Multimodal Similarity Gaussian Process Latent Variable Model

Guoli Song, Shuhui Wang, *Member, IEEE*, Qingming Huang, *Senior Member, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*— Data from real applications involve multiple modalities representing content with the same semantics from complementary aspects. However, relations among heterogeneous modalities are simply treated as observation-to-fit by existing work, and the parameterized modality specific mapping functions lack flexibility in directly adapting to the content divergence and semantic complicacy in multimodal data. In this paper, we build our work based on the Gaussian process latent variable model (GPLVM) to learn the non-parametric mapping functions and transform heterogeneous modalities into a shared latent space. We propose multimodal Similarity Gaussian Process latent variable model (m-SimGP), which learns the mapping functions between the intra-modal similarities and latent representation. We further propose multimodal distance-preserved similarity GPLVM (m-DSimGP) to preserve the intra-modal global similarity structure, and multimodal regularized similarity GPLVM (m-RSimGP) by encouraging similar/dissimilar points to be similar/dissimilar in the latent space. We propose m-DRSimGP, which combines the distance preservation in m-DSimGP and semantic preservation in m-RSimGP to learn the latent representation. The overall objective functions of the four models are solved by simple and scalable gradient decent techniques. They can be applied to various tasks to discover the nonlinear correlations and to obtain the comparable low-dimensional representation for heterogeneous modalities. On five widely used real-world data sets, our approaches outperform existing models on cross-modal content retrieval and multimodal classification.**

***Index Terms*— Multimodal learning, Gaussian processes, similarity preservation.**

Manuscript received September 24, 2016; revised March 4, 2017; accepted May 22, 2017. Date of publication June 7, 2017; date of current version June 23, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61672497, Grant 61332016, Grant 61620106009, Grant 61650202, Grant U1636214, Grant 61572488, and Grant 61429201, in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351802, and in part by the Key Research Program of Frontier Sciences of CAS under Grant QYZDJ-SSW-SYS013. The work of Q. Tian was supported by ARO under Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards by NEC Laboratories of America and Blippar. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christopher Wyatt. (*Corresponding author: Shuhui Wang*.)

G. Song and Q. Huang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China, with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: guoli.song@vipl.ict.ac.cn; qmhuang@ucas.ac.cn).

S. Wang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangshuhui@ict.ac.cn).

Q. Tian is with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249-1604 USA (e-mail: qi.tian@utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2713045

I. Introduction

DATA from real applications often involve multiple modalities representing content with the same semantics [1] and deliver rich information from complementary aspects. For example, users can write blogs to record their life, share photos with their friends, or find latent friends through tags. On video sharing websites, video content can also be learned from its associated textual descriptions such as user tags and comments. In knowledge sharing websites (*e.g.*, Wikipedia), an item page may include text, images and video to characterize the subject with different modalities. To leverage the information present in all the available modalities, learning from these multimodal data has become an important problem in many application areas [2]–[5]. A common solution is to learn the embedding function for multimodal data to obtain a shared latent space where we can align observations from different modalities and compare them directly.

One line of representative work is based on the generative model that provides intuitive probabilistic interpretation on multimodal correlation. In particular, Gaussian process latent variable models (GPLVMs) [6]–[9] have achieved great success in learning nonlinear low-dimensional embedding for multimodal data. Instead of specifying a set of deterministic (*e.g.*, CCA based methods [2], [10], [11]) or parametric (*e.g.*, univariate Gaussian [12]) mapping functions, a smooth non-parametric Gaussian process is defined in GPLVM on the probabilistic mapping from latent space to observation space. The flexibility of Gaussian process, determined by a variety of covariance functions, facilitates learning from real-world data with content divergence and complicated semantic relations. Despite that GPLVM better adapts to different modalities, there has been very few studies in describing multimodal relation using GPLVM.

We address cross-modal correlation learning with multimodal GPLVMs in this paper. In fact, GPLVMs discover the latent relation among multimodal data by introducing additive priors over latent space [7] or modality-specific covariance functions [8]. Given the latent representation, multimodal data are reconstructed by the learned Gaussian processes parameterized by covariance kernels. However, GPLVMs [6]–[8] suffer from high dimensionality of the multimodal data, which limit their applications on real-world problems.

One of the most critical problems of GPLVMs is that the topological structure in the observation space is not guaranteed to be preserved in the function embedding process, which leads to model degradation in processing high-dimensional multimodal data due to the *curse-of-dimensionality*. In existing

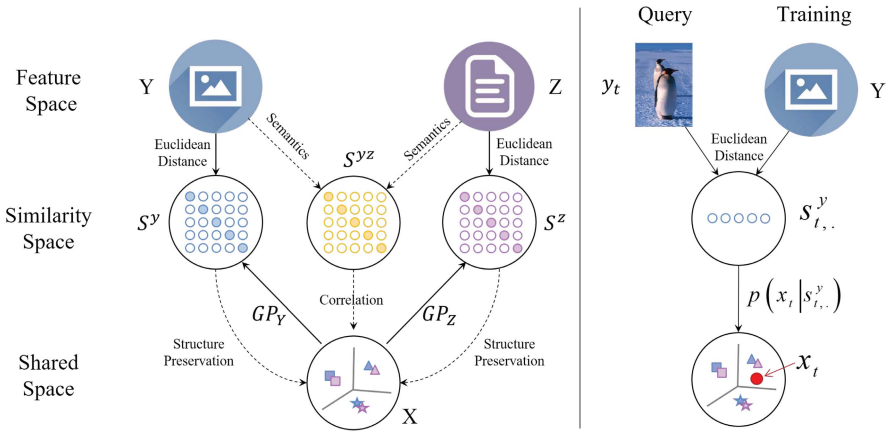


Fig. 1. The overview of the proposed similarity GPLVM. Left) Different modalities (*e.g.*, image and text) are first mapped from their respective natural spaces (Y and Z) to similarity spaces (S^y and S^z) by computing Euclidean distance. The shared space X is then learned using the framework of shared GPs. Constraints (dash lines) are placed over X to preserve cross-modal semantic correlation (S^{yz}) or intra-modal global similarity structure. Right) Inference for image query as an example. The corresponding latent position x_t for the query y_t is obtained by maximizing the posteriori probability $p(x_t|s^y_{t,.})$, where $s^y_{t,.}$ is the similarity matrix between y_t and the training images Y. Then cross-modal retrieval or classification of the query can be performed in the shared space.

study, similarity information is used to encode the modality-specific topological structure from perspectives of intra-modal content similarity [12], [13], context information [13] and semantic consistency [3]. It can be used as observation-to-fit [12] or mapping function regularizer [3] for learning a common latent space. To preserve the intra-modal topology, we construct Gaussian processes on the mappings between latent representations and multimodal observations at the similarity level rather than high-dimensional feature level [6]. To be specific, we propose to learn the nonlinear mapping functions between the intra-modal similarities and latent representations. From the maximum likelihood perspective, the nonlinear covariance matrices of multimodal mapping functions are learned to maximize the consistency to the modality-specific topologies. It better encodes the nonlinear semantic similarity in multimodal data. Compared to existing correlation models, our similarity-based GPLVM is robust to content divergence and high-dimensionality in multimodal representation.

In previous work, various constraints on the latent space have been devised in extensions of GPLVMs for specific purpose. For example, back-constraints [14] encodes the latent representation with the affinity information in the observation space. The locally linear embedding (LLE) [15] is combined with GPLVMs in [16] to preserve topological constraints. The discriminative shared-space prior [3], defined by a data-dependent weight matrix, enforces the topological structure preservation. In our similarity GPLVMs, we also impose specific constraints on the embedding to enhance the learning ability. In this paper, we consider two complementary constraints, as illustrated in Fig.1, from intra-modal similarity structure and inter-modal relationship perspectives.

First, different from existing local distance preservation models [14], [16], we intend to preserve the intra-modal global similarity structure, which is theoretically interpreted and can be easily unified to our similarity-based model. Our goal is to place each data object into the latent space such that the intra-modal similarities are maximally preserved. Motivated by the general idea of multidimensional scaling (MDS) [14], [17], where similarities among data objects are measured as distances among points in a low-dimensional space, we develop a distance-preserved constraint in each modality and add them into our similarity-based model. The constrained model makes full use of the intrinsic similarity information, and finds solution that encourages the divergence between the similarity structures of the original feature space and the latent space to be small. We call this constrained model multimodal Distance-preserved Similarity GPLVM (m-DSimGP).

Second, we explore inter-modal (dis)similarity relation in combination with the similarity-based model, while existing models only exploit simple inter-modal relation. For example, CCA-based models [2], [11] assume that the inter-modal relation is expressed by aligned multimodal data pairs. The inter-modal relation is also encoded as binary observation matrix to be fit [12], [18], [19]. The multimodal projections are learned to fit to observations of both intra-modal similarity and inter-modal relation among heterogeneous data objects [10], [12], [19]. In contrast, we directly impose two kinds of inter-modal relations (*i.e.*, semantic similarity and dissimilarity) as smooth priors on the output of multimodal similarity GPLVM. By using such regularization on the latent space, we enforce that the semantically similar/dissimilar cross-modal observations are also similar/dissimilar in the latent space, which provides goal-oriented solution to maximize the cross-modal semantic consistency. The proposed model is called multimodal Regularized Similarity GPLVM (m-RSimGP).

The outline of the proposed model is summarized in Fig.1. In particular, the following contributions have been made.

1) We propose multimodal Similarity Gaussian Process latent variable model (m-SimGP) for multimodal data analysis, which learns the non-parametric mapping functions between the intra-modal similarities and latent representation. It can be seen as a non-parametric generalization of the subspace-learning-based models.

