Generalized Coupled Dictionary Learning Approach with Applications to Cross-modal Matching

Devraj Mandal, Soma Biswas, Senior Member, IEEE

Abstract—Coupled dictionary learning has recently emerged as a powerful technique with wide variety of applications ranging from image synthesis to classification tasks. In this work, we extend the existing coupled dictionary learning approaches in two aspects to make them more suitable for the task of cross-modal matching. Data coming from different modalities may or may not be paired. For example, for image-text retrieval problem, 100 images of a class is available as opposed to only 50 samples of text data for training. Current coupled dictionary learning approaches are not designed to handle such scenarios, where classes of data points in one modality correspond to classes of data points in the other modality. Given the data from the two modalities, first two dictionaries are learnt for the respective modalities so that the data has a sparse representation with respect to their own dictionaries. Then the sparse coefficients from the two modalities are transformed in such a manner that data from the same class are maximally correlated, while that from different classes have very less correlation. This way of modeling the coupling between the sparse representations of the two modalities makes this approach work seamlessly for paired as well as unpaired data. The discriminative coupling term also makes the approach better suited for classification tasks. Experiments on different publicly available cross-modal datasets, namely CUHK photo-sketch face dataset, HFB visible and near-infra-red facial images dataset, IXMAS multiview action recognition dataset, wiki image and text dataset and Multiple Features dataset show that this generalized coupled dictionary learning approach performs better than the state-of-the-art for both paired as well as unpaired data.

Keywords—Coupled dictionary learning, Canonical correlation analysis and Cross-modal matching.

I. INTRODUCTION

Cross-modal data analysis is an evolving area in the field of computer vision and pattern recognition with diverse applications like photo-sketch synthesis and classification, multiview action recognition, text-image retrieval, etc. (Fig. 1). Data coming from different modalities can have very different representations. In addition, the data can also have large intra-class variability within their modality, making cross-modal matching extremely challenging. Several approaches have been proposed in the literature to address this problem. For example, Canonical Correlation Analysis (CCA) [1] [2] or Partial Least Squares (PLS) [3] methods, where a common feature space is learnt to represent the data of the different modalities. The basic idea of all these techniques is to learn salient relationships that exist between the data of the two modalities which is then used for matching the test cross-modal data.

Coupled dictionary learning (CDL) approaches [4] have recently evolved as a powerful technique and have achieved impressive results in different kinds of tasks, ranging from synthesis to classification. The main idea is to learn two dictionaries for the two modalities in a coupled manner such that the sparse coefficients are equal in the original or some transformed space. Like many of the other approaches eg. CCA [1] [2], the standard CDL formulation assumes that the modalities have paired data, i.e. each data point in the first modality is paired with a data point in the second modality. But, in realistic scenarios, the data coming from the two modalities may not be paired. Suppose the task is to match visible light facial images to near infra-red images, and in the training, 4 visible but only 2 near infra-red images of each subject are available. So obviously, there is no one-to-one correspondence between the data points of the two modalities, rather classes of data points in one modality correspond to classes of data points in the second modality. Currently, these methods cannot leverage the information in these kinds of settings. Also, since originally proposed for synthesis tasks, the standard CDL formulation does not have an explicit discriminative term, so they may not prove to be discriminative enough for certain classification tasks.

In this work, we build upon the recent success of the CDL approaches and propose a generalized coupled dictionary learning (GCDL) method. We assume that the correspondence of data in the two modalities is given in the form of classes, i.e. classes of data in one modality correspond to classes of data in the second modality, instead of one-to-one correspondence. Note that the correspondence between classes can be through class labels, though knowledge of class labels is not necessary. As in CDL, in this approach also, the dictionaries correspond-
ing to the two domains are learnt in a coupled manner. But we also learn two projection matrices for the two domains such that all the sparse coefficients of the same class from the two domains are maximally correlated. Thus incorporating the class information into the coupling term, GCDL can handle both paired as well as unpaired data seamlessly in the same framework. We also propose two modifications of the coupling term such that the transformed sparse coefficients of the same class are more correlated as compared to that of different class. This discriminative coupling term makes the approach better suited for classification tasks. The two dictionaries, two transformation matrices and the sparse coefficients are learnt using an iterative algorithm. During testing, the dictionaries are used to compute the sparse coefficients which are then transformed using the respective transformation matrices to be used for matching. We perform extensive evaluation to test the usefulness of the proposed approach. Specifically, we perform experiments on five cross-modal datasets,

- CUHK photo-sketch face dataset [5],
- HFB heterogeneous face dataset [6] for matching visible with near infra-red facial images,
- IXMAS multiview action dataset [7] for matching actions from different views,
- Wiki dataset [8] for image-text and text-image retrieval, and
- Multiple Features Dataset [9] for matching data using different kinds of features.

Comparisons with several state-of-the-art approaches show the effectiveness of the proposed approach for cross-modal matching.

The rest of the paper is organized as follows. Section II discusses the various related work in the literature. The proposed algorithm, optimization and the testing is described in Section III, IV and V respectively. The experimental results for different datasets are reported in Section VI. The paper concludes with a discussion section.

II. RELATED WORKS

In this section, we discuss the related work from literature for the problem of cross-modal matching. In general, the objective is to learn a common representative space in which the data coming from various modalities can be compared [10]. The data modality can be as similar as near infrared-visible images [6] and as different as text-image [8], etc.

