Deep Coupled Metric Learning for Cross-Modal Matching

Venice Erin Liong, Jiwen Lu, *Senior Member, IEEE*, Yap-Peng Tan, *Senior Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new deep coupled metric learning (DCML) method for cross-modal matching, which aims to match samples captured from two different modalities (e.g., texts versus images, visible versus near infrared images). Unlike existing cross-modal matching methods which learn a linear common space to reduce the modality gap, our DCML designs two feedforward neural networks which learn two sets of hierarchical nonlinear transformations (one set for each modality) to nonlinearly map samples from different modalities into a shared latent feature subspace, under which the intraclass variation is minimized and the interclass variation is maximized, and the difference of each data pair captured from two modalities of the same class is minimized, respectively. Experimental results on four different cross-modal matching datasets validate the efficacy of the proposed approach.**

***Index Terms*—Coupled learning, cross-modal matching, deep model, metric learning, multimedia retrieval.**

I. Introduction

CROSS-MODAL matching has been widely studied in computer vision and multimedia analysis in recent years due to the rapid growth of data in the form of images, videos and texts [1]–[9], which has many real-world applications such as multimedia retrieval [4], [10]–[13], image annotation and labeling [6], [11], [14], image classification [15], [16] and heterogeneous face recognition [17]–[19].

The objective of cross-modal matching is to determine whether a pair of samples from two different modalities represent the same object or not. One representative example of cross-modal matching is image-text retrieval which matches the most semantically relevant images for a given text query. This problem is challenging because there is usually an inherent heterogenous gap between two different sets of modalities (e.g. textual features vs image features). Another example is the cross-modal face recognition problem, which recognizes face images captured from two different environments, such as visible images vs. near-infrared images. Similarly, face representations captured from two modalities have large variations in illumination, camera view, and occlusions. In these two examples, a key challenge is how to effectively reduce the modality gap and exploit discriminative information across modalities so that the similarity of samples from different modalities can be effectively computed.

A variety of cross-modal matching methods [9] have been proposed in recent years, and the typical approach is to seek one common semantic space to reduce the modality gap. For example, canonical correlation analysis (CCA) [3] was applied for cross-modal matching where it projects two sets of features of different modalities into one common space where their correlation is maximized. Similarly, partial least square (PLS) [20] and semantic correlation matching (SCM) [4] used a similar idea to reduce the modality gap by using different statistical techniques and formulations. While these cross-modal matching methods have achieved encouraging performance, most of them employ direct projections from the original feature representations, which usually cannot truly capture the high-level semantics from nonlinear real-world data. While there are studies that provide nonlinear transformations based on kernels [21], [22], these models are not scalable for new training data. While more recent deep learning models have provided scalable nonlinear hierarchical transformations for discriminant feature representations, only few of them have been implemented particularly for cross-modal matching [23]–[25]. Hence, how to learn a model which can extract high-level semantic representations efficiently from nonlinear relationships across different modalities remains a challenging problem in cross-modal matching.

In this paper, we propose a new deep coupled metric learning (DCML) method for cross-modal matching. Unlike most existing methods modal-invariant feature learning methods such as CCA and PLS which learn a single linear latent space to reduce the modality gap, our DCML designs two neural networks to learn two sets of hierarchical nonlinear transformations (one set for each modality) to nonlinearly map data samples into a

Manuscript received February 1, 2016; revised May 23, 2016, October 24, 2016, and December 15, 2016; accepted December 26, 2016. Date of publication December 29, 2016; date of current version May 13, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61225008, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Balakrishnan Prabhakaran. *(Corresponding author: Jiwen Lu.)*

V. E. Liong is with the Interdisciplinary Graduate School, Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore 639798 (e-mail: veniceer001@e.ntu.edu.sg).

J. Lu and J. Zhou are with the Department of Automation, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua University, Beijing 100084, China, and also with the Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China (e-mail: lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Y.-P. Tan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eyptan@ntu.edu.sg).

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/TMM.2016.2646180

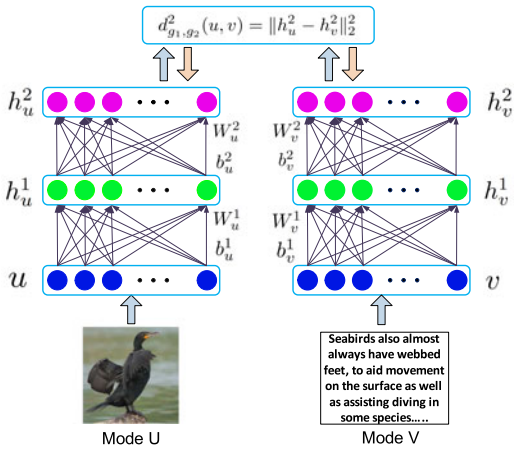

Fig. 1. Basic idea of the proposed DCML method for cross-modal matching. For each pair of samples which were captured from two different modalities (e.g. mode U and mode V), we first represent them as two feature vectors, u and v. Then, we pass them into the designed two feed-forward neural networks to nonlinearly map them into a shared feature subspace, where each network contains a set of hierarchical nonlinear transformations and u and v are represented as h^2_u and h^2_v, respectively. Finally, we compute the squared distance between these samples at the top layers of the networks and use the distance for cross-modal matching.

shared feature subspace, under which the intra-class variation is minimized and the inter-class variation is maximized, and the difference of each sample pair captured from two modalities of the same class is minimized, respectively. Fig. 1 illustrates the basic idea of the proposed approach. Experimental results on four different cross-modal matching applications demonstrate the effectiveness of the proposed method.

II. Related Work

In this section, we briefly review three related topics: 1) cross-modal matching 2) metric learning and 3) deep learning.

A. Cross-Modal Matching

Existing cross-modal matching methods [3]–[5], [12], [19], [26]–[28] can be categorized into two classes: homogenous data synthesis and cross-modal invariant feature learning. For the first class, data from one modality are synthesized into another modality so that the modality gap is reduced. For the second class, data instances from different modalities are mapped into a common latent space so that the modality difference is reduced and they can be measured directly. Most existing cross-modal matching methods learn a single latent space through a pair of linear transformations to reduce the modality gap. Real world data are typically nonlinear in nature, hence these methods may not be fully effective for reducing the modality gap during cross-modal matching. There are several kernel-based methods that provide nonlinear feature representations that are invariant for cross-modalities. For example, Hardoon *et al.* [21] used kernel canonical correlation analysis (KCCA) to learn a common semantic space for for cross-modal retrieval. Hwang *et al.* [29] and Ballan *et al.* [6] employed variants of KCCA for automatic image annotation. Gong *et al.* [8] performed nonlinear kernel embedding followed by a linear dimensionality reduction and CCA for content-based retrieval and tag-image search. While these methods may provide representative features, these methods may not be scalable when new training data are available. In addition, most kernel-based methods do not exploit label information which make them less discriminative. With this in mind, we develop a scalable framework using metric learning and deep learning techniques which provides strong nonlinear representations for each modality such that the gap between them is reduced.

