60, and 70, respectively. Best results are highlighted in bold.

	Number of Training Samples											
Num	num=40			num=50			num=60			num=70		
Methods	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
DPL	77.2%	75.4%	79.0%	79.9%	78.2%	81.5%	79.4%	77.7%	81.0%	82.4%	80.3%	84.4%
MTJSRC	79.0%	**82.2%**	76.0%	80.5%	80.6%	80.3%	81.0%	**86.0%**	76.1%	80.1%	82.1%	78.2%
RCR	78.6%	77.0%	80.3%	80.5%	80.3%	80.7%	80.9%	78.1%	83.6%	82.5%	80.6%,	84.4%
UMDL	77.4%	79.0%	75.8%	74.7%	73.4%	76.0%	79.1%	80.3%	78.0 %	79.6%	81.5%	77.8%
SDML	77.4%	79.0%	76.0%	73.6%	76.1%	71.2%	74.7%	70.4%	78.8 %	74.5%	75.9%	73.1%
JDCR	78.1%	75.8%	80.4%	80.1%	78.0%	82.0%	79.9%	78.8%	81.0 %	82.7%	80.7%	84.7%
CSRL	71.3%	73.5%	69.2%	72.2%	74.0%	70.5%	73.1%	77.9%	68.4%	75.4%	77.2%	73.7%
HMMF	**82.0%**	80.3%	**83.7%**	**82.7%**	**80.7%**	**84.6%**	**83.2%**	78.7%	**87.5%**	**84.9%**	**83.1%**	**86.6%**

Table 3: The area under curve (AUC) obtained by the different methods in DM detection.

Methods	AUC	Methods	AUC
DPL	0.8639	JDCR	0.8703
UMDL	0.8639	CSRL	0.8420
SMDL	0.8089	HMMF	**0.9027**

Table 4: The accuracy obtained by the different methods in the Wiki Text-Image dataset.

Method	DPL	MTJSCR	RCR	UMDL
Accuracy	65.1%	67.7%	66.1%	67.2%
Method	SMDL	JDCR	CSRL	HMMF
Accuracy	69.1%	68.5%	64.2%	71.1%

cation accuracy calculated by the proposed method reaches more than 2% improvement. Referring to the sensitivity and specificity, the presented model obtains superior values in most cases. For the specificity, the values gained by HMMF always outperforms that obtained by other comparison methods. For the sensitivity, although MTJSRC achieves a better performance our's when the training number is 40 and 60, our method arrives at the best point in other cases.

The ROC curves as well as their AUC values are further shown in Fig. 4 and Tab.3, respectively, when the training number is 70. From Fig. 4 we can see that the area covered by the ROC curve obtained by HMMF is remarkably larger than that calculated by SMDL and CSRL. In contrast to DPL, UMDL, and JDCR, HMMF also has the more or less improvement. From the Tab.3, it is easy to observe that our proposed hierarchical fusion model acquires higher AUC values. In comparison to SMDL and CSRL, HMMF achieves at least 6% improvement. Referring to the remaining strategies, HMMF is also much better. The area covered by HMMF is 0.9027, while the best result gained by these comparison strategies is only 0.8703.

Wiki Text-Image Dataset: The classification accuracy conducted on the Wiki Text-Image Dataset is listed in Tab. 4. Note that, we set the dimension of the latent variable to be 8 in our method. For CSRL, we set the dimension to be 10 since it achieves the best result in this case. It is easy to see that the proposed method achieves the best result compared with other approaches. In contrast to CSRL, DPL, RCR, MTJSRC and UMDL, HMMF gains a remarkable improvement. Compared with SMDL and JDCR, our strategy also obtains about 2.0% achievement. Particularly, the proposed method reaches as high as 71.1% in accuracy, while the best result obtained by SMDL and JDCR is only 69.1%.

Table 5: The accuracy obtained by HMMF with the change of the dimensionality of the latent variable.

Dimension	Accuracy	Dimension	Accuracy
D_j=1	44.0%	D_j=6	68.7%
D_j=2	53.4%	D_j=7	70.6%
D_j=3	59.0%	D_j=8	71.1%
D_j=4	64.4%	D_j=9	70.4%
D_j=5	66.8%	D_j=10	69.4%

The Tab. 5 further shows the accuracy with the changes of different dimensions of the latent variable. There is an obvious increase accuracy from 1 to 8 since a too low dimension of the latent variable would lose some valuable information. Subsequently, HMMF meets a slight performance degradation if D_j continues rising, indicating that a large dimensional subspace may introduce some information which does not contribute to the classification.

