Simultaneous Semi-Coupled Dictionary Learning for Matching RGBD data

Nilotpal Das, Devraj Mandal, Soma Biswas
Indian Institute of Science, Bangalore, India
{nilotpal30@ece, devraj89@ee, soma.biswas@ee}.iisc.ernet.in

Abstract

Matching with hidden information which is available only during training and not during testing has recently become an important research problem. Matching data from two different modalities, known as cross-modal matching is another challenging problem due to the large variations in the data coming from different modalities. Often, these are treated as two independent problems. But for applications like matching RGBD data, when only one modality is available during testing, it can reduce to either of the two problems. In this work, we propose a framework which can handle both these scenarios seamlessly with applications to matching RGBD data of Lambertian objects. The proposed approach jointly uses the RGB and depth data to learn an illumination invariant canonical version of the objects. Dictionaries are learnt for the RGB, depth and the canonical data, such that the transformed sparse coefficients of the RGB and the depth data is equal to that of the canonical data. Given RGB or depth data, their sparse coefficients corresponding to their canonical version is computed which can be directly used for matching using a Mahalanobis metric. Extensive experiments on three datasets, EURECOM, VAP RGB-D-T and Texas 3D Face Recognition database show the effectiveness of the proposed framework.

1. Introduction

In many practical scenarios, all the data that is available during the training stage may not be available during testing. This is analogous to human teaching approach, where during the training stage the teacher provides some additional help with figures, examples, etc. which are not available during testing, but which still help in the learning process. This is popularly known as Learning Using Privileged Information (LUPI) [26] or testing with hidden information (used in the remainder of the paper). Cross-modal matching is another important and challenging problem in the field of computer vision and pattern recognition with wide range of applications like photo-sketch recognition, text-image matching, etc. These two problems are generally treated as separate and different approaches have been proposed for handling both the scenarios. For example, algorithms like SVM+ [26], ITML+ [32] have been proposed to specifically handle testing with hidden information. A considerable research has also been done in the area of cross-modal matching, and Canonical Correlation Analysis [12], dictionary learning based methods such as Semi-Coupled Dictionary Learning [28] and Coupled Dictionary Learning [13] have been proposed.

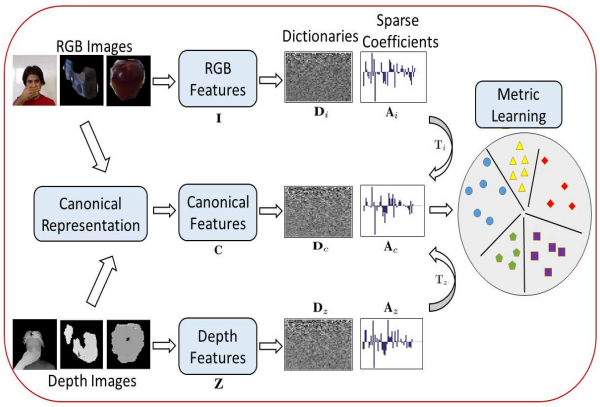

Figure 1. Flowchart of the proposed simultaneous semi-coupled dictionary learning (S^2CDL) approach.

Now-a-days, with the easy availability of RGBD sensors [2][1], we can simultaneously capture both RGB images as well as the depth information, which are together known as RGBD data. However, during testing, both the RGB and depth may not be available. In realistic scenarios, we may not even have prior information as to which modality will be available. Thus depending on the scenario, if only one modality is available, the problem can reduce to matching with hidden information (when RGB-RGB or depth-depth are available) or cross-modal matching (when RGB-depth are available). The same scenarios can occur in other applications also, like when we have both text and image data during training, but only one is available during testing. This calls for a framework which can handle both the scenarios simultaneously.

In this work, we propose a novel algorithm, S^2CDL (Si-

multaneous Semi-Coupled Dictionary Learning) which can handle matching with hidden information as well as cross-modal matching in the same framework. Given RGBD data of Lambertian objects, first a canonical representation is learnt from both the RGB and the depth information. We use the frontally illuminated re-lighted image as the canonical representation, since it incorporates the information about the albedo and the surface normals, which are the intrinsic characteristics of the object and is also robust to illumination variations. Separate dictionaries for the RGB, depth and canonical data are learnt such that the transformed sparse representation of the RGB and the depth data are same as that of the canonical representation. Finally, a Mahalanobis metric is used for matching the sparse coefficients of the canonical representation. Figure 1 shows a flowchart of the proposed approach. Extensive experiments performed on three RGBD datasets, namely EURECOM face database [19], VAP RGB-D-T dataset [20] and Texas 3D Face Recognition database [10] [9] [8] and comparisons with state-of-the-art approaches show the effectiveness of the proposed framework. The contributions of this work are as follows

- We have proposed a novel framework, named S^2CDL to simultaneously handle both matching with hidden information as well as cross-modal matching.
- The proposed approach handles the coupling between two domains through an intermediate robust, illumination invariant representation which incorporates its intrinsic characteristics.
- To the best of our knowledge, this is the first time that dictionary learning approach has been used to handle hidden information.