B. Metric Learning

In recent years, numerous metric learning methods have been proposed in computer vision and machine learning [30]. The aim of metric learning is to learn a distance metric such that the distance between semantically similar pairs are reduced, and dissimilar pairs are enlarged as much as possible. These methods can be mainly categorized into two classes: unsupervised and supervised. However, most existing metric learning methods are designed for intra-modal matching, which cannot effectively model the relationship of images captured from different modalities. More recently, several coupled metric learning algorithms have been proposed for cross-modal matching such as cross-modal metric learning (CMML) [31], maximum-margin coupled mappings (MMCM) [32], and coupled marginal fisher analysis (CMFA) [33]. However, these methods only learn a pair of linear transformations to map cross-modal samples into a new common feature space, which is not effective enough to discover the nonlinear relationship of samples. Our proposed DCML is a metric learning approach which learns two sets of nonlinear transformations to map data samples into common space a such that the intra-class variation is minimized and the inter-class variation is maximized, and the difference of each sample pair captured from two modalities of the same class is minimized.

C. Deep Learning

Deep learning aims to build high-level features by learning hierarchical feature representations from raw data. Over the past few years, a number of deep learning algorithms have been proposed [34] and some of them have been successfully employed in various computer vision applications such as image classification [35], object detection [36], and visual tracking [37]. Existing deep learning methods can be mainly categorized three classes: unsupervised, supervised, and semi-supervised. Representative deep learning models included deep stacked auto-encoder [38], deep convolutional neural networks [39], and deep belief network [40]. While many attempts have been made on deep learning in the literature, little progress has been made for cross-modal matching applications [7], [23]–[25], [41]. Kim *et al.* [41] used a deep belief network to learn the representation of each modality and then perform semantic transformation using CCA. More recently, Ngiam *et al.* [23] proposed a multi-modal deep learning framework based on an auto-encoder architecture to discover the correlation across modalities. Feng *et al.* [24] proposed a deep auto-encoder for cross-media retrieval which jointly reconstructs the representation for each

modality and maximize the correlation between modalities. Yan *et al.* [7] implemented the deep canonical correlation analysis (DCCA) [42] method for image and text matching through maximizing the canonical correlation objective through a deep network. Wang *et al.* [25] proposed a unified deep network to capture the high-level semantics and correlations between two modalities. Unlike these deep learning methods, our DCML method is a deep framework based on a coupled metric learning approach to jointly exploit the discriminative information among training samples and reduce the modality gap.

III. Proposed Approach

In this section, we first briefly review the coupled metric learning approach, then present the proposed DCML method and its implementation details.

A. Coupled Metric Learning

Let $U = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^{d_1 \times N}$ and $V = [v_1, v_2, \ldots, v_N] \in \mathbb{R}^{d_2 \times N}$ be two sets of samples captured from two different modalities, $u_i \in \mathbb{R}^{d_1}$ and $v_i \in \mathbb{R}^{d_2}$ are the ith corresponding pair, $1 \leq i \leq N$, and d_1 and d_2 are the corresponding feature dimension. Coupled metric learning approach aims to seek the following projection functions:

$$g_1 : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^d, \quad g_2 : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^d \tag{1}$$

where $d \ll \min(d_1, d_2)$ is the feature dimension of the learned latent common space, under which the similarity of these two sets is maximized so that the gap difference is reduced as much as possible.

Most existing coupled metric learning algorithms [3], [20] assume that g_1 and g_2 are linear parameterized functions which are usually defined as

$$g_1 = W_1 U, \quad g_2 = W_2 V \tag{2}$$

where $W_1 \in \mathbb{R}^{d \times d_1}$ and $W_2 \in \mathbb{R}^{d \times d_2}$ are two linear projection matrices.

Many criterions can be employed to measure the similarity of two sets U and V in the learned metric space, and the most two popular measures are canonical correlation maximization and Euclidean distance minimization. For the first one, CCA is the most popular method. For the second one, representative methods include CDFE [43], CSR [44], CMML [31], MvDA [17], GMLDA [45], and GMMFA [45].

B. Deep Coupled Metric Learning

Most existing coupled metric learning algorithms only learn a pair of linear transformation, which are not effective enough to capture the nonlinear manifolds where samples usually lie on. To address this limitation, several kernel-based coupled metric learning methods such as kernel CCA (KCCA) [3], [21] and its variants [6], [8], [29] have been proposed, which adopt the kernel trick to implicitly map samples into a high-dimensional kernel feature space and then apply coupled metric learning in that high-dimensional space. However, these methods suffer from the scalability problem because there is no explicit nonlinear mapping in these kernel-based methods. To address the limitation of these previous coupled metric learning methods, we propose a deep coupled metric learning (DCML) method by developing two deep neural networks to learn two sets of hierarchical nonlinear transformations (one set for each modality) to nonlinearly map cross-modal samples into a shared feature subspace, so that both the nonlinearity and scalability problems can be addressed simultaneously.

As shown in Fig. 1, we construct a coupled deep neural network to compute feature representation for cross-modal data, one network for one modality. Specifically, given each pair of data instance from two modalities, we pass them into the deep networks which consist of multiple stacked layers of nonlinear transformations. Assume there are $L+1$ layers in each of these two networks (denoted as Mode U and Mode V), and d_u^l and d_v^l units for the network U and network V in the lth layer, respectively, where $1 \leq l \leq L$. For two data instances u_i and v_i which are from two modalities, their outputs of the first layer in these two networks are computed as: $h_{\mathrm{iu}}^1 = s(W_u^1 u_i + b_u^1)$, $h_{iv}^1 = s(W_v^1 v_i + b_v^1)$, where W_u^1 and W_v^1 are the projection matrices to be learned in the first layer of these networks, and b_u^1 and b_v^1 are the bias vectors, s is a nonlinear active function which is applied component-wise. Then, the outputs of the first layer of these networks h_{iu}^1 and h_{iv}^1 are used as the inputs of the second layer. Hence, the outputs of the second layer are: $h_{\mathrm{iu}}^2 = s(W_u^2 h_{\mathrm{iu}}^1 + b_u^2)$, $h_{\mathrm{iv}}^2 = s(W_v^2 h_{\mathrm{iv}}^1 + b_v^2)$, where W_u^2 and W_v^2 are the projection matrices, and b_u^2 and b_v^2 are the bias vectors of the second layer, respectively. Similarly, the outputs for the lth layer are: $h_{\mathrm{iu}}^l = s(W_u^l h_{\mathrm{iu}}^{l-1} + b_u^l)$, $h_{\mathrm{iv}}^l = s(W_v^l h_{\mathrm{iv}}^{l-1} + b_v^l)$, and the outputs for the top layer are

$$g_1(u) = \ h_u^L = s(W_u^L h_u^{l-1} + b_u^L) \tag{3}$$

$$g_2(v) = \ h_v^L = s(W_v^L h_v^{l-1} + b_v^L) \tag{4}$$

where W_u^L and W_v^L are the projection matrices, b_u^L and b_v^L are the bias vectors for the top layer.

To improve the cross-modal matching accuracy, we have the following two objectives:

1) It is desirable to exploit more discriminative information from training samples.
2) It is expected to reduce the modality gap of the data pair captured from different modalities.