Canonical Correlation Analysis (CCA) [1][2] yields a feature space that can get the underlying common structure of the data. The lower dimensional feature space can be learned across the two modalities provided paired data is made available. To handle non-linear relationship between the data, "kernel trick" has been employed to devise Kernel CCA [2]. The additional constraint of paired data has been removed in the formulation of mean CCA, cluster CCA [11] and their kernelized versions. In this work, data is considered as given in classes, and correspondences between the classes are established. The heterogeneous domain adaption problem has been looked into in [12] where the correlation coefficient of the CCA method is used to effectively design a support vector machine (SVM) for classification purposes. The correlation transfer SVM (CT-SVM) [12] adapts the weights of the hyperplane by increasing (or decreasing) itself for higher (or lower) correlated feature dimensions. An extension using the Reduced Kernel [13] to handle non-linearity in data has also been provided [12]. In [8], an equivalence relation between cross-modal retrieval systems and isomorphic feature spaces has been mathematically devised which facilitates a two fold advantage of both low-level correlations and semantic abstraction.

Partial Least Squares (PLS) [3] is a regression model which uses linear maps to project input and output data into a low-dimensional subspace such that the covariance between the projections is maximized. In general, a linear mapping is used to match the input-output regression scores in the common subspace though extensions towards kernelized versions have shown significant improvement over the same. The Multi-view Discriminant Analysis [14] method jointly learns view-specific linear transformations so as to project the data into a common discriminant subspace. The algorithm is capable of handling instances from multiple views with the added benefit that the data may not be paired. Generalized Multiview Analysis (GMA) [15] mathematically formulates the cross-modal analysis problem as a constrained quadratic program and provides solution by generalized eigenvalue approach. GMA [15] is shown to be a supervised extension of CCA and an extension towards its kernel form has also been designed. Along the same philosophy as classical Linear Discriminant Analysis, the authors in [16] have proposed a novel method useful for set classification.

A different approach to this problem has been taken in [4][17][18] where sparse representation of the testing data in terms of the dictionaries learned jointly in the training stage is used for matching. The problem of super-resolution has been dealt with in [18], where, it is assumed that the low resolution and high resolution image patches have the same sparse representation. [19] propose Robust Sparse Coding algorithm which model the problem as a sparsity-constrained regression. The semi-coupled dictionary learning method [17] assumes that there exist a linear transformation between the sparse representation of the data of the two modalities. Studies in [20] and [21] have shown that a single linear operator may not be suitable for relating data across two different modalities, so the work in [4] uses two linear transformations to bring the sparse representations close together for comparison. Their method [4] is also applicable for synthesis problems. In StiM [22], a constrained dictionary learning problem is solved to handle multi-modal retrieval tasks. Label information [22] is utilized to discover the intra-class similarity for better learning of the mapping functions between the different modalities. Cross-modal data matching using both low rank transfer learning algorithms and auxiliary datasets have been found to be quite useful for matching task [23]. For document retrieval purpose, a Markov Random field of topic models is learned in [24] to link documents based on their similarity which allows efficient retrieval.

The problem of using unpaired data has been addressed in [25] and [26]. In [25], the algorithm learns consistent
feature representations from the two modalities that effectively utilizes class labels. To make better use of the block based image features, [25] also uses group prior knowledge. Though this method [25] can handle unpaired modality data it does not use the concept of dictionary learning. In [26], external face images are used to learn a common domain independent dictionary which can represent both the probe and gallery images. The concept is based on the fact that face images of different persons are more correlated in the same domain compared to correlation of the face images of same person across different domains. The work in [26] uses the maximum mean discrepancy theorem to learn a common subspace which rejects domain dependence.

Metric learning methods for cross modal data analysis have been studied in the works of [27], [28] and [29]. In [27], cross modal metric learning method uses both the positive and negative constraints of the dataset to map them into a common subspace for covariance maximization. Their approach performs better than the traditional CCA or PLS implementation due to its ability to leverage information from the negative pairs of data also. The work in [28] learns to project data from different modality into a hammering metric space for comparison by formulating the problem into a binary classification task using both positive and negative samples. A heterogeneous metric is learned in [29], where, by using the Joint Graph Regularized Heterogeneous Metric Learning algorithm, data from different modality along with the labels are integrated seamlessly to learn a high-level semantic metric for better cross-modal retrieval.

Cross modal data matching algorithms, pertaining to face recognition problems only, have been proposed in [30] [31] and [32]. In [30], the method maps images from the two different modalities into the common subspace by using two transformations. The work in [31] further expands this work by capturing the discriminative information from images of different classes. One advantage of this method [31] is that it is able to work with data from more than two modalities. The coupled discriminative feature learning method [32] learns the feature vectors adaptively from the training images itself. It extracts a series of filters which captures the inter-class and intra-class variance among the images effectively. Though the methods in [30][31] and [32] perform impressively for face recognition problems, it is yet to be seen how well they scale up with respect to other cross-domain matching tasks. The work in [33] uses hyperlingual words to capture the details of features across different modalities and within modalities to design superior face classification algorithms.

III. PROBLEM FORMULATION

In this section, we present the proposed Generalized Coupled Dictionary Learning (GCDL) approach for the task of cross-modal matching. The main idea is to learn the dictionaries for the two modalities in a coupled manner such that the sparse coefficients of the corresponding classes from the two modalities are maximally correlated in some transformed space and they are also discriminative enough to be used for matching applications.

Let the data corresponding to the two modalities be represented as $X$ and $Y$. Here, the data is of the form

$$X = \{X_1, X_2, \ldots, X_C\}$$

where $X_i$ is the subset of data $X$ which belongs to class $i$ and $C$ is the total number of classes. Note that the class labels need not be explicitly specified, since they are not used anywhere in the formulation. For example, we capture $M_1$ visible images of a subject and also $M_2$ near infra-red images in poor illumination conditions of the same subject. So the $M_1$ visible images from one modality and $M_2$ near infra-red images from the second modality form corresponding classes $X_i$ and $Y_i$. The class labels, i.e. the subject id cannot be used (and so is not required) since the training subjects are in general completely different from the subjects used during testing for this particular application. Similarly, we can write

$$Y = \{Y_1, Y_2, \ldots, Y_C\}$$

where $Y_i$ is the subset of data $Y$ which belongs to class $i$. Alternatively, $X_i$ and $Y_i$ can be represented as

$$X_i = \{x_1^i, x_2^i, \ldots, x_{|X_i|}^i\}$$

$$Y_i = \{y_1^i, y_2^i, \ldots, y_{|Y_i|}^i\}$$

where $|X_i|$ and $|Y_i|$ are the cardinalities of class $i$ in their respective modalities. Here $X \in \mathbb{R}^{d_1 \times N_1}$ and $Y \in \mathbb{R}^{d_2 \times N_2}$ are the feature matrices, with $d_1$ and $d_2$ are the length of the feature vectors and $N_1, N_2$ denotes the number of training images in the two modalities. The goal is to develop a dictionary learning technique which has the following properties:

1) The data from the two modalities can be sparsely reconstructed using the corresponding learnt dictionaries.
2) The approach should be able to seamlessly handle both paired as well as unpaired data from the two modalities.
3) The sparse coefficients should be discriminative enough for classification tasks.

Based on these criteria, we design different terms of the final objective function that we would like to optimize. Let $E_x$ and $E_y$ denote the energy terms associated with the data reconstruction error which address the first criterion. Let $E_{coup}$ denote the energy associated with the coupling term between the sparse coefficients of the two modalities which addresses the second and third criteria stated above. Now, we describe the specific forms of the different energy terms.

A. Data Reconstruction Term

The data reconstruction term ensures that the data in the two modalities are sparsely reconstructed using their respective dictionaries and corresponding sparse coefficients. Let $D_x \in \mathbb{R}^{d_1 \times K}$ and $D_y \in \mathbb{R}^{d_2 \times K}$ are the dictionaries corresponding to the two modalities with $K$ atoms each. The corresponding sparse coefficient matrices are denoted by $A_x \in \mathbb{R}^{K \times N_1}$ and $A_y \in \mathbb{R}^{K \times N_2}$ respectively. Thus, we would like to minimize the following objective functions

$$E_x = \operatorname{arg min}_{D_x, A_x} \| X - D_x A_x \|_F^2 + \alpha_x \| A_x \|_1$$

$$E_y = \operatorname{arg min}_{D_y, A_y} \| Y - D_y A_y \|_F^2 + \alpha_y \| A_y \|_1$$

(5)
Here, \( \| \cdot \|_F \) denotes the matrix Frobenius norm. The \( l_1 \) norm can be defined as follows: for \( x \in \mathbb{R}^n \), \( \| x \|_1 = \sum_{i=1}^{n} |x_i| \) where \( |x_i| \) stands for the absolute value operator. This ensures that the data \( X \) and \( Y \) can be reconstructed using their respective dictionaries and the sparse coefficients. For this part, the proposed approach shares the same philosophy as the work done in [4], [17] and [18].

### B. Sparse Coefficient Coupling Term

Different researchers have proposed different techniques for the coupling term to better describe the semantic relationship that exist between the different modalities. In [18], which deals with the super-resolution problem, the coupling term has been done in [4], [17] and [18].

The proposed variant of the standard CCA, the proposed coupling is also discriminative enough for classification tasks. Inspired can seamlessly handle both paired as well as unpaired data and their corresponding dictionaries and sparse coefficients. The coupling term is designed as discussed above. So we would require paired data, since they are concerned in getting corresponding classes to have low correlation. Thus, the energy term associated with the coupling term which is to be minimized is given as

\[
E_{\text{coup}} = \left( k_{i,j} - \frac{< T_x \lambda_{x,i}, T_y \lambda_{y,j} >}{\| T_x \lambda_{x,i} \|_2 \| T_y \lambda_{y,j} \|_2} \right)^2
\]

(10)

Here, \( k_{i,j} \in [-1, 1] \) is computed from the labeled data and it is close to 1 (i.e. high correlation) if \( i \) and \( j \) belongs to corresponding classes in the two modalities, otherwise, it is close to -1. Putting it all together, from (5) and (10), we have the final objective function as

\[
\begin{align*}
\text{arg min}_{D_x, D_y, \Lambda_x, \Lambda_y, T_x, T_y} & \quad \| X - D_x \Lambda_x \|_F^2 + \| Y - D_y \Lambda_y \|_F^2 + \alpha_x \| \Lambda_x \|_1 \\
& \quad + \alpha_y \| \Lambda_y \|_1 + \gamma \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left( k_{i,j} - \frac{< T_x \lambda_{x,i}, T_y \lambda_{y,j} >}{\| T_x \lambda_{x,i} \|_2 \| T_y \lambda_{y,j} \|_2} \right)^2 \\
& \quad \text{with } \| d_{x,i} \|_2 \leq 1, \| d_{y,j} \|_2 \leq 1
\end{align*}
\]

(11)

where, the extra constraints forces the dictionary elements to be of unit norm. We solve for the different unknowns in an iterative manner as described in the following section.

### IV. Optimization

Now, we describe how to solve the objective function in (11) for the different parameters. Specifically, we solve for the dictionaries \( D_x, D_y \), the sparse coefficients \( \Lambda_x, \Lambda_y \) and the two transformation matrices \( T_x, T_y \) iteratively by keeping all the other parameters fixed at the values of the previous iteration as described below.

#### A. Updating the dictionary atoms

Keeping all the other parameters, i.e. the sparse coefficients and the transformation matrices fixed, we solve for the two dictionaries \( D_x \) and \( D_y \) as follows

\[
\begin{align*}
\text{arg min}_{D_x} & \quad \| X - D_x \Lambda_x \|_F^2 \quad \text{s.t. } \| d_{x,i} \|_2 \leq 1 \quad \forall i \\
\text{arg min}_{D_y} & \quad \| Y - D_y \Lambda_y \|_F^2 \quad \text{s.t. } \| d_{y,j} \|_2 \leq 1 \quad \forall i
\end{align*}
\]

(12)

(13)

We have used the quadratic problem solver [34] to solve this constrained quadratic problem.