The rest of the paper is organized as follows. Section 2 discusses the related works in literature. The training and testing stages of the proposed approach are described in Section 3 and 4 respectively. The experimental results are given in Section 5 and the paper concludes with a summary.

2. Related Works

Both learning with hidden information and cross-modal matching are active areas of research in the field of computer vision. The LUPI scenario is well studied in the works of [24], [26] and [32] which deal with the transfer of information from the privileged data to the original modality. In [26], privileged information is used to improve the performance of SVM binary classification tasks. In addition to the objective function of SVM+ [26] being convex, an added advantage of increased speed of convergence towards the final solution, is also achieved. The work in [25] improves upon the student-teacher interaction [26] enabling the classifier to better learn which objects are similar with lighter computation loads. Extreme learning machines with privileged information (ELM+) [33] has been shown to work in the same scenario as SVM and SVM+ ([26]) and shows improvement over both of them. In [27], attributes such as face shape, face acne, etc. are used in a privileged manner to estimate the age of subjects. In [30], hidden information is used in two settings to design a better classifier by considering them either as secondary features or secondary targets. [24] proposes two maximum-margin techniques which enable the classifier to learn the hard and easy examples from the privileged information and use this additional knowledge to design a better predictor. Rank Transfer [24] is shown to be a generic method which can handle any form of privileged information such as attributes, annotator rationales, object bounding boxes or even textual descriptions. The work in [7] utilizes the distance metric learning algorithm ITML (Information Theoretic Metric Learning) [6] two times to handle privileged information - first to remove the outlier pairs whose distances are greater than some predefined threshold value by applying it on the privileged information and then again applying it on the reduced main dataset. The ITML with privileged information (ITML+) algorithm [32] designs a slack function to handle the additional data.

Several approaches have been proposed for matching data coming from different modalities, such as Canonical Correlation Analysis (CCA) [12][11] or dictionary learning based methods such as Semi-Coupled Dictionary Learning (SCDL) [28] and Coupled Dictionary Learning (CDL) [13]. CCA [12][11] learns a lower dimensional feature space for optimal representation of the data from two modalities provided paired data is made available. The non-linear relationships between data can be handled by using kernel trick while the constraint of paired data in either modalities has been removed in the Cluster CCA formulation [22]. Dictionary learning based methods [28][13] represent the given data in sparse format and try to find transformations that bridge the gap between these two sparse representations. The coupled dictionary learning method [13] has also been shown to handle synthesis problems well. Deep learning techniques have also been used for multi-modal retrieval in [29] where the deep learning architecture has been used to effectively learn the mapping function between heterogeneous sources for capturing intra-modal and inter-modal relationships.

3. Proposed Method

Here, we describe in detail the proposed framework for matching with hidden information and cross-modal matching for RGBD data of Lambertian objects. The approach has a training and testing stage. Given the RGB and the depth information in the training stage, we learn a canoni-

cal representation as described next.

3.1. Canonical Representation

Assuming that the objects under consideration are Lambertian, the intensity I at any pixel i is given by Lambert's Cosine law as follows [5],

$$I_i = \rho_i \max(\mathbf{n}_i^T \mathbf{s}, 0) \tag{1}$$

where, ρ_i is the albedo of pixel i, $\mathbf{n}_i$ denotes the surface normal and $\mathbf{s}$ the light source direction, assuming there is a single light source placed at infinity. But in realistic scenarios, the object can be illuminated by multiple light sources placed at arbitrary locations. Lee *et al.* [16] showed that an arbitrarily illuminated object can be approximated as the superposition of nine different images illuminated by nine sources placed at predefined positions. Thus if $\{\mathbf{s}_1, \ldots, \mathbf{s}_9\}$ be the nine illumination directions, the corresponding image formation equation becomes