Fig. 2 illustrates the basic idea of our proposed DCML method. To achieve the first objective, we employ a large margin criterion to minimize the intra-class variation and maximize the inter-class variation for feature representation at the top layer of these two networks, simultaneously. Specifically, for each pair of training samples u_i and v_j which are from two different modalities, we compute their squared distance $d_{g_1,g_2}^2(h_{\mathrm{iu}}^L, h_{\mathrm{jv}}^L)$ at the top layer of these two networks as follows:

$$d_{g_1,g_2}^2(h_{\mathrm{iu}}^L, h_{\mathrm{jv}}^L) = \|h_{\mathrm{iu}}^L - h_{\mathrm{jv}}^L\|_2^2. \tag{5}$$

We expect that $d_{g_1,g_2}^2(h_{\mathrm{iu}}^L, h_{\mathrm{jv}}^L)$ is as small as possible if u_i and v_j are the same class, and as large as possible if they are from

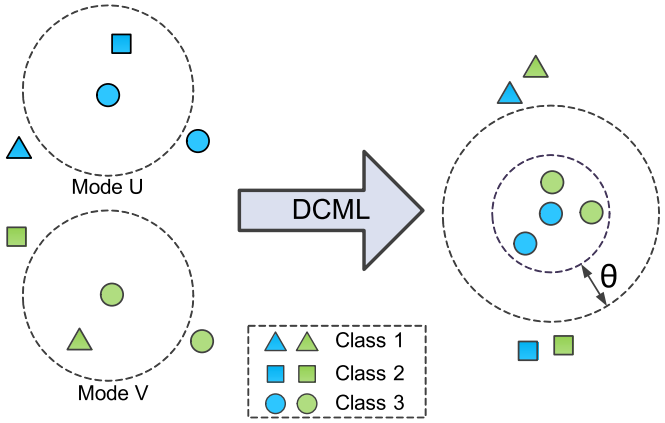

Fig. 2. Basic idea of the proposed DCML method. There are eight samples captured from two different modalities (e.g., Mode U and V), and each modality has four samples. All these eight samples come from three classes, denoted as circles, triangles and squares, respectively. In the original feature space of each modality, the similarity of the samples from class 1 is smaller than that of the samples from two different class (e.g., class 1 and class 3 in Mode U, and class 1 and class 3 in Mode V), which may reduce the recognition performance. In the learned latent feature space by DCML, we expect that the intraclass similarity is increased and interclass similarity is reduced, and the similarity of samples from the same class across different modalities is increased, so that discriminative information can be exploited and modality gap can be reduced, simultaneously. Having matched samples captured from two different modalities into the same sematic space, we can conduct matching for different applications, respectively.

different class, which are formulated as the following constrains:

$$d^2_{g_1,g_2}(h^L_{\rm iu}, h^L_{\rm jv}) \le \theta_1, \quad \text{if } l_{u_i v_j} = 1 \tag{6}$$

$$d^2_{g_1,g_2}(h^L_{\rm iu}, h^L_{\rm jv}) \ge \theta_2, \quad \text{if } l_{u_i v_j} = -1 \tag{7}$$

where $l_{u_i v_j} = 1$ means that u_i and v_j are the same class, and $l_{u_i v_j} = -1$ indicates that they are from different class, θ_1 and θ_2 are the small and large thresholds, respectively. To reduce the number of parameters, we adopt the following large margin optimization objective function to integrate these two constrains:

$$\min_{g_1,g_2} H_1 = 1 - l_{u_i v_j}\left(\theta - d^2_{g_1,g_2}(h^L_{\rm iu}, h^L_{\rm jv})\right) \tag{8}$$

where $\theta_1 = \theta - 1$ and $\theta_2 = \theta + 1$.

To achieve the second objective, we minimize the difference between each pair of data of the same class captured from different modalities over all layers

$$\min_{g_1,g_2} H_2 = \sum_{l=1}^{L-1}\sum_{i=1}^{N} \|h^l_{\rm iu} - h^l_{\rm iv}\|_2^2. \tag{9}$$

By applying the above two criterions for all samples in the training samples, we formulate the following optimization problem for our proposed DCML model:

$$\begin{aligned}\arg\min_{g_1,g_2} H = {} & H_1 + \lambda_1 H_2 + \lambda_2 H_3 \\ = {} & \frac{1}{2}\sum_{i,j=1}^{N} f(1 - l_{u_i v_j}(\theta - d^2_{g_1,g_2}(h^L_{iu}, h^L_{\rm jv})) \\ & + \frac{\lambda_1}{2}\sum_{i=1}^{N}\sum_{l=1}^{L-1}\delta_{\rm ij}\|h^l_{\rm iu} - h^l_{\rm iv}\|_2^2 \\ & + \frac{\lambda_2}{2}\sum_{l=1}^{L}\Big(\|W^l_u\|_F^2 + \|W^l_v\|_F^2 \\ & + \|b^l_u\|_2^2 + \|b^l_v\|_2^2\Big)\end{aligned} \tag{10}$$

where H_1 exploits the discriminative information using a large-margin criterion and label information to learn nonlinear projections, H_2 reduces the modality gap by preserving the similarity between each layer for similar training pairs, and H_3 represents the regularization of the parameters of the developed deep networks, λ_1 and λ_2 are two parameters to balance the effect of the different terms, $\delta_{\rm ij}$ is an indicator where it is 1 if u_i and v_j shares some common label and 0 otherwise, and $f(z)$ is a generalized logistic loss function to smoothly approximate the hinge loss function $z = \max(z, 0)$, and is defined as follows[1]:

$$f(z) = \frac{1}{\rho}\log(1 + \exp(\rho z)) \tag{11}$$

where ρ is the sharpness parameter.

We employ the stochastic sub-gradient descent algorithm to solve the optimization problem defined in (10) and obtain the parameters $\{W^l_u, W^l_v, b^l_u, b^l_v\}|_{l=1}^{L}$. The gradient of the objective function H with respect to these parameters can be computed as follows:

$$\begin{aligned}\frac{\partial H}{\partial W^l_u} = {} & \sum_{i,j=1}^{N}\Phi^l_{\rm iu}(h^{l-1}_{\rm iu})^T + \sum_{i=1}^{N}\lambda_1\Psi^l_{\rm iu}(h^{l-1}_{\rm iu})^T \\ & + \lambda_2 W^l_u\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial H}{\partial W^l_v} = {} & \sum_{i,j=1}^{N}\Phi^l_{\rm jv}(h^{l-1}_{\rm jv})^T + \sum_{i=1}^{N}\lambda_1\Psi^l_{\rm iv}(h^{l-1}_{\rm iv})^T \\ & + \lambda_2 W^l_v\end{aligned} \tag{13}$$

$$\frac{\partial H}{\partial b^l_u} = \sum_{i,j=1}^{N}\Phi^l_{\rm iu} + \sum_{i=1}^{N}\lambda_1\Psi^l_{\rm iu} + \lambda_2 b^l_u \tag{14}$$

$$\frac{\partial H}{\partial b^l_v} = \sum_{i,j=1}^{N}\Phi^l_{\rm jv} + \sum_{i=1}^{N}\lambda_1\Psi^l_{\rm iv} + \lambda_2 b^l_v \tag{15}$$

where Φ and Ψ are the updating functions. For the top layer when $l = L$, they can be computed as follows:

$$\begin{aligned}\Phi^L_{\rm iu} &= f'(\gamma)(h^L_{\rm iu} - h^L_{\rm jv}) \odot s'(y^L_{\rm iu}) \\ \Phi^L_{\rm jv} &= f'(\gamma)(h^L_{\rm jv} - h^L_{\rm iu}) \odot s'(y^L_{\rm jv}) \\ \Psi^L_{\rm iu} &= \delta_{\rm ij}(h^L_{\rm iu} - h^L_{\rm iv}) \odot s'(y^L_{\rm iu}) \\ \Psi^L_{\rm iv} &= \delta_{\rm ij}(h^L_{\rm iv} - h^L_{\rm iu}) \odot s'(y^L_{\rm iv})\end{aligned}$$

where

$$\begin{aligned}\gamma &\triangleq 1 - l_{u_i v_j}(\theta - d^2_{g_1 g_2}(h^L_{\rm iu}, h^L_{\rm jv})) \\ y^l_{\rm iu} &\triangleq W^l_u h^{l-1}_{\rm iu} + b^l_u\end{aligned}$$

[1]We performed empirical tests and found that our method yielded better performance when a smooth approximation for the hinge loss function was used.