#### B. Updating the sparse coefficients

In this step, the dictionaries and the transformation matrices are kept fixed, and the sparse coefficients \( \Lambda_x \) and \( \Lambda_y \) are updated. Note that \( \Lambda_x \) and \( \Lambda_y \) can have different number of entries for unpaired data. When \( \Lambda_x \) is updated, \( \Lambda_y \) is assumed to be fixed at the previous iteration and vice versa. For updating \( \Lambda_x \), we observe that if the normalizing factors in the denominator of (9), i.e. \( \| T_x \lambda_{x,i} \|_2 \) and \( \| T_y \lambda_{y,j} \|_2 \) are both equal to 1, the numerator in (10) can be written as

\[
(T_y \lambda_{y,j})^T T_x \lambda_{x,i} = k_{i,j}
\]

(14)
Here $k_{i,j}$ is high if $i$ and $j$ belong to corresponding classes in the two modalities. This can be written in matrix form as

$$(T_y A_y)^T T_x A_x = K_x \quad \text{or} \quad P_x A_x = K_x \quad (15)$$

Here, $P_x = (T_y A_y)^T T_x$. If the number of dictionary atoms is $M$, then the transformation matrices $T_y, T_x$ are of dimension $K \times K$ and $A_y, A_x$ are of dimension $K \times N$ where $N$ is the number of data in the training set. Then the size of the matrix $K$ is $N \times N$. Similarly, we get, $P_y A_y = K_y$. After every iteration, we normalize $\|T_x \lambda_{x,i}\|_2$ and $\|T_y \lambda_{y,j}\|_2$ to be unit norm, making the denominator in (9) to be unity so that (14) approximately holds.

Putting everything together, the sparse coefficients are updated as follows

$$\arg \min_{\Lambda_x} \|X - D_x A_x\|_2^2 + \gamma \|K_x - P_x A_x\|_2^2 + \alpha_x \|A_x\|_1$$

Here $\gamma$ is a scalar which controls the relative contribution of the different terms. Finally, the two terms of (16) can be combined to obtain the objective function to solve for $A_x$.

$$\arg \min_{\Lambda_x} \|X - D_x A_x\|_2^2 + \gamma \|K_x - P_x A_x\|_2^2 + \alpha_x \|A_x\|_1$$

Similarly, the sparse coefficients for the other domain $A_y$ can be solved using

$$\arg \min_{\Lambda_y} \|Y - D_y A_y\|_2^2 + \gamma \|K_y - P_y A_y\|_2^2 + \alpha_y \|A_y\|_1$$

The above formulations have the same form as the standard sparse coding. We have used SPAMS [35] to solve for the sparse coefficients $A_x$ and $A_y$.

C. Updating the transformation matrices

In this step, the dictionaries and the sparse coefficients are kept fixed and the transformation matrices corresponding to the two modalities, namely $T_x, T_y$ are updated. Since, the transformation matrices are present in only the coupling part of the objective function, we essentially need to solve (9). This is the same as finding the solution for the following objective function [1] [2]

$$\arg \max_{t_x, t_y} \frac{t'_x \Sigma_{xy} t_y}{t'_x \Sigma_{xx} t_x \sqrt{t'_y \Sigma_{yy} t_y}}$$

$$\arg \min_{t_x, t_y} \frac{t'_x \Sigma_{xy} t_y}{t'_x \Sigma_{xx} t_x \sqrt{t'_y \Sigma_{yy} t_y}}$$

Here, $t_x$ and $t_y$ corresponds to the columns of the transformation matrices for the two modalities, $\Sigma_{xx} = E[\lambda_x \lambda'_x]$ and $\Sigma_{yy} = E[\lambda_y \lambda'_y]$ are the within-set covariance matrices and $\Sigma_{xy} = E[\lambda_x \lambda'_y]$ is the between-set covariance matrix. This is the same formulation as CCA [1] [2], but it can be applied for paired data only. For unpaired data, different variants of CCA like Mean-CCA (MCCA) and Cluster-CCA (CCCA) [11] have been proposed. The main difference between MCCA, CCCA and the standard CCA is in the definition of the covariance matrices which takes into account the class correspondence information. In this work we use a modified version of CCA which has an added discriminative component to make the solution better suited for classification applications. First, we describe the CCCC formulation and then describe the proposed modifications to include the discriminative information.

For CCCC [11], the different covariance matrices are defined as follows.

$$\Sigma_{xy} = \frac{1}{M} \sum_{i=1}^{C} \sum_{j=1}^{C} \lambda^i_j \lambda^j_i$$

$$\Sigma_{xx} = \frac{1}{M} \sum_{i=1}^{C} \sum_{j=1}^{C} \lambda^i_j \lambda^j_i$$

$$\Sigma_{yy} = \frac{1}{M} \sum_{i=1}^{C} \sum_{j=1}^{C} \lambda^i_j \lambda^j_i$$

with $M = \sum_{i=1}^{C} |X_i| |Y_i|$ is the total number of pairwise correspondences. Here, we propose two modifications of CCCA to make the coupling term more discriminative. In both the modifications, we incorporate the discriminative information in the between set covariance matrix $\Sigma_{xy}$.

1) Modification 1: In this modification, we try to keep the data from the different classes which are most similar to the class at hand as far as possible. This is done by pushing the centroids of the different classes having the greatest similarity far apart. The new between set covariance matrix is thus given by

$$\Sigma''_{xy} = \Sigma_{xy} - \eta \frac{1}{M} \Sigma'_{xy}$$

where

$$\Sigma'_{xy} = \sum_{i=1}^{C} \left( \sum_{m \in N_i^x} \mu^i_m \mu^i_m + \sum_{n \in N_i^y} \mu^i_n \mu^i_n \right)$$

where, $\eta$ is the trade-off parameter and $N_i^x$ and $N_i^y$ are the sets of nearest neighbors for the two modalities as defined above. Here $\mu^i_m$ and $\mu^i_n$ denotes the class centers for class $i$ of data $X$ and $Y$ respectively.