2) We propose two general kinds of constraints over the latent representations: distance-preserved constraints for intra-modal global similarity structure preservation, and cross-modal similarity and dissimilarity constraints for inter-modal semantic consistency.
3) Two novel extensions of similarity GPLVM, *i.e.*, m-DSimGP and m-RSimGP, are presented by incorporating the proposed constraints into the m-SimGP model. Further, we derive a unified model m-DRSimGP, which utilizes the two complementary constraints. The conditional dependency among latent space and multimodal similarity observations can be easily learned with maximum a posteriori inference, and the overall objective functions can be solved by simple and scalable gradient decent techniques.
4) The proposed models can be applied to various tasks to discover the nonlinear correlations and obtain the comparable low-dimensional representation for heterogeneous modalities. On five widely used real-world multimodal datasets, we achieve significant improvement over the existing approaches in cross-modal content retrieval and classification tasks.

The remaining of this paper is organized as follows. Section II gives an overview of the related work. In Section III we review GPLVM and its multimodal extensions. In Section IV, we present the proposed Similarity Gaussian Process Latent Variable Model for multimodal data analysis. Section V presents extensive empirical evaluation on various datasets. Finally, Section VI concludes the paper.

II. Related Work

A. Multimodal Learning

The general problem for multimodal learning is how to bridge the heterogeneity gap between multimodal data for learning the correlations across different modalities. A wide variety of techniques have been developed for multimodal data analytic based on the modeling mechanism. One kind of representative works is to utilize latent variable models to learn a common semantic space for multimodal data. Based on different solution routines, these works can be roughly divided into statistical subspace learning models and probabilistic graphical models.

As the most typical subspace learning approach, canonical correlation analysis (CCA) based methods [2], [10], [11] project multimodal data into a shared subspace that guarantees different modalities are maximally correlated. In SCM [2], the CCA modeling is first applied to obtain two maximally correlated subspaces, and then logistic regressors are learned in each of these subspaces. GMA [11] is a supervised extension of CCA, which projects multiview data to a single subspace for cross-view classification and retrieval. There are some other statistical methods [4], [20], [21] that have also been proposed for multimodal learning problem. For example, the partial least squares (PLS) algorithm [22] is used to learn the common subspace by maximizing the covariance between different data modalities. Coupled dictionary learning (CDL) approaches [4], [23] have recently been proposed for multimodal data, which jointly learn a coupled dictionary as well as the corresponding mapping functions for different modalities.

The above subspace learning methods generally lack probabilistic interpretation to describe the semantic similarities. For probabilistic graphical modeling, representative works, *e.g.*, multimodal topic models, learn latent topics to describe the intrinsic semantic correlations in multimodal data. Based on latent Dirichlet allocation (LDA) [18], a variety of constraints are imposed. For example, mmLDA [24] enforces that all modalities share the same topic proportions, and corr-LDA [24] assumes one-to-one correspondence between the topics in each modality. To automatically determine the number of topics, non-parametric extensions of the multimodal topic models are proposed, which extends the Dirichlet prior by using a hierarchical Dirichlet process (HDP) [25] prior. Indian buffet process (IBP) is employed in [26] as the non-parametric prior distribution in Bayesian modeling for integrating multimodal data in a latent space. HDP and IBP model the prior distribution of the latent semantics in multimodal data analysis. In contrast, the Gaussian processes are used in multimodal GPLVMs [7]–[9] as prior probability distributions over the mappings from the shared space to data space. Due to the flexibility of Gaussian process, GPLVM-based methods can effectively obtain the informative latent representations and discover the nonlinear relationship among multimodal data with content divergence.

There is a close relationship between Gaussian processes and neural networks. Bayesian neural network models will converge to Gaussian processes in the limit of an infinite number of hidden units [27]. Due to the powerful representation learning ability [28], [29] of deep neural network models, recent works on multimodal embedding techniques [5], [30], [31] are based on deep learning. Deep extensions [32], [33] have also been developed for some shallow representation learning methods. For example, deep CCA [32] is developed to learn deep nonlinear mappings of two modalities that are maximally correlated. Inspired by deep neural networks, deep GP models [34], [35] are proposed by replacing every activation function in a Bayesian neural network with a Gaussian process transformation.

B. Dimensionality Reduction Methods

In the context of dimensionality reduction, the high-dimensional data $Y = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N\times d}$ is assumed to be generated by low-dimensional data $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N\times q}, q \ll d$ through a mapping function,

$$y_n = f(x_n) + \varepsilon_n, \tag{1}$$

where ε_n is the corrupting noise. The aim of dimensionality reduction is to preserve as much information from the original high-dimensional data as possible in the low-dimensional representation.

Dimensionality reduction methods can be categorized into linear methods and nonlinear methods. The linear methods assume that the mapping is a linear function, *e.g.*, $f(x_n) = Wx_n$, where $W \in \mathbb{R}^{d\times q}$ is the transformation weight matrix. Linear methods [10], [36] are simple and easy to implement. However, the linear assumption is overly smooth in preserving

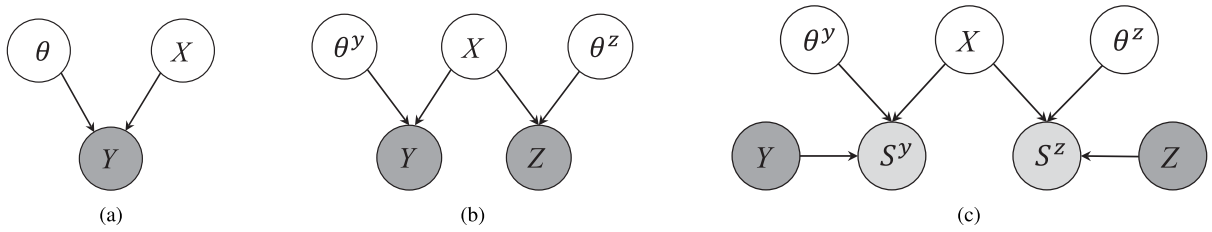


Fig. 2. GPLVMs *versus* the proposed m-SimGP. (a) is the original GPLVM proposed by Lawrence [6]. The observed data Y are assumed to be generated from a latent variable set X. (b) shows the shared GPLVMs for multimodal data. Two observed data modalities Y and Z are assumed to share the common latent space X. (c) is the proposed m-SimGP in this paper. To preserve the topological structure of each data modality, we build a multimodal latent variable model between the intra-modal similarities and latent space.

local structure in the latent space, so dimension reduction is natural to be extended to nonlinear mapping functions. A wide range of nonlinear methods [15], [37] have been suggested. It is possible for these methods to reflect the structure of the data through nonlinear embeddings. However, such methods may lack an intrinsic generative mechanism and do not explicitly include an assumption of underlying reduced data dimensionality [38].

Probabilistic dimensionality reduction approach [6] is to formulate latent variable models with smooth mappings from latent space to data space. If the mapping is chosen to be linear and the prior over the latent variables is taken to be Gaussian, the model is formulated as a probabilistic version of PCA [6]. GPLVM is a more recent nonlinear probabilistic dimensionality reduction method, which can be viewed as a nonlinear generalization of probabilistic PCA. Probabilistic models can be easily extended into a Bayesian framework for parameter learning, because of their generative nature and the forward mapping mechanism.

Our models can be recognized as the probabilistic dimension reduction on multimodal data. Compared to existing methods, our approaches gain more flexibility on real data with non-parametric and nonlinear mapping functions. By introducing the distance preservation and the semantic similarity constraints, we can reconstruct the structure and semantic relations more effectively with the generative mechanism over the learned latent representation.

III. Preliminary

In the GPLVM framework, the generative mapping function from latent to observation space is modeled using a Gaussian process as a prior distribution. As shown in Fig.2(a), the observation $Y \in \mathbb{R}^{N\times d}$ is assumed to be generated from a low dimensional latent space $X \in \mathbb{R}^{N\times q}$, through the mapping,

$$f(x) \sim \mathrm{GP}\left(\mu(x), k\left(x, x'\right)\right), \tag{2}$$

where the mean function $\mu(x)$ is typically taken to be zero for simplicity, and the covariance function $k\left(x, x'\right)$ is necessarily constrained to positive definite matrices.

The marginal likelihood of the observation Y with respect to the latent space X can be formulated as,

$$p(Y|X,\theta) = \frac{1}{\mathcal{A}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(K^{-1}YY^{\top}\right)\right), \tag{3}$$

where the normalization factor $\mathcal{A} = \sqrt{(2\pi)^{Nd}|K|^d}$, and $K \in \mathbb{R}^{N\times N}$ is the kernel matrix defined on X, *i.e.*, $K_{ij} = k(x_i, x_j)$. Considering that RBF kernel is simple and effective for high dimensional data [6], we use it with white noise as the covariance function,

$$k(x,x') = \sigma^2_{\mathrm{rbf}} \exp\left(-\frac{\left\|x-x'\right\|^2}{2l^2_{\mathrm{rbf}}}\right) + \sigma^2_{\mathrm{w}}\delta_{x,x'}, \tag{4}$$

where $\theta = \{\sigma^2_{\mathrm{rbf}}, \sigma^2_{\mathrm{w}}, l_{\mathrm{rbf}}\}$ denotes the parameters of the covariance matrix, which govern the variance of the RBF kernel, the variance of additive noise, and the RBF bandwidth, respectively.