Algorithm 1: DCML

Input: Training set U and V, network layer number $L+1$, threshold θ, learning rate η, iterative number R, parameter λ_1 and λ_2, and convergence error ε.
Output: Parameters $\{W_u^l, b_u^l\}_{l=1}^L$ and $\{W_v^l, b_v^l\}_{l=1}^L$.
Step 1 (Initialization):
Initialize $\{W_u^l, b_u^l\}_{l=1}^L$ and $\{W_v^l, b_v^l\}_{l=1}^L$.
Step 2 (Optimization by back propogation):
for $r = 1, 2, \cdots, R$ **do**
  Randomly select a sample pair $(u_i, v_j; l_{u_i v_j})$ from the training set.
  Set $h_{iu}^0 = u_i$ and $h_{jv}^0 = v_j$, respectively.
  for $l = 1, 2, \cdots, L$ **do**
    Compute h_{iu}^l and h_{jv}^l using the deep networks.
  end
  for $l = L, L-1, \cdots, 1$ **do**
    Obtain the gradients according to (12)-(15).
  end
  for $l = 1, 2, \cdots, L$ **do**
    Update W_u^l, W_v^l, b_u^l and b_v^l according to (16)-(17).
  end
  Calculate H_r using (10).
  If $r > 1$ and $|H_r - H_{r-1}| < \varepsilon$, go to **Return**.
end
Return: $\{W_u^l, b_u^l\}_{l=1}^L$ and $\{W_v^l, b_v^l\}_{l=1}^L$.

$$y_{\mathrm{jv}}^l \triangleq W_v^l h_{\mathrm{jv}}^{l-1} + b_v^l$$
$$y_{\mathrm{iv}}^l \triangleq W_v^l h_{\mathrm{iv}}^{l-1} + b_v^l.$$

For all other layers, $1 \le l \le L-1$, Φ and Ψ are computed as follows:

$$\Phi_{\mathrm{iu}}^l = (W_u^{l+1})^T \Phi_{\mathrm{iu}}^{l+1} \odot s'(y_{\mathrm{iu}}^l)$$
$$\Phi_{\mathrm{jv}}^l = (W_v^{l+1})^T \Phi_{\mathrm{jv}}^{l+1} \odot s'(y_{\mathrm{jv}}^l)$$
$$\Psi_{\mathrm{iu}}^l = ((W_u^{l+1})^T \Psi_{\mathrm{iu}}^{l+1} + \delta_{\mathrm{ij}}(h_{\mathrm{iu}}^l - h_{\mathrm{iv}}^l)) \odot s'(y_{\mathrm{iu}}^l)$$
$$\Psi_{\mathrm{iv}}^l = ((W_v^{l+1})^T \Psi_{\mathrm{iv}}^{l+1} + \delta_{\mathrm{ij}}(h_{\mathrm{iv}}^l - h_{\mathrm{iu}}^l)) \odot s'(y_{\mathrm{iv}}^l)$$

where the operation "$\odot$" denotes the element-wise multiplication.

Then, we use the the following gradient descent algorithm to update the parameters W_u^l, W_v^l, b_u^l and b_v^l of our deep networks:

$$W_u^l = W_u^l - \eta \frac{\partial H}{\partial W_u^l}, \quad W_v^l = W_v^l - \eta \frac{\partial H}{\partial W_v^l} \tag{16}$$

$$b_u^l = b_u^l - \eta \frac{\partial H}{\partial b_u^l}, \quad b_v^l = b_v^l - \eta \frac{\partial H}{\partial b_v^l} \tag{17}$$

where η is the learning rate in our gradient descent algorithm.

Algorithm 1 summarizes the detailed procedure of the proposed DCML method.

C. Implementation Details

Our deep network consists of several fully-connected layers of different dimensions, where the learning rate η, parameter λ_1 and λ_2 were set as 0.0001, 0.01 and 0.0001, respectively[2] for all experiments. The parameters W_u^l and W_v^l of our DCML model were initialized as $\mathbf{I} \in \mathbb{R}^{d_l \times d_{l-1}}$ (d_l is the feature dimension of the lth layer), which is a matrix with ones on the diagonal and zeros elsewhere. The bias vectors b_u^l and b_v^l were initialized as zero vectors. For the activation function, we used the *tanh* function. We set the layers to $L = 2$ for all experiments to prevent over-fitting. We performed empirical tests which show that networks with $L > 2$ are comparable with $L = 2$. It is expected that if more training pairs are used, a deeper network would be more effective [46].

In the training stage, we randomly choose sample pairs and iteratively passed through them to the network. For each epoch, the positive and negative pairs are of equal quantity. The training stage converges when the objective function does not change within a certain threshold $\epsilon = 0.0001$ for an epoch.

IV. Experiments

We conducted experiments on three different cross-modal matching applications which includes text-image matching, tag-image retrieval, and heterogeneous face recognition on four datasets to demonstrate the effectiveness of the proposed DCML method. The followings describe the details of the experiments and results.

A. Text-Image Matching on the Wiki Dataset

We applied the Wiki image-text dataset [4] for cross-modal text-image retrieval. The dataset consists of 2866 image-text pairs where each pair consists of an image and the corresponding complete text article annotated with a label from 10 semantic classes (i.e. sport, music, warfare, etc). We evaluated our DCML using two image descriptors in our experiments. First, each image is represented by a 128-dimensional SIFT descriptor by following the settings in [4], [45]. Second, we extracted a deep convolutional neural network (CNN) feature where the model is pre-trained in the ImageNet dataset [35] with the deep architecture in [47] where we extract the features from the fc7 layer and employed PCA to reduce it into 512 dimension.[3] For each text, we represented it as a 10-dimensional feature vector with the Latent Dirichlet allocation (LDA) model. We randomly used 1300 pairs in our experiments for training, 130 pairs per class, and used the remaining 1566 pairs for testing. For a fair comparison, we implemented other cross-modal matching methods with their publicly available codes under the same protocol. We repeated our experiments 10 times and took the average as the final matching results.