2) Modification 2: In this approach, instead of subtracting the centroid of the nearest neighbor class while constructing $\Sigma_{xy}$, we try to make similar data from different classes as far apart as possible. The intuition behind the proposed modification is that there can be data belonging to different classes which are similar to the current data, which will cause classification errors. So, the classification accuracy can improve if those data samples can be pushed far apart. So the new between set covariance matrix is given by

$$\Sigma''_{xy} = \Sigma_{xy} - \eta \frac{1}{M} \Sigma'_{xy}$$

where

$$\Sigma'_{xy} = \sum_{i=1}^{C} \left[ \sum_{j_1=1}^{C} \sum_{k_1 \in S_{j_1}} \lambda^i_{j_1} \lambda^j_{k_1} + \sum_{k_2=1}^{C} \sum_{j_2 \in S_{k_2}} \lambda^i_{j_2} \lambda^j_{k_2} \right]$$

with $\eta$ as is defined before. The sets $S_{j_1}$ is the collection of all the sparse coefficients in the $Y$ domain containing the nearest neighbors of $\lambda^j_{k_1}$ which does not belong to class $i$. The set $S_{k_2}$ can be similarly defined pertaining...
to elements in the X domain. In (24), the superscript refers to the class in which the data belongs and the subscript refers to the instance within that class. In both the modifications, $\Sigma'_{xy}$ replaces $\Sigma_{xy}$ in (19).

The problem in (19) can be solved by the generalized eigenvalue [2] approach. To solve (19), the corresponding Lagrangian is constructed by considering the constraints $t'_x \Sigma_{xx} t_x = 1$ and $t'_y \Sigma_{yy} t_y = 1$ (which makes the transformations scale invariant) and the Lagrangian is minimized by taking its derivative. Then the problem of finding $t_x$ reduces to solving the following equation [2]

$$\Sigma^{-1}_{xx} \Sigma_{xy} \Sigma^{-1}_{yy} \Sigma_{yx} t_x = \beta^2 t_x$$

(25)

where, $\beta$ is the Lagrangian multiplier and the corresponding $t_y$ can be calculated by substitution. Fig. 2 shows the flowchart of the training stage of the proposed approach. The output of the training stage is the two dictionaries $D_x$ and $D_y$ for the two modalities and the corresponding transformation matrices $T_x$ and $T_y$.

V. TESTING

The dictionaries and the transformation matrices that are learnt in the training stage are used for matching data from two different modalities during testing. First, given the data from the two modalities, $x$ and $y$, their sparse coefficients $\lambda_x$ and $\lambda_y$ are computed. The sparse coefficients are then transformed using their respective transformation matrices and the correlation is computed in the transformed space as follows

$$\rho_{x,y} = \frac{\langle T_x \lambda_x, T_y \lambda_y \rangle}{\|T_x \lambda_x\|_2 \|T_x \lambda_y\|_2}$$

(26)

Higher correlation implies greater similarity between the two data, while lower correlation implies less similarity.

For computing recognition score, first the sparse coefficient of the probe ($\lambda_p$) and those of the gallery ($\lambda_{g_i}, i = 1, 2, ..., N$) are computed. $N$ is the total number of data in the gallery. The correlation score $\rho_{p,g_i}$ using (26), between the probe with the $i^{th}$ gallery data is computed. The gallery data with the highest correlation with the probe data is its closest match. So all the $N$ correlation scores are sorted in descending order and the first one (having highest value) is taken as the correct match as given by the algorithm. This is repeated for all the probe data and the recognition score is computed.

VI. EXPERIMENTAL RESULTS

Here, we report the results of extensive experiments performed to evaluate the usefulness of the proposed approach for the task of cross-modal matching. We have used several publicly available cross-modal datasets for the evaluation, namely, CUHK dataset [5], HFB dataset [6], IXMAS multi-view action dataset [7], wiki dataset [11] and Multiple Features Dataset [9]. First, we use the standard experimental protocols for the various datasets, which by design use paired data, and compare the performance of the proposed approach with the state-of-the-art approaches. Next, we show the effectiveness of the proposed approach for unpaired data in the two modalities and also the importance of incorporating the discriminative term for classification tasks. In all our experimental results, we denote the proposed approach with modification 1 as GCDL1 and with modification 2 as GCDL2. Details of all the datasets are summarized in Table I.
The recognition rate obtained using the proposed approaches along with the state-of-the-art results are reported in Table II. All the other accuracies are taken directly from [4]. We see that the proposed approach performs favorably as compared to the state-of-the-art results.

B. Results on HFB database

Here, we test the performance of the proposed approach for the task of heterogeneous face recognition, specifically for the task of matching visible (VIS) images to near infrared (NIR) facial images. We use the HFB dataset [6] which consists of VIS and NIR images of 100 subjects. Example images from this dataset are shown in Fig. 3(B). For this experiment, we follow the same protocol as in [11], and consider 70 subjects for training and the rest 30 subjects for testing. 14 fiducial points are selected on each image and SIFT descriptors are computed at each location to get the feature vector of length 1792. PCA is used to reduce the feature dimension to 200. The parameters for our implementation are selected as: dict_size = 70, number of nearest neighbor η = 1 and η = 0.1. The results of our proposed approaches are reported in Table III. The performance metric used for comparison is rank-1 recognition (in fraction for comparison with existing approaches, with 0 denoting 0% recognition percentage and 100% indicating 100% recognition percentage). We have compared the results with the state-of-the-art approaches, specifically against CCA [11], MCCA [11], CCCA [11], CDL [4], MvDA, MvDA-VC [14] and SiL [22]. All the numbers for the other approaches are taken directly from the respective papers. We see that the proposed algorithms gives the highest recognition performance as compared to the state-of-the-art methods.
C. Results on IXMAS Multi-view action dataset