$$I_i = \sum_{k=1}^{9} \alpha_k I_i^k, \qquad I_i^k = \rho_i \max(\mathbf{n}_i^T \mathbf{s}_k, 0) \tag{2}$$

where the nine illumination directions are given by [16] (0,0), (49,17), (-68,0), (73,-18), (77,37), (-84,47), (-84,-47), (82,-56), (-50,-84) in spherical coordinates. I_i^k is the intensity of pixel i for the k^{th} light source direction. The combining coefficients can be computed in the least square manner using the surface normals computed from the depth information and constant albedo [5]. Using this estimate $\hat{\alpha_k}$ and the surface normals, we recompute the albedo as follows

$$\rho_i = \frac{I_i}{\sum_{k=1}^{9} \hat{\alpha_k} \max(\mathbf{n}_i^T \mathbf{s}_k, 0)} \tag{3}$$

Though there are approaches to obtain a more robust estimate of the albedo [5], we have used the computed albedo directly without any further processing.

In this work, we consider the re-lighted image under frontal illumination condition as the canonical representation. This representation is not only robust to illumination variations, but also incorporates the intrinsic characteristics of the underlying Lambertian object, i.e. the albedo and the surface normal. So given the surface normal and the estimated albedo, we obtain the canonical representation C_i as follows

$$C_i = \rho_i \max(\mathbf{n}_i^T \mathbf{s}_0, 0) \tag{4}$$

where $\mathbf{s}_0 = (0, 0, 1)$ is the frontal light source.

3.2. S^{2}CDL Algorithm

Here we describe the proposed S^2CDL algorithm which utilizes the canonical representation for matching RGBD data in both the scenarios, i.e. matching with hidden information and cross-modal matching. Since in the testing stage, the data can come from any modality which is not known a priori, we propose to

- Learn the relation between the RGB data and the canonical representation.
- Learn the relation between the depth data and the canonical representation.
- Learn the above relations simultaneously so that cross-modal matching can be incorporated in the same framework.

During testing, if the data comes from the same modality (i.e. RGB-RGB or depth-depth matching), the first two conditions ensure that information from the other hidden modality is utilized, whereas the third condition enables the proposed algorithm to work in the cross modal scenario. Considering the above noted objectives, we propose a dictionary learning algorithm (S^2CDL) for which the objective function is given by

$$E = E_d + E_c \tag{5}$$

Here the first term represents the data reconstruction term and the second term the data coupling term, which we will describe in details below.

Data Representation Term: The first term in the objective function (5) ensures that the input data can be reconstructed using the learnt dictionaries. Here we learn three dictionaries, corresponding to the RGB data, depth data and canonical data. So the data representation terms have the form

$$E_{d,x} = ||\mathbf{X} - \mathbf{D}_x \mathbf{\Lambda}_x||_F^2 + \alpha ||\mathbf{\Lambda}_x||_1 \tag{6}$$

where $\mathbf{X} = \{\mathbf{I}, \mathbf{Z}, \mathbf{C}\}$ denote that $\mathbf{X}$ can be the image data $\mathbf{I}$, depth data $\mathbf{Z}$ or the canonical data $\mathbf{C}$. $\mathbf{D}_x$ and $\mathbf{\Lambda}_x$ denote the corresponding dictionary and the sparse coefficients.

Coupling Term: The second term in the objective function (5) denotes how the sparse coefficients are related to one another. Here we have the sparse coefficients for $\mathbf{I}$, $\mathbf{Z}$ and $\mathbf{C}$. As discussed above, we would like to learn the relationship between the RGB and depth data with the canonical representation simultaneously. For this, we propose to learn two transformation matrices $\mathbf{T}_i$ and $\mathbf{T}_z$ simultaneously, such that when applied to the RGB and the depth data, they get transformed to the same coefficients as that of the canonical representation. Thus the coupling terms are given as below

$$E_{c,i} = ||\mathbf{T}_i \mathbf{\Lambda}_i - \mathbf{\Lambda}_c||_F^2 \tag{7}$$

$$E_{c,z} = ||\mathbf{T}_z \mathbf{\Lambda}_z - \mathbf{\Lambda}_c||_F^2 \tag{8}$$