We compared our DCML with eight cross-modal matching methods: CCA [3], PLS [20], MvDA [17], GMLDA [45], GMMFA [45], SCM [4], KCCA [21] and LCFS [5]. CCA and PLS are cross-modal models which reduce the modality gap via subspace analysis using pairwise information. MvDA, GMLDA and GMFA learned a single common discriminative representations for two modes by using the fisher criterion. KCCA used

[2]We obtained these parameters by using the 10-fold cross-validation strategy.
[3][Online]. Available: http://www.vlfeat.org/matconvnet/models/imagenet-vgg-f.mat.

TABLE I
MAP (%) COMPARISON OF OUR DCML AND STATE-OF-THE-ART CROSS-MODAL MATCHING METHODS ON THE WIKI DATASET USING THE SIFT FEATURE

Method	Image query	Text query	Average
PLS [20]	21.49	17.07	19.28
CCA [3]	24.60	19.18	21.89
GMMFA [45]	24.20	17.97	21.09
GMLDA [45]	18.08	13.83	15.96
MvDA [17]	15.99	13.26	14.63
SCM [4]	23.84	22.23	23.04
KCCA [21]	26.85	21.34	24.10
LCFS [5]	26.83	21.77	24.30
CDL* [28]	27.76	23.11	25.44
LGCFL* [12]	27.90	21.77	24.80
JFSSL* [48]	30.63	22.75	26.69
DCML	**35.04**	**25.55**	**30.03**

*The results are from the original papers.

TABLE II
MAP (%) COMPARISON OF OUR DCML AND STATE-OF-THE-ART CROSS-MODAL MATCHING METHODS ON THE WIKI DATASET USING THE CNN FEATURE

Method	Image query	Text query	Average
PLS [20]	31.28	26.45	28.86
CCA [3]	35.42	32.50	33.96
GMMFA [45]	27.76	12.85	20.30
GMLDA [45]	18.01	14.09	16.05
MvDA [17]	16.17	12.00	14.09
SCM [4]	38.74	37.03	37.88
KCCA [21]	37.34	34.61	35.98
LCFS [5]	39.39	38.09	38.74
LRBS* [49]	44.41	37.70	41.06
RE-DNN* [25]	34.04	35.26	34.65
JFSSL* [48]	42.79	39.57	41.18
DCML	**55.36**	**53.81**	**54.59**

*The results are from the original papers.

kernels and find a common space between two modalities based on the pairwise correlation information.[4] SCM obtained representations for two modalities to maximize their correlations and transform them in a semantic space. LCFS learned a coupled projections to transform each modal to a common subspace defined by class labels and low-rank constraint. Among these methods, only KCCA performed nonlinear representations while other methods performed linear regression or projection. The source codes of these compared methods are provided by the authors and we carefully tuned the parameters of different methods to obtain the best results for a fair comparison.

In our DCML method, we trained our model using three layers ($L = 2$) on the training set and the feature dimension for these layers were set as 128→ 50 →20 and 10→ 50 →20 for the image and text when the SIFT feature is used, and 512→ 100 →50 and 10→ 100 →50 when the CNN feature is employed, respectively. For experiments with hand-crafted features, we also compared our method with CDL [28] which performed dictionary learning, LGCFL [12] which used a variant of subspace learning method to reduce the modality gap, and JFSSL [48] which performed linear regressions and multimodal graph regularization for feature selection. For experiments with the CNN features, we also compared our method with JFSSL, LRBS [49] and RE-DNN [25]. LRBS [49] employed bilinear transformation and RE-DNN used deep networks for reducing the modality gap, respectively. Here, CDL and RE-DNN are nonlinear representations. Different from our deep network, RE-DNN performed multi-modal learning by using a Euclidean distance criterion with two stages of intra-modal pre-training and inter-modal full training, while our method performed direct full training with cross-modal specific criterions. Tables I and II show the mean average precision (mAP) of our DCML and other cross-modal methods on the Wiki image-text dataset when the hand-crafted and CNN features were used, respectively. It can be seen that our DCML still achieves the best performance for both CNN and SIFT, respectively. It is important to note that it is difficult to perform an exact comparison due to certain experimental differences. Particularly, LRBS and RE-DNN employed 200 and 20 LDA topics as text features respectively, while LRBS and RE-DNN used different CNN models for their experiments. In addition, for LGCFL, LRBS, and Re-DNN, the number of the train and test splits are also different from ours which followed [5] experimental set-up. We see that our DCML outperforms other compared cross-modal matching methods by 13% and 3% in average when the CNN and SIFT features were used, respectively. Fig. 3 shows the precision-scope curves of different coupled metric learning methods with the SIFT feature on the Wiki image-text dataset. The scope refers to the number of top-ranked samples. Similarly, the precision-scope evaluation is consistent with the mAP for both the image and text query tasks, where our DCML outperforms the other compared methods significantly.

[4]In our experiments, we made use of Gaussian Kernels which provided the best results among different kernels.

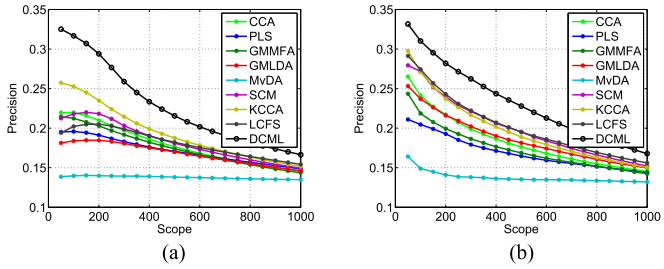

Fig. 3. Precision-scope curve versus different values of M for different coupled metric learning methods from the Wiki Experiment using SIFT image features, where M is the number of top-retrieved instances. (a) Image → Text. (b) Text → Image.

B. Tag-Image Retrieval on the Pascal VOC 2007 Dataset

In this subsection, we conducted tag-image retrieval experiments on the Pascal VOC 2007 [50]. The dataset contains 5011 image-text pairs for training, and 4952 image-text pairs for testing. Each image-text pair is annotated from 20 categories (i.e. aeroplane, bottle, horse, sofa). Unlike the Wiki dataset which utilizes text information from articles, the Pascal VOC 2007 dataset only makes use of tag information. We also evaluated

TABLE III
MAP (%) COMPARISON OF OUR DCML AND STATE-OF-THE-ART CROSS-MODAL MATCHING METHODS ON THE PASCAL VOC 2007 DATASET USING THE HAND-CRAFTED IMAGE FEATURE

Method	Image query	Text query	Average
PLS [20]	17.40	14.01	15.71
CCA [3]	14.89	14.28	14.59
GMMFA [45]	29.65	26.52	28.08
GMLDA [45]	30.58	25.47	28.02
MvDA [17]	12.30	9.15	10.72
SCM [4]	23.08	19.38	21.23
KCCA [21]	29.24	24.66	26.95
LCFS [5]	41.37	33.48	37.43
CDL* [28]	37.41	29.44	33.42
LGCFL* [12]	40.10	32.00	36.00
JFSSL* [48]	36.07	28.01	32.04
DCML	**44.49**	**36.26**	**40.38**

*The results are from the original papers.