The proposed algorithm is further evaluated for cross-domain action recognition tasks. For this scenario, we consider the IXMAS [7] dataset. This multi-view action dataset consist of eleven actions each of which is performed three times by twelve actors. The actions are captured synchronously by five cameras named cam0 to cam4 (four side views and one top view) respectively (Fig. 3(C)). The task is to recognize the performed action in one view by studying the similar examples in another view. For extracting the features from the video frames in the IXMAS [7] dataset, bag of features model has been used like in [4]. Here, at most 200 cuboids from each video is extracted and represented by a 100-dimensional length vector. For reducing the dimensionality of the representation, principal component analysis is used. The reduced version of these descriptors are then clustered together into a codebook of 1000 visual words by using the k-means algorithm. The same experimental set up involving the leave-one-action-out strategy [4] has been used to evaluate the proposed algorithm.

Fig. 4 lists the results of the cross-view action recognition task obtained using the proposed algorithms GCDL1 and GCDL2. For the purpose of clarity, the average recognition rate for a particular target view over all the other source views is given in the figure. The performance is compared against several state of the art methods. All the other results have been directly taken from the author’s respective papers. It can be observed that the average recognition rate across all the different views for the methods CCA+CTSVM [12], Rd KCCCA+CTSVM [12], CDL [4], CCCA [11], SliM [22] and our proposed method comes out to be $0.67, 0.68, 0.627, 0.596, 0.497, 0.632$ respectively. The results from the proposed algorithm are highlighted in Table IV. From the results in Table IV, we observe that the average recognition rate of GCDL1 and GCDL2 across the two modalities compare favorably with the other state-of-the-art methods. SliM [22] performs slightly better than the proposed approaches for Image-to-Text experiments but overall highest performance is obtained using the proposed approaches.

D. Results on Wiki dataset

1) Full Wiki dataset: The wiki dataset [11] consists of text and image data pairs (2173 pairs for training and 693 pairs for testing) from ten different categories namely art, biology, geography, history, literature, media, music, royalty, sport and warfare. The data has been extracted from the Wikipedia’s ‘‘featured articles’’ section where each article normally has an accompanying image from the Wikipedia Commons directory. The labels for the text documents are represented as bag of words (BoW) and the images as bag of visual words (BoVw) [22]. For comparative studies we have used the same scenario as in [22] where two set ups have been used consisting of (500-dim BoVw, 1000-dim BoW) and (1000-dim BoVw, 5000-dim BoW). The evaluation method used in this case is the mean average precision (MAP) scores. MAP score is defined as the measure of whether the retrieved data belong to the same class (relevant) or does not belong to the same class (irrelevant) [22]. The parameters have been set as dict size = 50, nearest neighbor $k = 1$ and $\eta = 0.1$. The results from the proposed algorithm are shown in Table V.

2) Reduced Wiki dataset: To evaluate our algorithm with respect to the method in [12], we also test it on the reduced wikipedia dataset. In this, only five subject categories namely art and architecture, biology, literature, sport and warfare have been selected for evaluation. 100 instances from each category have been taken and the training set consist of two-thirds of the data while the rest have been used as the testing set [12]. The text is represented by the Latent Dirichlet Allocation (LDA) model [38] and the images by the bag of visual words model (BoVw). The dimension of the two feature representations are few examples from two different classes “warfare” and “sport” are shown in Fig. 5. For feature vector representation, the text documents are represented as bag of words (BoW) and the images as bag of visual words (BoVw) [22]. For comparative studies we have used the same scenario as in [22] where two set ups have been used consisting of (500-dim BoVw, 1000-dim BoW) and (1000-dim BoVw, 5000-dim BoW). The evaluation method used in this case is the mean average precision (MAP) scores. MAP score is defined as the measure of whether the retrieved data belong to the same class (relevant) or does not belong to the same class (irrelevant) [22]. The parameters have been set as dict size = 50, nearest neighbor $k = 1$ and $\eta = 0.1$. The results from the proposed algorithm are shown in Table V. From the results in Table V, we observe that the average recognition rate of GCDL1 and GCDL2 across the two modalities compare favorably with the other state-of-the-art methods. SliM [22] performs slightly better than the proposed approaches for Image-to-Text experiments but overall highest performance is obtained using the proposed approaches.
in [12] that use transfer learning techniques. The results in Table V for the methods in [12] have been directly taken from the paper and the best results from the paper have been reported in the table. For the image to text retrieval results, we observe that the best recognition rate of 91.71% is obtained by the proposed method. The text to image experiment shows that the approach gives an average recognition rate of 54.54% which is better than the other state-of-the-art approaches.

E. Results on the Multiple Features Dataset

The multiple features dataset [9] consist of a corpus of ten classes of handwritten digits. Each class has 200 images, thus the complete dataset has 2000 examples. Some example images from the dataset are shown in Fig. 6. The experimental set up for this dataset is a little different from the other experiments presented so far. Here, the images are represented in its different feature domains such as Fourier coefficients (76), profile correlations (216), Karhunen-Love coefficients (64), pixel average (240), Zernike moments (47) and morphological features (6). The numbers in brackets represent the feature dimensions. In this case, instead of learning across different modalities, we are required to learn across different feature domains. The source data represents images in one feature domain and the target data has images in some other feature domain.

As our method is based on dictionary learning technique, we do not consider the morphological features as its feature dimension is very low. For evaluation, we follow the same protocol as in [12] and randomly split the dataset into two-thirds (and one-thirds) per class to constitute the training (and testing) set respectively.