In practice, a maximum a posteriori (MAP) probability estimation is used to learn the latent space X. The posterior distribution can be written as

$$p(X,\theta|Y) \propto p(Y|X,\theta)\, p(X). \tag{5}$$

Different forms of prior knowledge, *i.e.*, $p(X)$, can be introduced into the GPLVM to enhance the flexibility [6], [14], [39]. For example, spherical Gaussian prior [6] is used over the latent variables to enforce the smoothness of X and prevent the GPLVM from placing latent positions infinitely far apart.

GPLVM can be generalized to multimodal data, which are assumed to share a common latent space [8], [9], as shown in Fig.2(b). These models have achieved success in numerous applications, such as human pose estimation [7], tracking [40], facial expression recognition [3], *etc*. We apply GPLVM to multimodal data analysis for cross-modal correlation learning and classification.

IV. The Proposed Approach

The goal of our work is to discover a general latent representation shared by observations from multiple modalities. To achieve this, GPLVM is constructed on the modalities for its flexibility in probabilistic modeling on the conditional dependency between observation and latent space. In Section IV-A, we introduce multimodal similarity Gaussian process latent variable model, using similarity information in each data modality to preserve the intra-modal consistency. We further impose intra-modal distance constraint and cross-modal similarity/dissimilarity constraint on the latent space, respectively, which further enhances the model generality of our multimodal similarity GPLVM approach as presented in corresponding Section IV-B and IV-C.

Specifically, we consider a set of bi-modal data objects $\mathcal{O} = \{o_i\}_{i=1}^N$, each comprising of observations from two modalities, *i.e.*, $o_i = \{y_i, z_i\}$. Let $Y \in \mathbb{R}^{N\times d_y}$ and $Z \in \mathbb{R}^{N\times d_z}$ represent two data modalities, respectively. The objective is to relate these two modalities to the same latent space. In this work, Gaussian kernel is used to measure the intra-modal similarities for simplicity. Specifically, the similarity matrices $S^y \in \mathbb{R}^{N\times N}$ and $S^z \in \mathbb{R}^{N\times N}$ are defined as follows,

$$S^y\left(y_i, y_j\right) = \exp\left(-d^2\left(y_i, y_j\right)/2\gamma_y\right),$$
$$S^z\left(z_i, z_j\right) = \exp\left(-d^2\left(z_i, z_j\right)/2\gamma_z\right), \tag{6}$$

where $d\left(y_i, y_j\right) = \left\|y_i - y_j\right\|_2$ and $d\left(z_i, z_j\right) = \left\|z_i - z_j\right\|_2$. $\gamma_y, \gamma_z > 0$ are bandwidth parameters.

A. Multimodal Similarity GPLVM (m-SimGP)

As shown in Fig.2(c), we assume that the intra-modal similarity matrices S^y and S^z are generated from a shared q-dimensional latent manifold, where $q \ll \min\left(d_y, d_z\right)$. Each similarity matrix can be represented by the mappings with respect to a common latent space $X \in \mathbb{R}^{N\times q}$:

$$S_{ij}^y = f_{ij}^y\left(X\right) + \varepsilon_{ij}^y, \quad S_{ij}^z = f_{ij}^z\left(X\right) + \varepsilon_{ij}^z, \tag{7}$$

where $f_{ij}^y = f_j^y\left(x_i\right)$ and $f_{ij}^z = f_j^z\left(x_i\right)$ map the latent variable to the corresponding similarity. Each x_i generates the i-th row of S^y and S^z with f_j, $j = 1, \ldots, N$. The noise terms ε^y and ε^z are typically taken to be Gaussian with zero mean.

Similar as GPLVM, to find the latent representation X and the mappings $\{f_j^y\}_{j=1}^N$ and $\{f_j^z\}_{j=1}^N$, we place Gaussian process priors over the mappings:

$$f^y \sim GP\left(\mu^y\left(X\right), K^y(X, X)\right),$$
$$f^z \sim GP\left(\mu^z\left(X\right), K^z(X, X)\right). \tag{8}$$

As in section III, the mean functions are taken to be zero, and the covariance functions are generated by RBF kernel. The definition allows the mappings to be marginalized out analytically, and the marginal likelihood with respect to the latent variable can be computed as,

$$p\left(S^y, S^z \left| X, \theta^y, \theta^z\right.\right) = p\left(S^y \left| X, \theta^y\right.\right) p\left(S^z \left| X, \theta^z\right.\right), \tag{9}$$

$$p(S^y \left| X, \theta^y\right.) = \frac{1}{\mathcal{A}^y} \exp\left(-\frac{1}{2}\mathrm{tr}\left(K_y^{-1} S^y (S^y)^\top\right)\right), \tag{10}$$

$$p(S^z \left| X, \theta^z\right.) = \frac{1}{\mathcal{A}^z} \exp\left(-\frac{1}{2}\mathrm{tr}\left(K_z^{-1} S^z (S^z)^\top\right)\right). \tag{11}$$

If Gaussian prior is used over the latent variable, the objective function can be written as:

$$\arg\min_X \mathcal{L}^y + \mathcal{L}^z + \sum_{i=1}^N \frac{1}{2}\|x_i\|^2, \tag{12}$$

where $\mathcal{L}^y$ and $\mathcal{L}^z$ are the negative log-likelihood associated with Eq. (10) and (11), respectively. The model optimization can be solved by scaled conjugate gradient (SCG) technique [41].

In the proposed m-SimGP model, a simple spherical Gaussian prior is placed over the latent variable. For better preserving the intra-modal structure or cross-modal correlation among data observations, we can further impose more explicit constraints on the embeddings. In the following sections, we propose two kinds of better-behaved models through imposing restrictions on latent points.

B. Multimodal Distance-Preserved Similarity GPLVM (m-DSimGP)

In the embedding process of unimodal GPLVM, we aim to maximally preserve the intra-modal global similarity structure of the data. To be specific, the empirically similar objects are encouraged to be near to each other in the latent space, whereas those dissimilar objects are encouraged to be far apart. In this paper, we borrow the general idea of multidimensional scaling [14], [42] and measure similarities among pairs of objects as distances among points in the low-dimensional latent space. We bring a distance-preserved constraint for each modality by imposing restrictions on distance or similarity between data points, as shown in Fig.3(a).

Let $S^x \in \mathbb{R}^{N\times N}$ be a similarity matrix, where S_{ij}^x between latent variables x_i and x_j is computed according to

$$S^x(x_i, x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\gamma_x}), \tag{13}$$

where the bandwidth parameter γ_x is set to 1 for simplicity. We restrict the latent similarity matrix S^x to be in the proximity of the input similarity matrices S^y and S^z:

$$\left\|S^y - S^x\right\|_F^2 \le \rho_y, \quad \left\|S^z - S^x\right\|_F^2 \le \rho_z, \tag{14}$$

where the distance parameters $\rho_y, \rho_z > 0$, and $\|\cdot\|_F$ denotes the Frobenius norm. The norms exploit the global structure by preserving the pairwise sample similarity.

As functions of latent positions X, the norms in Eq. (14) can be easily unified to our similarity-based model due to its simple formulation. We replace the similarity constraints by penalty terms for the latent variable X, and combine them with the negative log-likelihood of the modalities Y and Z in Eq. (12). The new objective is formulated as:

$$\arg\min_X \mathcal{L}^y + \mathcal{L}^z + \mu_y \left\|S^y - S^x\right\|_F^2 + \mu_z \left\|S^z - S^x\right\|_F^2, \tag{15}$$

where μ_y and μ_z are the tradeoff parameters. The constrained model makes full use of the intrinsic similarity information. The regularization terms enforce that the searching for the latent variable should be in the vicinity of the input similarity matrices S^y and S^z. Latent positions are encouraged to make the divergence small between the similarities in the data and the latent space. We can solve the problem by gradient based optimization technique.

C. Multimodal Regularized Similarity GPLVM (m-RSimGP)

Nothing in GPLVM encourages the semantically dissimilar observations to be far in the latent space, nor semantically similar observations to be close in the latent space [39]. The learned latent space may not appropriately reflect the true cross-modal correlation in the observation space in the context

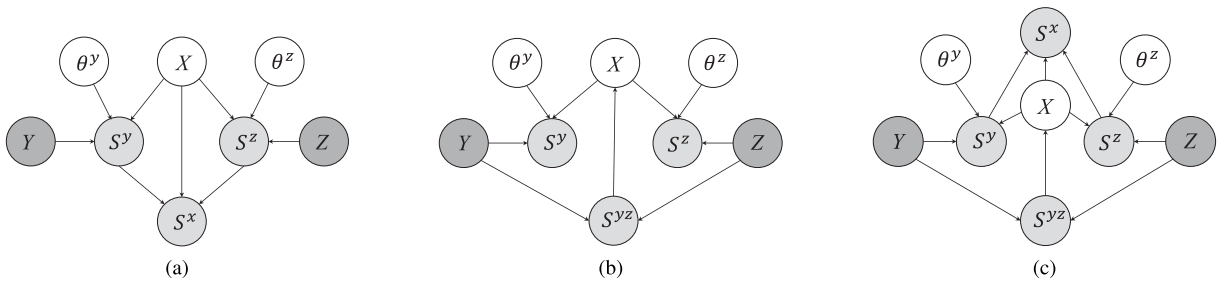


Fig. 3. Extensions of m-SimGP. (a) Multimodal Distance-Preserved Similarity GPLVM (m-DSimGP): The added constraints are from intra-modal similarity structure (S^y, S^z). (b) Multimodal Regularized Similarity GPLVM (m-RSimGP): The added constraints are from inter-modal semantic relationship (S^{yz}). (c) m-DRSimGP: The model consists of the proposed intra-modal and inter-modal constraints.

of multimodal correlation learning. To address this problem, we propose a regularized similarity GPLVM in this section.