TABLE IV
MAP (%) COMPARISON OF OUR DCML AND STATE-OF-THE-ART CROSS-MODAL MATCHING METHODS ON THE PASCAL VOC 2007 DATASET USING THE CNN IMAGE FEATURE

Method	Image query	Text query	Average
PLS [20]	49.47	52.53	51.00
CCA [3]	30.49	31.28	30.89
GMMFA [45]	64.73	68.86	66.80
GMLDA [45]	67.19	72.47	69.83
MvDA [17]	70.17	7171	70.94
SCM [4]	69.09	68.52	68.81
KCCA [21]	67.15	67.66	67.41
LCFS [5]	70.94	74.74	72.84
LRBS* [49]	65.10	68.69	66.90
DCML	**73.77**	**75.01**	**74.39**

*The results are from the original papers.

our DCML using two set of descriptors to represent the images in our experiments. The first image feature representation is the concatenation of bag of visual words (BOVW), GIST features and color histograms provided by [51], which is a 776-dimension feature vector. The second is the CNN feature which is extracted from a similar setting as the previous experiment. The text feature representation is based on the absolute rank feature also provided by [51] and were tagged by the Amazon Mechanical Turk. We used the original train and test split provided but removing image-text pairs that have multiple labels, resulting to 2808 and 2841 train and test set, respectively.

In our DCML method, we trained our model using three layers on the training set and the feature dimension for these layers were set as 776→ 200 →100 and 399→ 200 →100 for the image and text when the BOVW-GIST-COLOR feature is used, and 512→ 300 →50 and 399→ 300 →50 when the CNN feature is employed, respectively. We also compared our method with JFSSL, CDL, LGCFL, and LRBS. Tables III and IV show the mean average precision (mAP) of our DCML and other cross-modal methods on the Pascal dataset with image and text query, respectively. We see that our DCML outperforms other compared cross-modal matching methods by approximately 2% and 3% when the CNN and handcrafted features were used,

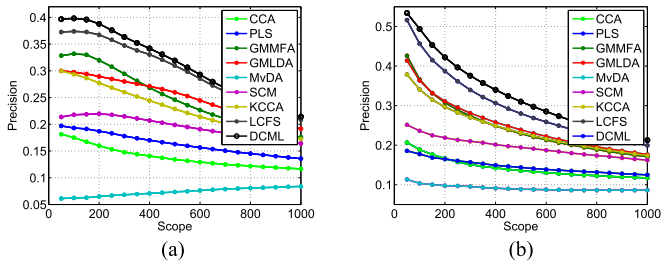

Fig. 4. Precision-scope curve versus different values of M for different coupled metric learning methods from the PASCAL VOC 2007 Experiment using BOVW-GIST-COLOR image features, where M is the number of top-retrieved instances for the Pascal VOC 2007 dataset. (a) Image → Text. (b) Text → Image.

respectively. As expected, the CNN features generally achieved better performance because of its strong representation. Fig. 4 shows the precision-scope curves of different coupled metric learning methods using the handcrafted feature. We see that our DCML still outperforms the other compared methods in both forms of the cross-modal matching tasks.

C. Tag-Image Retrieval on the NUS-WIDE Dataset

In this subsection, we conducted another tag-image retrieval experiment on the NUS-WIDE dataset [52]. This dataset contains approximately 270 000 images with concepts of 81 categories. Following the same setting in [24], we also selected 10 categories having the largest quantity and extracted 1000 image-tag pairs for each category. Hence, we have a subset of 10 000 pairs for experiments. In our experiments, we randomly split this subset into three parts: 8000 pairs for training, 1000 pairs for validation, and 1000 pairs for testing, where each part contains each equal number for samples for each category. We also evaluated our DCML with two types of feature descriptors: the combined local feature provided in [52] which is represented as a 1134-dimension feature vector, and CNN feature which is reduced to 512 dimension by PCA. Different from the PASCAL VOC 2007, tag information provided is more representative using larger amount of words for each image. Each tag information is represented by a 1000-dimensional bag-of-words model provided in [52]. To evaluate the performance of different methods, we used the mAP and top 20% percentage measures as in [24]. In our DCML, we performed PCA to map each sample into a 512-dimensional feature vector, and trained the model with a three layer deep model on the training set. The feature dimension for these layers were set as 512→ 200 →100 for both image and text descriptor.

We compared our DCML with eight state-of-the-art cross-modal retrieval methods. Table V shows the mean average precision (mAP) and Top 20% criteria on the image-text and text-image query experiments. We see that our DCML method outperforms all the other compared methods including some deep models which were proposed for cross-modal retrieval [23], [24], [41]. Our method with CNN features is better by 0.5% and 2% in the mAP and 4% and 8% in the top20% evaluation metric for the image query and text query, respectively.

TABLE V
MAP (%) AND TOP 20% RESULTS ON THE NUS-WIDE-10K DATASET

Method	mAP			Top 20%		
	Image query	Text query	Average	Image query	Text query	Average
CCA-AE [41]	32.6	26.8	23.4	37.0	33.8	35.4
CCA-Cross-AE [41]	13.7	34.4	27.2	29.0	47.7	38.4
CCA-Full-AE [41]	14.8	24.2	24.2	37.1	38.2	37.7
Bimodal AE [23]	25.0	29.7	27.4	30.2	35.4	32.8
Bimodal DBN [23]	17.3	20.3	18.8	25.3	27.0	26.2
Corr-AE [24]	31.9	37.5	34.7	47.1	53.5	50.3
Corr-Cross-AE [24]	34.9	34.8	34.9	53.1	59.7	56.4
Corr-Full-AE [24]	33.1	37.9	35.5	49.6	56.5	53.5
DCML	**38.5**	**40.5**	**39.5**	**61.5**	**62.5**	**62.0**
CCA-CNN	69.5	67.4	68.4	65.5	64.8	65.2
SCM-CNN	82.7	77.8	80.2	68.2	66.9	67.6
KCCA-CNN	81.7	80.6	81.1	71.1	70.5	70.8
LCFS-CNN	85.1	80.3	82.7	72.7	72.8	72.8
DCML-CNN	**85.6**	**82.6**	**84.1**	**76.4**	**80.4**	**78.4**

TABLE VI
RECOGNITION PERFORMANCE COMPARISON OF OUR DCML METHOD AND STATE-OF-THE-ART COUPLED METRIC LEARNING METHODS ON THE CASIA VIS-NIR (VERSION 2.0) DATASET

Method	Rank-one	VR1	VR2
CCA [3]	76.08 ± 1.86	43.47	64.53
PLS [20]	33.90 ± 2.99	9.30	30.61
MvDA [17]	42.90 ± 3.60	14.99	39.30
GMLDA [45]	43.61 ± 3.06	13.57	39.17
GMMFA [45]	66.65 ± 2.03	30.32	62.83
KCCA [21]	67.20 ± 3.03	29.77	58.89
DCML	$\mathbf{82.19 \pm 1.11}$	**46.71**	**66.85**

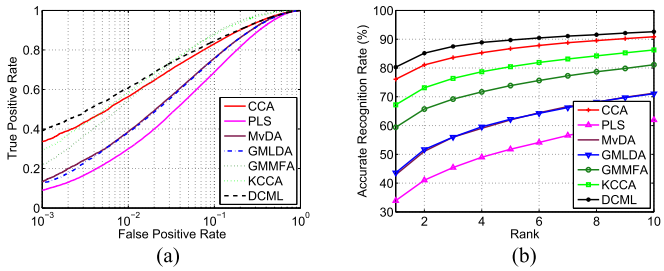

Fig. 5. Performance curves of different coupled metric learning methods son the CASIA VIS-NIR database (version 2). (a) ROC curve. (b) CMC curve.