The recognition rate for a particular feature averaged over all the other features is given in Fig. 7. The performance metric used is rank-1 recognition percentage. In this experiment, the proposed algorithms GCDL1 and GCDL2 perform comparably with the CCA-CTSVM approach in [12]. The average recognition result across all the feature domains for the various methods namely CCA-CTSVM, RdKCCA-CTSVM [12], CCA+SV [11], CDL+SV [4], SliM [22], GCDL1 and GCDL2 are 82.16, 76.76, 79.46, 80.22, 79.44, 81.72 and 81.46 respectively.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Image to Text (Rank - 1%)</th>
<th>Image to Text (Rank - 1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear CCA + linear SVM [12]</td>
<td>87.00</td>
<td>44.75</td>
</tr>
<tr>
<td>Linear CCA + nonlinear SVM [12]</td>
<td>86.50</td>
<td>43.50</td>
</tr>
<tr>
<td>Linear CCA + CTSVM [12]</td>
<td>89.50</td>
<td>44.50</td>
</tr>
<tr>
<td>RdKCCA + linear SVM [12]</td>
<td>82.25</td>
<td>49.00</td>
</tr>
<tr>
<td>RdKCCA + CTSVM [12]</td>
<td>82.00</td>
<td>49.75</td>
</tr>
<tr>
<td>CCA + linear SVRM</td>
<td>90.50</td>
<td>51.71</td>
</tr>
<tr>
<td>CDL + linear SVM</td>
<td>81.41</td>
<td>46.26</td>
</tr>
<tr>
<td>SliM + linear SVM</td>
<td>91.51</td>
<td>43.03</td>
</tr>
<tr>
<td>GCDL1 + linear SVM</td>
<td>91.71</td>
<td>54.54</td>
</tr>
<tr>
<td>GCDL2 + linear SVM</td>
<td>90.30</td>
<td>53.73</td>
</tr>
</tbody>
</table>

Fig. 6. Sample examples of the digits from Multiple Features database [9]. Each row shows ten examples of a particular digit.
In this experimental setup, the number of training images per subject is taken to be different for the two modalities. This is a reasonable setting because in practical scenarios, getting equal number of images per subject in both modalities may not be always possible. Let \( n_{xtr} \) and \( n_{ytr} \) denote the number of images per subject in the two modalities, and in this setting, \( n_{xtr} \neq n_{ytr} \). To test the efficacy of the proposed algorithm for unpaired data, we have considered two different combinations = \( \{(n_{xtr}, n_{ytr}) = (2, 4), (4, 2)\} \). The testing set remains the same as in the original experimental setting in Subsection VI-B. The results of the proposed approach for this experimental setting are reported in Table VI. We compare the performance of the proposed approach with CCCA, KCCCA and CDL which are most closely related to our approach. CCCA and KCCCA can seamlessly handle unpaired data in two modalities. For CDL, we consider the less number of data among the two modalities and then consider the remaining data as paired. We see that for this data and experimental setting, the performance of the proposed approach compares favorably with the other state-of-the-art approaches. The performance metric used in Table VI is rank-1 recognition accuracy in percentage. It should also be pointed out that the recognition rate is lower than the one obtained in Table III but the degradation in our algorithm’s performance is graceful when switching from the paired-modality to the unpaired-modality scenario. Another possible reason for this decrease in performance for all the algorithms may be reduction in size of the training set.

Now, we test the performance of the proposed approach on unpaired data for the IXMAS [7] dataset. For this experiment, the number of training samples per action in one view (modality) is six and the number of training samples per action in the second view (modality) is four. So there is no one-to-one correspondence between the training data of the two modalities. Rather, there is correspondence between the classes of data in the two modalities. The average recognition rate is shown in Fig. 8. The performance metric used is rank-1 recognition percentage. We see that for this data and experimental setting, the performance of the proposed approach compares favorably with the other state-of-the-art approaches.

### F. Performance for unpaired data

In all the experiments so far, the data in the two modalities is essentially paired, with one-to-one correspondence between the data points. But the proposed GCDL approach is designed to work seamlessly for both paired as well as unpaired data. In this section, we test the performance of the proposed approach for unpaired data.

First, we test the performance on unpaired HFB dataset [6]. In this experimental setup, the number of training images per subject is taken to be different for the two modalities. This is a reasonable setting because in practical scenarios, getting equal number of images per subject in both modalities may not be always possible. Let \( n_{xtr} \) and \( n_{ytr} \) denote the number of images per subject in the two modalities, and in this setting, \( n_{xtr} \neq n_{ytr} \). To test the efficacy of the proposed algorithm for unpaired data, we have considered two different combinations = \( \{(n_{xtr}, n_{ytr}) = (2, 4), (4, 2)\} \). The testing set remains the same as in the original experimental setting in Subsection VI-B. The results of the proposed approach for this experimental setting are reported in Table VI. We compare the performance of the proposed approach with CCCA, KCCCA and CDL which are most closely related to our approach. CCCA and KCCCA can seamlessly handle unpaired data in two modalities. For CDL, we consider the less number of data among the two modalities and then consider the remaining data as paired. We see that for this data and experimental setting, the performance of the proposed approach compares favorably with the other state-of-the-art approaches. The performance metric used in Table VI is rank-1 recognition accuracy in percentage. It should also be pointed out that the recognition rate is lower than the one obtained in Table III but the degradation in our algorithm’s performance is graceful when switching from the paired-modality to the unpaired-modality scenario. Another possible reason for this decrease in performance for all the algorithms may be reduction in size of the training set.

Now, we test the performance of the proposed approach on unpaired data for the IXMAS [7] dataset. For this experiment, the number of training samples per action in one view (modality) is six and the the number of training samples per action in the second view (modality) is four. So there is no one-to-one correspondence between the training data of the two modalities. Rather, there is correspondence between the classes of data in the two modalities. The average recognition rate is shown in Fig. 8. The performance metric used is rank-1 recognition percentage. We see that for this data and experimental setting, the performance of the proposed approach compares favorably with the other state-of-the-art approaches.