Inspired by [43], where the topology preserving constraints regularize the latent space for multimodal distance metric learning, we aim to minimize the distance between similar data pairs and maximize the distance between dissimilar data pairs. We develop a multimodal regularized similarity Gaussian process latent variable model (m-RSimGP), where the prior characterized by a cross-modal similarity matrix is placed over the latent space, as shown in Fig.3(b).

Given a set of data objects $\mathscr{O} = \{o_i\}_{i=1}^N$ with two feature modalities y_i and z_i, the cross-modal similarity matrix $S^{yz} \in \{0,1\}^{N\times N}$ is defined as follows:

$$\left(S^{yz}\right)_{ij} = \begin{cases} 1, & \text{if } \left(o_i, o_j\right) \in \mathscr{S} \\ 0, & \text{if } \left(o_i, o_j\right) \in \mathscr{D}, \end{cases} \tag{16}$$

where $i, j = 1, 2, \ldots, N$. $\mathscr{S} = \{\left(o_i, o_j\right)\}$ denotes the set of pairs with similar semantics, and $\mathscr{D} = \{\left(o_i, o_j\right)\}$ denotes the set of pairs with dissimilar semantics. To make sure that semantically similar observations are close to each other and semantically dissimilar observations are far from each other in the embedded latent space, we impose the similarity and dissimilarity priors on the latent representation. The corresponding learning problem with respect to the latent variable X is formulated as follows:

$$\begin{aligned} &\min_X \ \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \left\|x_i - x_j\right\|^2 \\ &\text{s.t. } \left\|x_i - x_j\right\|^2 \geqslant 1, \quad \forall \left(o_i, o_j\right) \in \mathscr{D}, \end{aligned} \tag{17}$$

where x_i is the representation of the data point $o_i = \{y_i, z_i\}$ in the latent space. Euclidean distance is used as the distance measure for the embedded latent representation. The dissimilar points are separated by a margin of 1 in the latent space.

The optimization problem in Eq. (17) can be interpreted as a prior over the latent variable and combined with the likelihood maximization problem, where the smooth Gaussian prior constraint in Eq. (12) is substituted with the cross-modal similarity and dissimilarity constraints. As a result, our proposed m-RSimGP model is formulated as:

$$\begin{aligned} &\min_X \ \mathcal{L}^y + \mathcal{L}^z + \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \left\|x_i - x_j\right\|^2 \\ &\text{s.t. } \left\|x_i - x_j\right\|^2 \geqslant 1, \quad \forall \left(o_i, o_j\right) \in \mathscr{D}. \end{aligned} \tag{18}$$

The dissimilar constraints in Eq. (18) can be further relaxed with a convex hinge loss. Thus we obtain an unconstrained problem that is much easier to optimize:

$$\begin{aligned} \min_X \ \mathcal{L}^y + \mathcal{L}^z &+ \lambda_1 \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \left\|x_i - x_j\right\|^2 \\ &+ \lambda_2 \sum\nolimits_{(o_i,o_j)\in\mathscr{D}} \max\left(0, 1 - \left\|x_i - x_j\right\|^2\right), \end{aligned} \tag{19}$$

where λ_1 and λ_2 are the tradeoff parameters. They can be assigned with the same value, indicating equal importance of similar pairs and dissimilar pairs.

D. Multimodal Distance-Preserved Regularized GPLVM (m-DRSimGP)

As shown in Fig.3(c), we further combine the proposed two kinds of constraints to measure the interaction between them in the unified model. The distance constraints in Eq. (14) are introduced into the m-RSimGP model and combined with the cross-modal semantic regularization terms in Eq. (19). The new model is denoted as m-DRSimGP, and the resulting objective is formulated as:

$$\begin{aligned} \arg\min_X \ \mathcal{L}^y + \mathcal{L}^z &+ \mu_y \left\|S^y - S^x\right\|_F^2 + \mu_z \left\|S^z - S^x\right\|_F^2 \\ &+ \lambda_1 \sum\nolimits_{(o_i,o_j)\in\mathscr{S}} \left\|x_i - x_j\right\|^2 \\ &+ \lambda_2 \sum\nolimits_{(o_i,o_j)\in\mathscr{D}} \max\left(0, 1 - \left\|x_i - x_j\right\|^2\right), \end{aligned} \tag{20}$$

where μ_y, μ_z, λ_1 and λ_2 are the tradeoff parameters. The experiments show that these two different kinds of constraints can affect each other. For cross-modal correlation learning, the inter-modal similarity and dissimilarity constraints play a more important role than the intra-modal similarity constraints.

E. Generalization

When the observed data are coming from more than two modalities, the model can be readily extended. Take a new data modality $W \in \mathbb{R}^{N\times d_w}$ for example, our objective is to relate the three modalities (Y, Z, W) to the same latent space X. Then the new marginal likelihood with respect to the latent variable is $p\left(S^y, S^z, S^w \,|X, \theta^y, \theta^z, \theta^w\right) = p\left(S^y \,|X, \theta^y\right) p\left(S^z \,|X, \theta^z\right) p\left(S^w \,|X, \theta^w\right)$.

Similar to Eq. (10) and (11), the new modality-specific negative log-likelihood $\mathcal{L}^w$ can be computed and added to the objective function. For the m-DSimGP model, we can obtain the modality-specific similarity constraint as in Eq.(14), *i.e.*, $\|S^w - S^x\|_F^2 \leq \rho_w$. The m-RSimGP model can also be

easily extended and scaled to multiple modalities. For data with three different modalities, *e.g.*, $o_i = (y_i, z_i, w_i)$, the observations of different modalities share the common latent variables x_i. Therefore, the pairwise semantic relation is still applied between o_i and o_j, and the new cross-modal similarity matrix S^{yzw} can be obtained according to Eq. (16).

F. Optimization and Inference

It is obvious that the problems to be solved are highly non-linear functions of the latent variables, and there are no closed-form solutions. Therefore, we employ the SCG technique to optimize the objective functions. The optimization procedure for the m-RSimGP model is given in [44], and the solution for the m-DSimGP model can be obtained according to the similar procedure.

After the optimization procedure, we obtain the Gaussian processes for generating multimodal observations with the shared space X. When inferring the new set of observed test points, the inference procedure is straightforward in our solution framework. We take image observation as an example. The procedure for the text is similar. Given the image observation y_t, we can learn the corresponding latent representation x_t by maximizing the posteriori probability $p(x_t|s^y_{t,\cdot})$, where $s^y_{t,\cdot}$ is the similarity matrix between y_t and the training images Y according to Eq. (6).

V. Experiments

A. Datasets and Experimental Settings

Five popular multimodal datasets are used in experiments:

PASCAL Sentence [45] contains a total of 1000 images collected from 20 categories of PASCAL 2008. For each of the categories, 50 images are randomly selected. Each image is annotated with 5 sentences by Amazon Turkers. We use the same feature representation as in [46]. After SIFT features are extracted, each image is represented as a 1024-dim feature vector with the bag of visual words (BoVW) model. The text representation is based on 100-topic latent Dirichlet allocation (LDA) model. A random 70/30 split of the dataset is used for training/testing.

Wiki [2] is collected from Wikipedia consisting of 2,866 image-text documents. Each image is represented by a 128-dim bag-of-words based on SIFT descriptor and each text is represented by a 10-dim LDA feature. Totally 10 categories are considered and each document is labeled with one of them. A random 80/20 split of the dataset is used to produce a training set and a testing set.

TVGraz [47] contains 2,058 image-text pairs from 10 visual object categories of the Caltech-256 dataset. It is collected from webpages retrieved by Google image search for each of the 10 categories. We still use the same data provided by [46], where each image is represented by a 1024-dim feature vector based on SIFT BoVW, and the text is represented by a 100-dim LDA feature. The dataset is randomly divided into a training set of 1,558 documents and a test set of 500 documents.

NUS-WIDE-5.7K is a subset selected from NUS-WIDE dataset [48], consisting of 5730 paired objects. Each pair includes an image represented by a 500-dim bag-of-words based on SIFT descriptor and 1000-dim tag text. The class labels of image-text pairs are selected as the classes with the top-10 largest numbers of images. We randomly choose 85% of the data for training and the remaining 15% for testing.

Pascal VOC2007 [49] consists of 9963 image-tag pairs with 20 categories. We follow the same dataset used in [21] and [20], where the text is represented by a 399-dim word frequency feature, and each image is represented by a 512-dim GIST feature. In the experiment, the images containing only one object are selected. Finally, we obtain a training set with 2808 pairs and a test set with 2841 pairs.

Unless specified, we use the optimal settings of the parameters tuned by a parameter validation process for all the experiments. The bandwidth parameters of similarity matrices are set to 1, *i.e.*, $\gamma_y = \gamma_z = 1$. In all experiments, the tradeoff parameters μ_y and μ_z are assigned with the same value, indicating equal importance of two data modalities. The tradeoff parameters λ_1 and λ_2 are also assigned with the same value, indicating equal importance of similar and dissimilar semantic information. For each dataset, PCA [50] is used first to reduce the dimensions for textual and image features to 10 and 128, respectively. Then CCA [10] is used to obtain a low-dimensional initial representation of the latent space shared by two data modalities.