Conclusion

In this paper, a generative and hierarchical model is proposed for multi-view and multi-feature fusion. A latent variable is first learned for each view to fuse multiple features. The label information is also imposed on these variables across various views to jointly exploit the correlation among them. In this way, the multi-view and multi-feature data can be hierarchically and effectively fused. The EM algorithm is then applied to optimize the proposed method and a closed-form solution for each parameter is calculated. The experimental results on both synthetic and two real-world datasets substantiate the superiority of the presented method.

Acknowledgments

The work is partially supported by the NSFC fund (61332011, 61272292, 61271344, 61602540), Shenzhen Fundamental Research fund (JCYJ20150403161923528,

JCYJ20140508160910917), and the Science and Technology Development Fund (FDCT) of Macau 124/2014/A3.

References

Andrew, G.; Arora, R.; Bilmes, J. A.; and Livescu, K. 2013. Deep canonical correlation analysis. In *ICML (3)*, 1247–1255.

Archambeau, C., and Bach, F. R. 2009. Sparse probabilistic projections. In *Advances in neural information processing systems*, 73–80.

Bach, F. R., and Jordan, M. I. 2005. A probabilistic interpretation of canonical correlation analysis.

Bahrampour, S.; Nasrabadi, N. M.; Ray, A.; and Jenkins, W. K. 2016. Multimodal task-driven dictionary learning for image classification. *IEEE Transactions on Image Processing* 25(1):24–38.

Bishop, C. M. 2006. Pattern recognition. *Machine Learning* 128.

Ceci, M.; Pio, G.; Kuzmanovski, V.; and Džeroski, S. 2015. Semi-supervised multi-view learning for gene network reconstruction. *PloS one* 10(12):e0144031.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, 129–136. ACM.

Diethe, T.; Hardoon, D. R.; and Shawe-Taylor, J. 2010. Constructing nonlinear discriminants from multiple data views. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 328–343. Springer.

Eleftheriadis, S.; Rudovic, O.; and Pantic, M. 2015. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing* 24(1):189–204.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *Advances in neural information processing systems*, 793–801.

Guo, Y. 2013. Convex subspace representation learning from multi-view data. In *AAAI*, volume 1, 2.

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

Jing, X.-Y.; Hu, R.; Wu, F.; Chen, X.-L.; Liu, Q.; and Yao, Y.-F. 2014. Uncorrelated multi-view discrimination dictionary learning for recognition. In *AAAI*, 2787–2795.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 38(1):188–194.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, J.; Zhang, D.; Li, Y.; Wu, J.; and Zhang, B. 2016. Joint similar and specific learning for diabetes mellitus and impaired glucose regulation detection. *Information Sciences*.

Li, J.; Zhang, B.; and Zhang, D. 2017a. Joint discriminative and collaborative representation for fatty liver disease diagnosis. *Expert Systems with Applications*.

Li, J.; Zhang, B.; and Zhang, D. 2017b. Shared autoencoder gaussian process latent variable model for visual classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Nicolaou, M. A.; Panagakis, Y.; Zafeiriou, S.; and Pantic, M. 2014. Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1522–1526.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260. ACM.

Song, G.; Wang, S.; Huang, Q.; and Tian, Q. 2015. Similarity gaussian process latent variable model for multi-modal data analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, 4050–4058.

Tao, H.; Hou, C.; Nie, F.; Zhu, J.; and Yi, D. 2017. Scalable multi-view semi-supervised classification via adaptive regression. *IEEE Transactions on Image Processing* 26(9):4283–4296.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. A. 2015. On deep multi-view representation learning. In *ICML*, 1083–1092.

Yang, M.; Zhang, L.; Zhang, D.; and Wang, S. 2012. Relaxed collaborative representation for pattern classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2224–2231.

Yuan, X.-T.; Liu, X.; and Yan, S. 2012. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing* 21(10):4349–4360.

Zhang, L., and Zhang, D. 2016. Visual understanding via multi-feature shared learning with global consistency. *IEEE Transactions on Multimedia* 18(2):247–259.

Zheng, S.; Cai, X.; Ding, C. H.; Nie, F.; and Huang, H. 2015. A closed form solution to multi-view low-rank regression. In *AAAI*, 1973–1979.