TABLE VII
RECOGNITION PERFORMANCE COMPARISON OF OUR DCML METHOD AND STATE-OF-THE-ART HETEROGENEOUS FACE RECOGNITION METHODS ON THE CASIA VIS-NIR (VERSION 2.0) DATASET

Method	Rank-one	Year
FaceVACS [55]	58.56 ± 1.19	2012
PCA+Sym+HCA [53]	23.70 ± 1.89	2013
HOG+LDA+Cosine [56]	73.28 ± 1.10	2014
Reconstruction+UDP(DLBP) [57]	78.46 ± 1.67	2015
Gabor+RBM+Remove 11PCs [58]	$\mathbf{86.16 \pm 0.98}$	2015
CDFL (s=3) [59]	71.50 ± 1.40	2015
DCML	$\mathbf{82.19 \pm 1.11}$	

D. Heterogeneous Face Recognition on the CASIA VIS-NIR Dataset

In this subsection, we performed VIS-NIR heterogeneous face recognition on the CASIA VIS-NIR (version 2.0) [53]. There are 275 subjects in the CASIA VIS-NIR (version 2.0) dataset. For each subject, there are 1-22 VIS and 5-50 NIR images. All face images in both the visual and near infrared face sets were aligned and cropped into 128×128 pixels according to the provided eye coordinates. For each face image, we divided it into 16×16 non-overlapped blocks and extracted the SIFT [54] feature from each block. Then, the SIFT features from all blocks within the same image were concatenated into a longer feature vector. Finally, we applied WPCA to learn a project matrix to map each concatenated feature into a 2000-dimensional feature vector.

We followed the standard protocol of the CASIA VIS-NIR (version 2.0) dataset, where VIS images were used as the gallery set and the NIR images were used as the probe set. We trained our deep model using three layers on the "View 1" subset and the dimensions for these layers were set as $2000 \rightarrow 1000 \rightarrow 500$, respectively. All other parameters of our model were as the same as those used in the previous experiments.

Table VI shows the recognition performance of different coupled metric learning methods on the CASIA VIS-NIR (version 2.0) dataset, where three different evaluation measures including the rank-one recognition rate, the verification rate at 0.1% false acceptance rate (VR1), and the verification rate at 1.0% false acceptance rate (VR2) were evaluated and compared. Fig. 5 shows the receiver operating characteristic (ROC) curve and the cumulative match characteristic (CMC) curve of different coupled metric learning methods. We clearly see that our DCML consistently outperforms the current state-of-the-art coupled metric learning methods, and the minimal improvement is 6.11% in terms of the rank-one recognition rate.

We also compared our DCML with state-of-the-art heterogeneous face recognition methods, and Table VII shows the

TABLE VIII
RANK-ONE RECOGNITION RATE ON THE CASIA VIS-NIR (VERSION 2.0) DATASET AND AVERAGE MAP ON THE WIKI, PASCAL VOC 2007, NUS-WIDE DATASET USING DIFFERENT DCML METHODS

Dataset	DCML1	DCML2	DCML
Wiki	26.99	29.52	**30.03**
Wiki (CNN)	45.21	53.14	**54.59**
PASCAL VOC 2007	18.97	39.08	**40.38**
PASCAL VOC 2007 (CNN)	29.39	72.27	**74.39**
NUS-WIDE	38.66	39.00	**39.50**
NUS-WIDE (CNN)	79.35	82.59	**84.00**
CASIA VIS-NIR	79.80	81.76	**82.19**

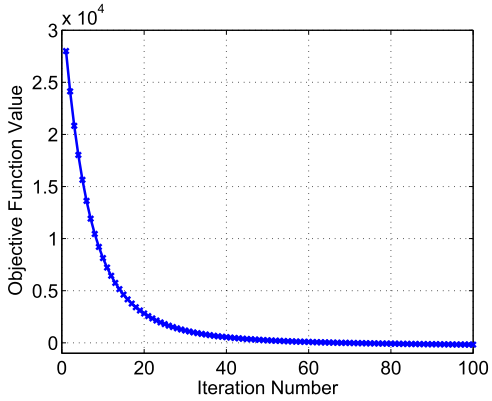

Fig. 6. Convergence curve of DCML on the CASIA VIS-NIR database (version 2) dataset.

performance of different methods. As seen, our method is very competitive and achieved comparable performance with the state-of-the-arts. While our method is worse than the method in [58] which yields the best performance, it is to note that these methods extracted more powerful features for cross-modal face matching besides coupled metric learning while ours used basic handcrafted SIFT features. It is expected that using stronger features would also lead to better performance for our DCML method.

E. Parameter Analysis

Influence analysis of different components in DCML: To investigate the contributions of different terms in our DCML model, we developed two variations of our method: DCML1 and DCML2. For DCML1, the discriminative part H_1 is removed. For DCML2, the correlation part H_2 is removed. The optimization procedure would be similar our DCML method which performs stochastic gradient descent. Table VIII shows the performance of different DCML methods on different datasets. We see that both the discriminative and correlation terms contribute to the final recognition rate. We also see the DCML2 consistently perform better than DCML1 which means the discriminative part H_1 contributes more to the overall recognition rate of our DCML method.

Convergence analysis: We evaluated the convergence of our DCML method. Fig. 6 plots the value of the objective function of DCML versus different number of iterations on the CASIA VIS-NIR (version 2.0) dataset. We see that the proposed DCML method converges in $40 \sim 45$ iterations.

TABLE IX
CPU TIME (SECONDS) USED BY DIFFERENT COUPLED METRIC LEARNING METHODS ON THE CASIA VIS-NIR DATABASE (VERSION 2) DATASET

Method	Training	Testing
CCA [3]	0.30	0.01
PLS [20]	208.50	0.08
MvDA [17]	0.27	0.03
GMLDA [45]	0.59	0.01
GMMFA [45]	0.41	0.01
KCCA [21]	14.10	6.24
DCML	22.17	0.02

F. Computation Time

Lastly, we investigated the computational time of our DCML and compared it with those of existing coupled metric learning methods. Our computer is configured with a 3.40 GHz CPU and 24.0 GB RAM. Table IX shows the computational time for training and testing on the CASIA VIS-NIR (version 2.0) dataset. We see that the computational time of our method for training is generally larger than other coupled metric learning methods, and the testing time is comparable to other coupled metric learning methods.

G. Discussion

The above experimental results suggest the following three key observations:

1) Our DCML method achieves strong performance and beat the state-of-the-art cross-modal methods in three image-text/tag retrieval experiments (Wiki, PASCAL VOC 2007, and NUS-WIDE). This is because our approach trained a deep network to model the nonlinearity of real-world data by exploiting both the discriminative and nonlinear information, simultaneously. Moreover, our DCML can be also extended for the tag annotation task.
2) For text/tag-image retrieval experiments, we have used different types of image features in our experiments and observed that the CNN features provided a boost in the retrieval performance. This also shows that our DCML method is flexible for feature representations with varying dimensions.
3) Our DCML method also achieves competitive performance with heterogenous face recognition experiment, which shows that our model can be used for different cross-modal applications.
4) We have investigated the contribution of each term in our DCML and have shown that both the discriminative and correlation terms contributed to the overall performance. Particularly, the discriminative part of our objective formulation have a larger contribution.