B. Image-Text Retrieval

Image-text retrieval is a typical cross-modal problem, consisting of two tasks: (1) image query *vs.* text database, (2) text query *vs.* image database. The observations of different modalities can be projected into the unified latent space X, and then cross-modal retrieval is performed by measuring the distance between their latent representations. A retrieved result is considered correct if it belongs to the same class as the query. We use 11-point interpolated precision-recall (PR) curve and mean average precision (MAP) [51] to measure the retrieval performance.

We compare our methods with the following baselines:

- **CCA** [10]: CCA learns a shared subspace by maximizing the correlation between the projections of images and text documents. Then the latent subspace is used to perform retrieval with standard distance based querying.
- **SCM** [2]: SCM first uses the CCA modeling to learn two maximally correlated subspaces, and then Logistic regressors are learned in each of these subspaces to produce the semantic mappings.
- **MLBE** [12]: As a generative model, MLBE uses binary hash codes as latent variables to generate intra-modal and inter-modal similarities. The code length for MLBE is set to 8 in our experiments.
- **GMLPP** [11]: GMLPP is the multiview extension of locality preserving projections (LPP), which is a linear dimensionality reduction algorithm with locality preserving quality.
- **DCCAE** [30]: DCCAE is a DNN-based multimodal feature learning algorithm which combines CCA [32] and autoencoder [28] based terms. In the experiment, the parameters for DCCAE are set according to [30], where

TABLE I
THE MAP COMPARISON FOR CROSS-MODAL RETRIEVAL TASK ON FIVE DATASETS. THE RESULTS SHOWN IN BOLDFACE ARE THE BEST

Dataset	Task	Method									Ours			
		CCA [10]	SCM [2]	MLBE [12]	GMLPP [11]	DCCAE [30]	LGCFL [21]	RL-PLS [20]	GCDL [4]	SGPLVM [7]	m-SimGP	m-DSimGP	m-RSimGP	m-DRSimGP
PASCAL	img-query	0.1938	0.2267	0.2543	0.1864	0.1988	0.2570	0.2140	0.2104	0.1591	0.2860	0.2903	0.3301	**0.3332**
	txt-query	0.1570	0.1730	0.2215	0.1567	0.1670	0.2379	0.1659	0.1787	0.1423	0.2818	0.2848	0.3275	**0.3318**
	avg	0.1754	0.1999	0.2379	0.1716	0.1829	0.2475	0.1900	0.1946	0.1507	0.2839	0.2876	0.3288	**0.3320**
Wiki	img-query	0.2453	0.2684	0.3787	0.2657	0.2542	0.2736	0.3087	0.2242	0.2054	0.4336	0.4470	**0.4697**	0.4690
	txt-query	0.2010	0.2276	0.4109	0.2056	0.1916	0.2241	0.2435	0.1823	0.1628	0.4188	0.4242	0.4418	**0.4880**
	avg	0.2232	0.2480	0.3948	0.2357	0.2229	0.2489	0.2761	0.2033	0.1841	0.4262	0.4356	0.4558	**0.4785**
TVGraz	img-query	0.3626	0.4280	0.3468	0.3175	0.3879	0.4366	**0.5737**	0.3413	0.3458	0.4467	0.4659	0.5102	0.5113
	txt-query	0.3334	0.3944	0.3849	0.4037	0.3736	0.4140	**0.5478**	0.3094	0.3205	0.4453	0.4667	0.5079	0.5373
	avg	0.3480	0.4112	0.3659	0.3606	0.3808	0.4253	**0.5608**	0.3254	0.3332	0.4460	0.4663	0.5091	0.5243
NUS-WIDE	img-query	0.2072	0.3282	0.2533	0.3029	0.4013	0.2911	0.4312	0.3415	0.2666	0.3384	0.4265	0.4304	**0.4363**
	txt-query	0.2003	0.2187	0.3232	0.2706	0.2804	0.2205	0.2932	0.2186	0.1429	0.3418	0.3941	0.4104	**0.4253**
	avg	0.2038	0.2735	0.2883	0.2868	0.3409	0.2558	0.3622	0.2801	0.2048	0.3401	0.4103	0.4204	**0.4308**
VOC2007	img-query	0.2835	0.4003	0.2281	0.2357	0.3302	0.3886	0.3869	0.2641	0.2360	0.4385	0.4483	0.5253	**0.5364**
	txt-query	0.2155	0.2869	0.3692	0.2183	0.2518	0.3161	0.2678	0.1926	0.1890	0.4374	0.4537	0.5291	**0.5350**
	avg	0.2495	0.3436	0.2987	0.2270	0.2910	0.3524	0.3274	0.2284	0.2125	0.4380	0.4510	0.5272	**0.5357**

feature mappings and reconstruction mappings are both implemented by networks of 3 hidden layers.

- **LGCFL** [21]: As a supervised cross-modal matching approach, LGCFL utilizes class labels to learn consistent feature representations from heterogeneous modalities, and introduces a local group based priori for better utilizing block based image features.
- **RL-PLS** [20]: RL-PLS takes the class label as the assistant modality, and builds two kernel PLS [52] models to project both images and text into the label space, which can reserve the label information and local structure at the same time.
- **GCDL** [4]: GCDL learns coupled dictionaries for the two modalities such that the transformed sparse coefficients of the same class are maximally correlated and they are also discriminative enough to be used for cross-modal matching.
- **SGPLVM** [7]: SGPLVM uses back-constraints to learn a shared latent representation that captures the correlations among different modalities.

Table I summarizes the performance results of the cross-modal retrieval in terms of MAP over PASCAL Sentence, Wiki, TVGraz, NUS-WIDE-5.7K, and Pascal VOC2007, respectively. First, we can see that our similarity based GPLVMs are much better than other methods in overall performance. Compared to SGPLVM and GMLPP, m-SimGP gains significant performance improvement, which indicates that similarity information is important in capturing the correlation structure of multimodal data. It is clear that our non-parametric models outperform the parametric MLBE model. To be specific, the best performance achieved by m-RSimGP outperforms MLBE by 15% higher MAP on the Wiki dataset. Compared to our models, MLBE models the dependence between similarity observations and the latent variables by univariate Gaussian mapping functions, which shows the effectiveness of Gaussian process in discovering the nonlinear relationship among multimodal data. Our methods also give better performance than the deterministic CCA-based models (CCA, SCM, DCCAE). For example, the performance achieved by m-SimGP outperforms SCM by 42% higher MAP on the PASCAL dataset. DCCAE performs badly for all the tasks except the image-to-text retrieval on NUS-WIDE, because that DNN-based models are appropriate for relatively large datasets. On most of the datasets, our unsupervised m-SimGP and m-DSimGP outperform the supervised algorithms (LGCFL, RL-PLS, GCDL) that utilize class label information to reduce the semantic gap across modalities. For TVGraz, our proposed supervised approaches (*i.e.*, m-RSimGP and m-DRSimGP) achieve a comparable performance. The main reason for the outstanding performance of RL-PLS is that the authors introduce a complex label space using real values instead of the simple binary class labels. Moreover, we see that TVGraz is a less challenging dataset than the other four datasets because all the algorithms achieve relative high MAP scores on TVGraz.

Second, the m-DSimGP method outperforms m-SimGP on all the datasets for all the retrieval tasks. It shows that the distance constraints in (14) are helpful for latent variable modeling. Besides, m-DRSimGP gains performance improvement compared to m-RSimGP on almost all the tasks, which further shows the proposed similarity constraints can facilitate the cross-modal correlation learning. Third, we can see m-RSimGP achieves the best performance compared to our m-SimGP and m-DSimGP models. It shows that the cross-modal similarity and dissimilarity constraints in Eq. (17) over the latent variables contribute significantly to the cross-modal correlation learning.

Fig. 4 shows the PR curves on all the datasets demonstrating the promising performance of our methods. Fig.4(a-1) (a-2) and Fig. 4(e-1) (e-2) show that our methods, especially m-RSimGP and m-DRSimGP, perform consistently better than any other methods on PASCAL and VOC2007, respectively. For TVGraz dataset, it can be observed from Fig. 4(c-1) and (c-2) that m-RSimGP and m-DRSimGP outperform all the other methods except RL-PLS.

On Wiki dataset, Fig.4(b-1) and (b-2) show that our methods achieve significant improvement over other methods except MLBE for both retrieval tasks. However, MLBE