### G. Performance for additional discriminative component

We have proposed two modifications to the standard CCCA [11] for the coupling term which relates the sparse coefficients of the two modalities. In this section, we evaluate the importance of these two discriminative terms for classification tasks. Specifically, we see how the proposed modifications GCDL1 and GCDL2 perform with respect to the GCDL (which is the default coupling term without the modifications).

We perform two experiments on the HFB [6] and IXMAS [7] using the same experimental setting as used in the unpaired data case in the previous section. Comparisons of the proposed modifications GCDL1 and GCDL2 with respect to GCDL for the two datasets are included in Table VI and Fig 8 respectively. For both the datasets, we observe that in general, both the modifications perform better than GCDL for classification tasks thus signifying the importance of the added discriminability in the coupling term.

### VII. Analysis of the algorithm

To analyze the effect of the different parameters \( \alpha_x \), \( \alpha_y \) and \( \gamma \) in the performance of the proposed algorithms, we performed

---

**Table VI. Results on HFB dataset (unpaired data). The performance metric used is rank-1 recognition accuracy in percentage.**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NIR Query Rank - 1%</th>
<th>VIS Query Rank - 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n_{xtr}, n_{ytr})</td>
<td>Avg. (n_{xtr}, n_{ytr})</td>
</tr>
<tr>
<td>CCCA [11]</td>
<td>66.66 67.50 47.50 57.08</td>
<td>54.16 46.66</td>
</tr>
<tr>
<td>KCCCA [11]</td>
<td>60.00 61.66 60.83 57.08 54.16 46.66</td>
<td>47.50</td>
</tr>
<tr>
<td>CDL [17]</td>
<td>65.00 63.33 64.16 60.00 60.00 55.00</td>
<td>57.50</td>
</tr>
<tr>
<td>GCDL</td>
<td>63.33 64.16 63.74 63.33 63.33 65.83</td>
<td></td>
</tr>
<tr>
<td>GCDL1</td>
<td>69.16 65.00 67.03 69.16 65.83 67.50</td>
<td></td>
</tr>
<tr>
<td>GCDL2</td>
<td>70.83 62.50 66.66 70.00 67.50 68.75</td>
<td></td>
</tr>
</tbody>
</table>
parameter analysis on different datasets by varying one of the parameters, and keeping the other two constant. We observed that the performance of the proposed algorithms is quite stable for a wide range of values of the parameters.

The proposed GCDL approach will have an added computation complexity compared to the other dictionary based methods due to the presence of the coupling term i.e., while updating the transformation coefficients. In each run of the algorithm, it will need to solve a generalized eigenvalue problem to get the transformation matrices $T_x$ and $T_y$. To do so it needs to compute three covariance matrices per iteration as shown in Eq. (20-24). Computation of the covariance matrix $\Sigma_{xy}$ is the most expensive step and hence we calculate its cost. As the number of dictionary atoms in $D_x$ and $D_y$ is $K$, the dimension of each sparse coefficient $\lambda_x$ or $\lambda_y$ will be of $K \times 1$ length. For computations of the terms in Eq. (20-24), the computational complexity is given as $O(MK^3)$, where $M$ is the total number of pairwise correspondence as defined above. The complexity remains the same even under proposed modifications due to them being basic addition and subtraction operations which will ideally be much less than computing the $\Sigma_{xy}$ term. To solve Eq. (19), an added computational cost of $O(K^3)$ is generated for matrix multiplication, inversion and eigenvalue decomposition as shown for Eq. (25). So the total complexity while updating the transformation coefficients is given as $O(MK^2) + O(K^3)$. As this process is subsequently run over many iterations (consider it to be $n$) until the convergence of the GCDL algorithm, the total complexity becomes $O(nMK^2) + O(nK^3)$.

We ran the experiments 10 times for different datasets and observed that the average computation time which includes both training and testing time is only slightly higher than CDL (2.2 sec. and 2.8 sec. for HFB dataset for CDL and GCDL1 respectively). The system configuration used is: Intel Core i7-3770 CPU @ 3.40 GHz 32 GB RAM. We also observe that the objective function (11) decreases monotonically with increasing iterations for both the proposed algorithms and we stop the iterations when the change in the objective function value in successive iterations falls below a predefined threshold epsilon ($\epsilon = 0.0050$). We observe that the algorithms usually converge within 10 iterations.

We also computed the average and standard deviation performance of the proposed algorithms over the 10 runs to observe how the performance varies with different runs. For the HFB dataset, the mean (standard deviation) of GCDL1 is 70.99 (5.28) as compared to 61.99 (4.92) and 65.41 (5.04) for CCCA and CDL respectively.

In the experiments, we have compared the proposed approaches with several supervised as well as unsupervised approaches. We made the following observations by analyzing the results of all the datasets: (1) In general, the supervised approaches like SliM performs better as compared to the unsupervised approaches like CCA, since they utilize the additional class information; (2) Though unsupervised, the dictionary learning approaches (SCDL and CDL) gave good performance for most datasets. (3) The proposed approaches gave consistent good performance on all the different datasets, though for some datasets, like Reduced Wikipedia (image to text), its performance is comparable with that of SliM; (4) We feel the better performance of the proposed approach can be attributed to increased flexibility in relating the sparse components from the two modalities using the maximizing correlation criterion, and use of class information to improve the discriminability.

VIII. Conclusion

In this paper, we present a generalized coupled dictionary learning technique for the task of cross-modal matching. The data in both modalities have sparse representation with respect to their own dictionaries. The coupling between the sparse coefficients is learnt in such a manner that the coefficients belonging to the same class for both modalities have high correlation in some transformed space. The approach can seamlessly handle both paired as well as unpaired data in the two modalities in the same framework. We also proposed two modifications to the coupling term so as to make it discriminative for classification tasks. Extensive experiments on different datasets such as HFB, IXMAS, Wikipedia text-image, CUHK and Multiple Features dataset show the effectiveness of the proposed approach.

REFERENCES