V. Conclusion

In this paper, authors have proposed a new deep coupled metric learning (DCML) method for cross-modal matching. Their method develops two deep neural networks to learn two sets of hierarchical nonlinear transformations to exploit both discriminative information and reduce the modality gap, which significantly improve the performance of different cross-modal matching applications. Experimental results on four cross-modal datasets have clearly demonstrated the effectiveness of the proposed method.

References

[1] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[2] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.

[3] W. Yang, D. Yi, Z. Lei, J. Sang, and S. Z. Li, "2D-3D face matching using CCA," in *Proc. 8th IEEE Int. Conf. Automatic Face & Gesture Recog.*, Sep. 2008, pp. 1–6.

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

[5] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis*, Dec. 2013, pp. 2088–2095.

[6] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo, "A cross-media model for automatic image annotation," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, p. 73.

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

[8] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.

[9] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[10] J. Costa Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[11] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[12] 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.

[13] A. Habibian, T. Mensink, and C. G. Snoek, "Discovering semantic vocabularies for cross-media retrieval," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 131–138.

[14] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 966–973.

[15] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 902–909.

[16] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1903–1910.

[17] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[18] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1707–1716, Dec. 2012.

[19] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 593–600.

[20] A. Li, S. Shan, X. Chen, and W. Gao, "Cross-pose face recognition based on partial least squares," *Pattern Recog. Lett.*, vol. 32, no. 15, pp. 1948–1955, 2011.

[21] D. R. Hardoon and J. Shawe-Taylor, "KCCA for different level precision in content-based image retrieval," in *Proc. 3rd Int. Workshop Content-Based Multimedia Indexing*, 2003, pp. 1–6.

[22] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Trans Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.

[23] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[24] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[25] C. Wang, H. Yang, and C. Meinel, "A deep semantic framework for multimodal representation learning," *Multimedia Tools Appl.*, vol. 75, pp. 9255–9276, 2016.

[26] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.

[27] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 1005–1010.

[28] X. Xu, A. Shimada, R.-I. Taniguchi, and L. He, "Coupled dictionary learning and feature mapping for cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun.-Jul. 2015, pp. 1–6.

[29] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, 2012.

[30] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2012.

[31] A. Mignon *et al.*, "CMML: A new metric learning approach for cross modal matching," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 1–14.

[32] S. Siena, V. N. Boddeti, and B. Kumar, "Maximum-margin coupled mappings for cross-domain matching," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep.-Oct. 2013, pp. 1–8.

[33] S. Siena, V. N. Boddeti, and B. V. Kumar, "Coupled marginal fisher analysis for low-resolution face recognition," in *Proc. 12th Int. Conf. Comput. Vis.*, 2012, pp. 240–249.

[34] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[36] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2553–2561.

[37] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 809–817.

[38] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 3361–3368.

[39] Y. Sun *et al.*, "Hybrid deep learning for face verification," in *Proc. 2013 IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1489–1496.

[40] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[41] J. Kim, J. Nam, and I. Gurevych, "Learning semantics with deep belief network for cross-language information retrieval," in *Proc. Comput. Linguistics*, 2012, pp. 579–588.

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

[43] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 13–26.

[44] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1123–1128.

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

[46] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[47] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[48] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
[49] C. Kang *et al.*, "Cross-modal similarity learning: A low rank bilinear formulation," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1251–1260.
[50] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[51] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–12.
[52] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, p. 48.
[53] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 348–353.
[54] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[55] "FaceVACS software developer kit, Cognitec Systems," 2012, accessed on May 2012. [Online]. Available: http://www.cognitec-systems.de
[56] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1788–1793.
[57] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2015, pp. 141–150.
[58] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Shared representation learning for heterogeneous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, May 2015, pp. 1–8.
[59] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

Venice Erin Liong received the B.S. degree from the University of the Philippines Diliman, Quezon City, Philippines, in 2010, the M.S. degree from the Korea Advanced Institute of Science and Technology, Daejeon City, South Korea, in 2013, and is currently working toward the Ph.D. degree at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Her research interests include computer vision and pattern recognition.

Jiwen Lu (S'10–M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012.

He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From March 2011 to November 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He has authored or coauthored more than 130 scientific, 35 of them IEEE transactions papers. His current research interests include computer vision, pattern recognition, and machine learning.

Prof. Lu is a Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He serves/has served as an Associate Editor of *Pattern Recognition Letters, Neurocomputing*, and IEEE ACCESS, a Managing Guest Editor of *Pattern Recognition and Image and Vision Computing*, and a Guest Editor of *Computer Vision and Image Understanding*. He is/was a Workshop Chair/Special Session Chair/Area Chair for more than ten international conferences. He was the recipient of the National 1000 Young Talents Plan Program in 2015.

Yap-Peng Tan (S'95–M'98–SM'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, USA, in 1995 and 1997, respectively, all in electrical engineering.

From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, USA, and Sharp Laboratories of America, Camas, WA, USA. In November 1999, he joined the Nanyang Technological University of Singapore, where he is currently an Associate Professor and the Associate Chair (Academic) of the School of Electrical and Electronic Engineering. He is the Principal Inventor or Co-Inventor of 15 U.S. patents in the areas of image and video processing. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, and data analytics.

Prof. Tan served as the Chair of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society from 2012 to 2014, a Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society from 2009 to 2013, Voting Member of the IEEE International Conference on Multimedia & Expo Steering Committee from 2011 to 2012, and the Chairman of the IEEE Signal Processing Singapore Chapter from 2009 to 2010. He has also served as an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (since 2016), IEEE TRANSACTIONS ON MULTIMEDIA (since 2014) and IEEE ACCESS (since 2013), an Editorial Board Member of the EURASIP Journal on Advances in Signal Processing and EURASIP Journal on Image and Video Processing, a Guest Editor for special issues of several journals including the IEEE TRANSACTIONS ON MULTIMEDIA, and a member of the Multimedia Systems and Applications Technical Committee and Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He is the Tutorial Co-Chair of the 2016 IEEE International Conference on Multimedia and Expo (ICME 2016) and Technical Program Co-Chair of the 2019 IEEE International Conference on Image Processing, and was the Finance Chair of the 2004 IEEE International Conference on Image Processing, the General Co-Chair of the 2010 IEEE International Conference on Multimedia and Expo, the Technical Program Co-Chair of the 2015 IEEE International Conference on Multimedia and Expo, and the General Co-Chair of the 2015 IEEE International Conference on Visual Communications and Image Processing.

Jie Zhou (M'01–SM'04) received the B.S. and M.S. degrees in mathematics from Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995.

From then to 1997, he was a Postdoctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored or coauthored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 40 papers have been published in top journals and conferences such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *Computer Vision and Pattern Recognition*. His research interests include computer vision, pattern recognition, and image processing.

Prof. Zhou is an Associate Editor for the *International Journal of Robotics and Automation* and two other journals. He was the recipient of the National Outstanding Youth Foundation of China Award.