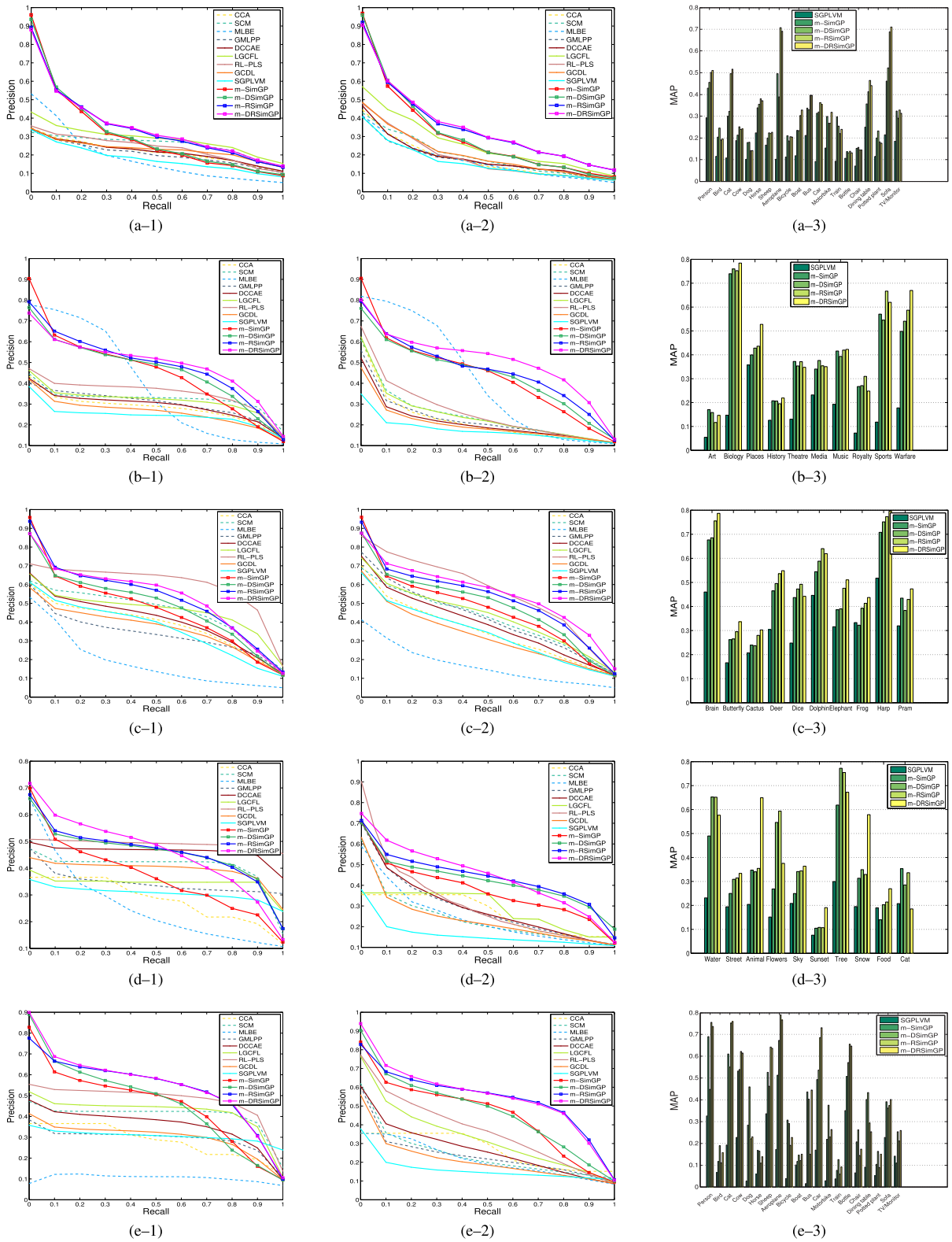


Fig. 4. Precision-Recall curves of cross-modal retrieval using both image and text queries, and average per-class MAP scores across image and text queries. (a-1) PASCAL: PR curves for image query (a-2) PASCAL: PR curves for text query (a-3) PASCAL: average per-class map (b-1) Wiki: PR curves for image query (b-2) Wiki: PR curves for text query (b-3) Wiki: average per-class map (c-1) TVGraz: PR curves for image query (c-2) TVGraz: PR curves for text query (c-3) TVGraz: average per-class map (d-1) NUS-WIDE: PR curves for image query (d-2) NUS-WIDE: PR curves for text query (d-3) NUS-WIDE: average per-class map (e-1) VOC2007: PR curves for image query (e-2) VOC2007: PR curves for text query (e-3) VOC2007: average per-class map.

achieves better precision at low recall rates compared to our methods. As we can see from the PR curves of PASCAL, TVGraz, NUS-WIDE and VOC2007, it is in fact much more difficult for MLBE to obtain good precision rate when detecting more relevant documents. We attribute this to its intrinsic property. As a parametric model, MLBE is pre-specified with

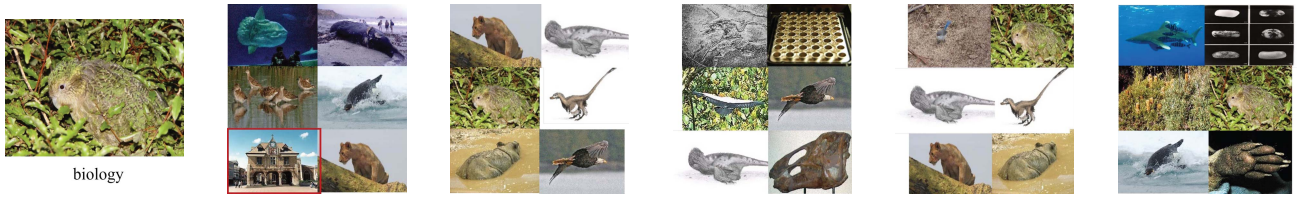


Fig. 5. Image-to-text retrieval on Wiki. (a) The image query. Here we present images that corresponding to the top retrieved texts. Red border indicates a false positive. (b) SGPLVM (c) m-SimGP (d) m-DSimGP (e) m-RSimGP (f) m-DRSimGP.

Fig. 6. Text-to-image retrieval on Wiki. (a) The text query and ground-truth image. Here we present six retrieved relevant images. Red border indicates a false positive. (b) SGPLVM (c) m-SimGP (d) m-DSimGP (e) m-RSimGP (f) m-DRSimGP.

the nearly optimal latent feature dimension for retrieving documents that are related specifically to the query. However, the parametric MLBE lacks the ability to widen the scope of matches, and thus tends to achieve low precision at high recall levels.

On NUS-WIDE-5.7K dataset, the PR curves shown in Fig.4(d-2) indicate that our methods perform much better than any other baseline methods for the text-to-image retrieval. For the image-to-text retrieval task, as shown in Fig.4(d-1), the PR curves of m-RSimGP and m-DRSimGP clearly outperform other methods except DCCAE and RL-PLS, further verifying the results shown in Table I. It can be observed that our methods achieve higher precision than DCCAE and RL-PLS at low recall, which is more applicable in practice. Fig.4(d-1) and (d-2) also show that m-RSimGP achieves the best performance compared with the proposed m-SimGP and m-DSimGP. However, it requires a more careful study of the corresponding PR curves to compare the effectiveness of m-RSimGP and m-DRSimGP models. This observation is consistent with the MAP results shown in Table I, where the corresponding MAP scores for m-RSimGP and m-DRSimGP are close to each other, especially in image-to-text retrieval. Fig. 4 also shows the per-class MAP scores of our methods compared to SGPLVM. For all the datasets, our four methods have higher MAP scores than the GPLVM-based baseline SGPLVM on almost all the classes, and m-DRSimGP has the best overall performance.

Our methods consistently achieve promising performance on both retrieval tasks, which verifies the effectiveness of our methods in reducing the semantic gap between modalities. As shown in Table I, other latent variable models either achieve better MAP performance of image query (*e.g.*, SGPLVM) or better MAP of text query (*e.g.*, MLBE). For our methods, the MAP scores of both retrieval tasks are pretty close to each other. Therefore, our models can better achieve the semantic consistency among cross-modal data and the learned latent representation can better reflect the cross-modal correlation in the observation space.

Finally, we show some examples of cross-modal retrieval on Wiki. Fig. 5 shows an example of image-to-text retrieval. We use the corresponding images of the retrieved texts to demonstrate the results. We see that the retrieved results by our methods are all about “biology", but the fifth retrieved text of SGPLVM comes from the “geography" category. For the example of text-to-image, the query text is presented with its corresponding image. Fig. 6 shows some of the top retrieved images by SGPLVM are from other categories. However, all of the retrieved images by our methods are from the “warfare" category same as the query text, and our m-DRSimGP can retrieve the ground-truth image of the query text.

C. Classification

Our work aims to discover a general latent representation shared by multimodal observations. Therefore, the resulting posterior of our framework is the latent space instead of the class information. In other words, the classification problem is not directly modeled in our methods. To obtain the class prediction, we apply a classifier to the learned latent space. In our experiments, classification is accomplished by using the k-nearest neighbor (k-NN) classifier to find the closest latent representation to the test data.

The proposed models are compared to 1-NN, CCA, SGPLVM, Discriminative GPLVM (D-GPLVM) [39] and Discriminative Shared GPLVM (DS-GPLVM) [3]. As a single-view method, D-GPLVM restricts the latent space with a prior based on Linear Discriminant Analysis (LDA). In our experiments, we extend it to learn from multimodal observations. DS-GPLVM generalizes the Gaussian Markov Random Field (GMRF) prior for single view to multiview learning. In [3], DS-GPLVM is performed in two scenarios for inference. In the first, each modality is independently back-projected to the

TABLE II
AVERAGE CLASSIFICATION ACCURACY ON FIVE DATASETS. THE RESULTS SHOWN IN BOLDFACE ARE THE BEST

Datasets	Methods									
	1-NN	CCA [10]	SGPLVM [7]	D-GPLVM [39]	DS-IBP [3]	DS-SBP [3]	m-SimGP	m-DSimGP	m-RSimGP	m-DRSimGP
PASCAL	0.1767	0.1267	0.1700	0.1833	0.1229	0.5067	0.4833	0.4667	**0.5233**	0.5200
Wiki	0.1746	0.1948	0.1457	0.1934	0.1499	**0.6921**	0.6003	0.5382	0.6652	0.6205
TVGraz	0.4580	0.4540	0.4740	0.4500	0.5020	0.6840	0.6360	0.6100	**0.6980**	0.6380
NUS-WIDE	0.1956	0.1713	0.2072	0.1562	0.1988	0.6898	0.6765	0.6667	**0.7396**	0.7025
VOC2007	0.2665	0.2288	0.2675	0.2763	0.2087	0.7378	0.6352	0.6238	**0.7543**	0.6900

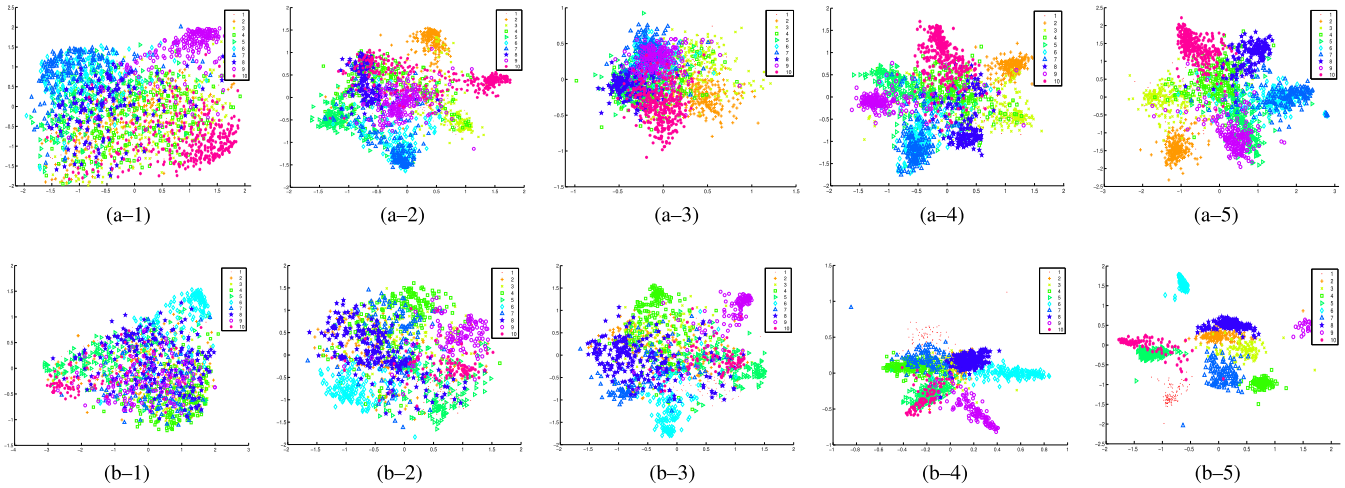


Fig. 7. Visualization of the latent representations discovered by (a-1) SGPLVM (a-2) m-SimGP (a-3) m-DSimGP (a-4) m-RSimGP (a-5) m-DRSimGP on the Wiki dataset, and (b-1) SGPLVM (b-2) m-SimGP (b-3) m-DSimGP (b-4) m-RSimGP (b-5) m-DRSimGP on the TVGraz dataset. The data points with the same colors and shapes indicate that they come from the same category (Better viewed in color).

latent space, where the back-constraints are defined on each modality separately. In the second, a single back-projection to the latent space is performed for all the modalities, where the back-constraint is defined on the set of all the modalities. We denote the former approach as DS-IBP and the latter one as DS-SBP. We build 1-NN classifier baseline in the image feature space. In the testing stage, 1-NN classifier is applied to the learned latent space to obtain the prediction results.

Table II presents the average classification accuracy on all the five datasets. The results show that our models can effectively learn a discriminative latent space from multimodal observations. We can see that our methods are either comparable or better than other methods. Though DS-IBP and D-GPLVM are designed for the classification problem, their poor performance reflects that these two methods lack the ability to effectively capture the semantically consistent representation of multimodal data. Since the major difference between DS-IBP and DS-SBP is the pattern of back-projection, we attribute the superior performance of DS-SBP to the fact that DS-SBP back-projects complementary information from all the data modalities during the inference process. Different from DS-SBP, the inference procedure of our methods is much simpler, where only the image information is used to estimate a posteriori to infer the testing latent representations.

Furthermore, we can see that m-RSimGP achieves the best classification accuracy among our proposed methods on all the datasets. The proposed m-DSimGP and m-DRSimGP performs relatively worse than the other two models, which indicates that the distance constraints in Eq.(14) do not take an active part in class prediction. Different from the retrieval task, as we know, classification needs labels and thus is not a purely distance-based approach. Therefore, m-DSimGP and m-DRSimGP with heavy emphasis on global structure tends to weaken the discriminative manifold learning, which requires the same-class samples to be close and those of different classes to be far. And yet, the m-RSimGP model can achieve better classification performance by enforcing that the semantically similar/dissimilar cross-modal observations are also similar/dissimilar in the latent space.

D. The Latent Representation Visualization

In this section, we visualize the discovered latent representation. The experiments are performed on Wiki and TVGraz, each with 10 categories. On both datasets, the 10-dim latent representations are embedded into a 2-dim space using the t-SNE [53] algorithm for visualization. The results show that the proposed models perform much better on producing a low-dimensional embedding compared to the original SGPLVM. As shown in Fig. 7, the latent representations discovered by SGPLVM provide little insight into the category structure of the data objects. In contrast, the latent representations

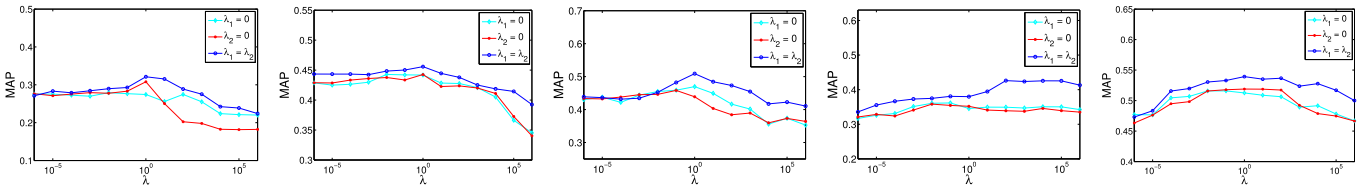


Fig. 8. Sensitivity test on the tradeoff parameters in m-RSimGP *w.r.t.* the performance of image-text retrieval. Here we denote the parameter variables λ_1 and λ_2 as λ for simplicity. (a) PASCAL (b) Wiki (c) TVGraz (d) NUS-WIDE (e) VOC2007.

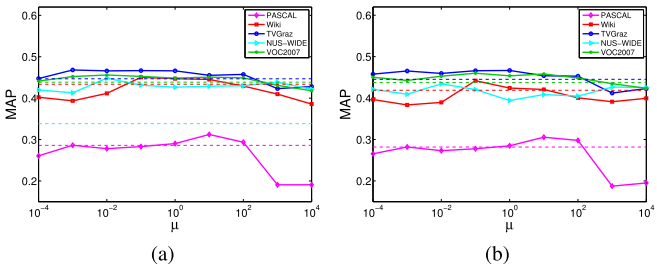


Fig. 9. Sensitivity test on the tradeoff parameters in m-DSimGP w.r.t. the performance of image-text retrieval. (a) Image to text (b) Text to image. The dashdot line shows the best results obtained by m-SimGP.

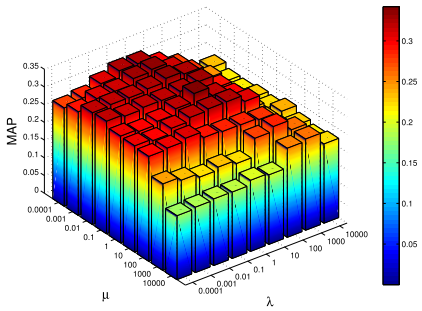


Fig. 10. Sensitivity test on the tradeoff parameters in m-DRSimGP *w.r.t.* the performance of image-text retrieval: An example on the PASCAL dataset.

discovered by our models exhibit a more clear grouping pattern for the data from the same category. Therefore, our similarity-based GPLVMs are capable of discovering the discriminative information among multimodal data.

E. Parameter Sensitivity Analysis

1) μ: In our work, the tradeoff parameters μ_y and μ_z in (15) are assigned with the same value, *i.e.* $\mu_y = \mu_z = \mu$, indicating equal importance of the observation modalities. We conduct sensitivity analysis on them to test how they impact the performance of the proposed m-DSimGP model. Fig.9 shows the curves of average MAP scores for image-text retrieval with different tradeoff parameters.

It can be seen that m-DSimGP can achieve superior performance over m-SimGP under some range of the parameter value. For example, m-DSimGP performs better on PASCAL when the value of the parameter μ is limited to $[10^0, 10^2]$, and for Wiki the best interval of the parameter value is $[10^{-1}, 10^1]$. For TVGraz and VOC2007, the performance decreases as the value of μ increases to 10^2 for both retrieval tasks. However, m-DSimGP is robust to the change of the parameter on the NUS-WIDE-5.7K dataset, and it performs consistently better than m-SimGP. The figure shows that the constraints on the local structure in (15) are helpful for latent variable modeling, but overly strong constraints may cause overfitting problem especially for the small-scale datasets. For consistency, we fix $\mu_y = \mu_z = 1$ for m-DSimGP in all the experiments.

2) λ_1 and λ_2 : We conduct sensitivity analysis on the tradeoff parameters λ_1 and λ_2 in (19) to test how they impact the cross-modal correlation learning. Fig.8 shows the curves of average MAP scores of image-text retrieval with different tradeoff parameters. We consider three different settings: (1) λ_1 is fixed to 0, (2) λ_2 is fixed to 0, (3) λ_1 and λ_2 are set to the same value. For simplicity, the parameter variables are denoted as λ, as shown in Fig.8.

We can observe that both similar and dissimilar semantic information have great impact on the performance of the m-RSimGP model. The curves of λ_1 and λ_2 are pretty similar to each other, which indicates their equal importance in cross-modal correlation learning. In all the settings, the average MAP is low for small λ, and it is improved by increasing λ. The performance is much better, when λ is increased to 10^0 or 10^1. Fig.8 clearly shows that the average MAP performance is on a downward trend, *e.g.*, the average MAP score on PASCAL decreases significantly when λ reaches 10^1. These phenomenons are possibly due to the fact that an overly large λ will improve the risk of overfitting and cause the model to become trapped in a local minima. Taken together, λ_1 and λ_2 are set to 1 for m-RSimGP in all the experiments.

3) λ and μ : We also conduct sensitivity analysis on the tradeoff parameters in (20) to evaluate the two kinds of constraints in the m-DRSimGP model. The m-DRSimGP performance is tested on the PASCAL dataset in the case of cross-modal retrieval. We assume the observation modalities are equally important, and both similar and dissimilar semantic information are present in Eq. (20), *i.e.*, $\mu_y = \mu_z = \mu$, and $\lambda_1 = \lambda_2 = \lambda$. We have used 9 different values of μ and λ, resulting in a total of 81 pairs of (μ, λ), as shown in Fig.10.

Seen from Fig.10, m-DRSimGP can achieve consistently good performance as long as the value of μ or λ is not too large. The best performance is achieved when μ and λ are increased to around 10^0. Therefore, the proposed two kinds of constraints, *i.e.*, intra-modal distance constraints and inter-modal semantic constraints, are helpful for the m-DRSimGP model in cross-modal correlation learning. Overall, the performance is slightly better when λ is set to a larger value

than μ. Again, it indicates that the inter-modal similarity and dissimilarity constraints play a more important role than the intra-modal distance constraints. For all the experiments, we fix the tradeoff parameters of m-DRSimGP and also set $\mu = \lambda = 1$ on the other four datasets.

VI. Conclusion

Most existing models [10], [12], [19] assume that both inter-modal relation and intra-modal relation are independent. In this paper, we assume that the intra-modal relation is conditionally independent to each other given the latent representation. The intra-modal similarities are sampled from multimodal Gaussian processes determined by the modality-specific covariance functions on the latent representation. Accordingly, we extend the similarity GPLVM [44] by exploiting the relationship between latent output space and mutimodal similarity input space. For better preserving the nonlinear similarity structure, we construct m-DSimGP by putting restrictions on intra-modal similarity structure. For encoding semantic similarity in multimodal data, we come up with m-RSimGP by forcing similar/dissimilar points to be similar/dissimilar in the latent space. Then these two extensions are combined to formulate m-DRSimGP by constraining both intra-modal and inter-modal similarity correlation. The proposed models can be applied to various tasks to discover the nonlinear correlations and obtain the comparable low-dimensional representation for heterogeneous modalities. In future work, we will investigate on constructing hierarchical/deep structure for latent variable model [34], [35], [54] to better capture the intrinsic semantic consistency of heterogeneous modalities. We will also address the problems of correspondence missing and information imbalance in real-world data based on our similarity-based GPLVM.

References

[1] N. Chen, J. Zhu, F. Sun, and E. P. Xing, "Large-margin predictive latent subspace learning for multiview data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2365–2378, Dec. 2012.

[2] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 251–260.

[3] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[4] D. Mandal and S. Biswas, "Generalized coupled dictionary learning approach with applications to cross-modal matching," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3826–3837, Aug. 2016.

[5] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3441–3450.

[6] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, Nov. 2005.

[7] C. H. Ek, J. Rihan, P. H. S. Torr, G. Rogez, and N. D. Lawrence, "Ambiguity modeling in latent spaces," in *Machine Learning for Multimodal Interaction*. Utrecht, The Netherlands: Springer, 2008, pp. 62–73.

[8] A. C. Damianou, C. H. Ek, M. Titsias, and N. D. Lawrence, "Manifold relevance determination," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2012, pp. 145–152.

[9] A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao, "Learning shared latent structure for image synthesis and robotic imitation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1233–1240.

[10] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[11] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.

[12] Y. Zhen and D. Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 940–948.

[13] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1577–1584.

[14] N. D. Lawrence and J. Quiñonero-Candela, "Local distance preservation in the GP-LVM through back constraints," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 513–520.

[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[16] R. Urtasun, D. J. Fleet, A. Geiger, J. Popović, T. J. Darrell, and N. D. Lawrence, "Topologically-constrained latent variable models," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 1080–1087.

[17] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. New York, NY, USA: Springer, 2005.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[19] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2407–2414.

[20] J. He, B. Ma, S. Wang, Y. Liu, and Q. Huang, "Cross-modal retrieval by real label partial least squares," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 227–231.

[21] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[22] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. Bohinj, Slovenia: Springer, 2005, pp. 34–51.

[23] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2496–2503.

[24] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th ACM Int. SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 127–134.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[26] B. Ozdemir and L. S. Davis, "A probabilistic framework for multimodal retrieval using integrative Indian buffet process," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2384–2392.

[27] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. New York, NY, USA: Springer, 2012.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[29] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.*, to be published, doi:10.1109/TFUZZ.2016.2574915.

[30] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multiview representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[31] S. Rastegar, M. Soleymani, H. R. Rabiee, and S. M. Shojaee, "MDL-CW: A multimodal deep learning framework with cross weights," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2601–2609.

[32] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

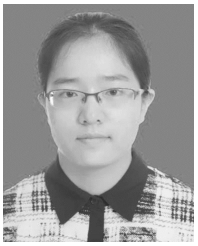
[33] Y. Deng, F. Bao, X. Deng, R. Wang, Y. Kong, and Q. Dai, "Deep and structured robust information theoretic learning for image analysis," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4209–4221, Jul. 2016.

[34] A. C. Damianou and N. D. Lawrence, "Deep Gaussian processes," in *Artificial Intelligence and Statistics*. Scottsdale, AZ, USA: AISTATS, 2013, pp. 207–215.

[35] Z. Dai, A. Damianou, J. González, and N. D. Lawrence. (2015). "Variational auto-encoded deep Gaussian processes." [Online]. Available: https://arxiv.org/abs/1511.06455

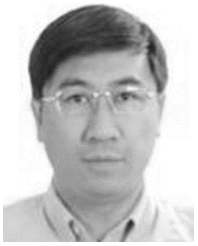
[36] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Francisco, CA, USA: Academic, 2013.

[37] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[38] A. C. Damianou, M. K. Titsias, and N. D. Lawrence, "Variational inference for latent variables and uncertain inputs in Gaussian processes," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1425–1486, 2015.
[39] R. Urtasun and T. Darrell, "Discriminative Gaussian process latent variable model for classification," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 927–934.
[40] V. A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2185–2192.
[41] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, no. 4, pp. 525–533, Nov. 1993, doi: 10.1016/S0893-6080(05)80056-5.
[42] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952.
[43] P. Xie and E. P. Xing, "Multi-modal distance metric learning," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1806–1812.
[44] G. Song, S. Wang, Q. Huang, and Q. Tian, "Similarity Gaussian process latent variable model for multi-modal data analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4050–4058.
[45] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical Turk," in *Proc. NAACL HLT Workshop*, 2010, pp. 139–147.
[46] J. C. Pereira and N. Vasconcelos, "On the regularization of image semantics by modal expansion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3093–3099.
[47] I. Khan, A. Saffari, and H. Bischof, "TVGraz: Multi-modal learning of object categories by combining textual and visual features," in *Proc. AAPR Workshop*, 2009, pp. 213–224.
[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 48.
[49] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). *The Pascal Visual Object Classes Challenge 2007 (voc2007) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[50] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Edu. Psychol.*, vol. 24, no. 6, p. 417–441, 1933.
[51] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[52] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Mar. 2001.
[53] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[54] N. D. Lawrence and A. J. Moore, "Hierarchical Gaussian process latent variable models," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 481–488.



Guoli Song received the B.S. degree in mathematics and applied mathematics, the M.S. degree in operational research and cybernetics from Zhengzhou University, in 2009 and 2012, respectively. She is currently pursuing the Ph.D. degree with the School of Computer and Control Engineering, University of Chinese Academy of Sciences. Her current research interests include machine learning and cross-media information retrieval.

Shuhui Wang received the B.S. degree in electronics engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2012. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His current research interests include semantic image analysis, image and video retrieval and large-scale Web multimedia data mining.



Qingming Huang received the B.S. degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor and the Deputy Dean with the School of Computer and Control Engineering, University of Chinese Academy of Sciences. He has authored over 300 academic papers in international journals, such as the IEEE TIP, the TKDE, the TMM and the TCSVT, and top level international conferences, including the ACM Multimedia, the ICCV, the CVPR, the ECCV, the VLDB, the AAAI, and the IJCAI. His current research interests include multimedia computing, image/video processing, pattern recognition, and computer vision.

Qi Tian (F'16) received the B.E. degree in electronic engineering from Tsinghua University in 1992 and the M.S. degree in ECE from Drexel University in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign (UIUC) in 2002. He was a Tenured Associate Professor from 2008 to 2012 and a Tenure-Track Assistant Professor from 2002 to 2008. From 2008 to 2009, he took one-year Faculty Leave at Microsoft Research Asia as a Lead Researcher in the Media Computing Group. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA). and His research interests include multimedia information retrieval, computer vision, pattern recognition and bioinformatics and published over 370 refereed journal and conference papers.

Dr. Tian received the 2017 UTSA Presidents Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from the College of Science, UTSA. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, the *Multimedia System Journal*, and in the Editorial Board of the *Journal of Multimedia*, and the *Journal of Machine Vision and Applications*. He is a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the *Journal of Computer Vision and Image Understanding*.