The two modalities are related through the robust intermediate canonical representation, which incorporates all the intrinsic characteristics of the underlying Lambertian object. Thus combining both the data representation as well as the coupling terms, we get the final objective function as

$$\begin{aligned}\arg\min_{A} &||\mathbf{I}-\mathbf{D}_i\mathbf{\Lambda}_i||_F^2+||\mathbf{Z}-\mathbf{D}_z\mathbf{\Lambda}_z||_F^2+||\mathbf{C}-\mathbf{D}_c\mathbf{\Lambda}_c||_F^2\\ &+\gamma_1||\mathbf{T}_i\mathbf{\Lambda}_i-\mathbf{\Lambda}_c||_F^2+\gamma_2||\mathbf{T}_z\mathbf{\Lambda}_z-\mathbf{\Lambda}_c||_F^2\\ &+\alpha_1||\mathbf{\Lambda}_i||_1+\alpha_2||\mathbf{\Lambda}_z||_1+\alpha_3||\mathbf{\Lambda}_c||_1\\ &+\beta_1||\mathbf{T}_i||_F^2+\beta_2||\mathbf{T}_z||_F^2\\ &s.t., ||d_{i,j}||_2\leq 1, ||d_{z,j}||_2\leq 1 \;\&\; ||d_{c,j}||_2\leq 1\;\forall j \end{aligned} \quad (9)$$

where $A=\{\mathbf{D}_i,\mathbf{D}_z,\mathbf{D}_c,\mathbf{\Lambda}_i,\mathbf{\Lambda}_z,\mathbf{\Lambda}_c,\mathbf{T}_i,\mathbf{T}_z\}$. We solve the above objective function and obtain the values of the three dictionaries $(\mathbf{D}_i,\mathbf{D}_z,\mathbf{D}_c)$ and their respective transformations $(\mathbf{T}_i,\mathbf{T}_z)$ as described next.

3.3. Optimization for the Proposed Algorithm

Here we describe how to solve for the different unknowns in the objective function (9), namely the three dictionaries, the corresponding sparse coefficient matrices and the two transformation matrices. Though the above objective function is not jointly convex, it is convex with respect to each of the terms while keeping all the other terms fixed. So to solve for the unknowns, we follow an iterative approach as described below:

Initialization: We initialize the dictionaries $\mathbf{D}_i$, $\mathbf{D}_z$ and $\mathbf{D}_c$ using the standard KSVD formulation [3] as given below.

$$\min_{\mathbf{D}_i,\mathbf{D}_z,\mathbf{D}_c,\mathbf{\Lambda}}\left|\left|\begin{bmatrix}\mathbf{I}\\ \mathbf{Z}\\ \mathbf{C}\end{bmatrix}-\begin{bmatrix}\mathbf{D}_i\\ \mathbf{D}_z\\ \mathbf{D}_c\end{bmatrix}\mathbf{\Lambda}\right|\right|_F^2+\alpha||\mathbf{\Lambda}||_1 \quad (10)$$

The respective sparse coefficients $\mathbf{\Lambda}_i$, $\mathbf{\Lambda}_z$ and $\mathbf{\Lambda}_c$ can be initialized by solving the sparse coding separately for $\mathbf{I}$, $\mathbf{Z}$ and $\mathbf{C}$. The transformation matrices $\mathbf{T}_i$ and $\mathbf{T}_z$ are set to the identity matrix.

Updating the dictionaries: In this step, the transformation matrices and the sparse coefficients are kept fixed to the value at the previous iteration and the dictionaries are updated as follows:

$$\arg\min_{\mathbf{D}_x}||\mathbf{X}-\mathbf{D}_x\mathbf{\Lambda}_x||_F^2 \qquad s.t.\ ||d_{x,j}||_2\leq 1;\forall j \quad (11)$$

for all $\mathbf{X}=\{\mathbf{I},\mathbf{Z},\mathbf{C}\}$. This is a constrained quadratic problem with respect to $\mathbf{D}_x$ and the solution can found by using Lagrange dual techniques [15]. We have used the SPAMS package [18] for computing the dictionaries.

Updating the sparse representations: The sparse representations are computed by keeping all the other terms, i.e. the dictionaries and the two transformations fixed at the value of the previous iteration. Thus the sparse coefficients for the RGB data can be computed by solving the following objective function

$$\min_{\mathbf{\Lambda}_i}||\mathbf{I}-\mathbf{D}_i\mathbf{\Lambda}_i||_F^2+\gamma_1||\mathbf{T}_i\mathbf{\Lambda}_i-\mathbf{\Lambda}_c||_F^2+\alpha_1||\mathbf{\Lambda}_i||_1 \quad (12)$$

This is equivalent to solving the following function

$$\min_{\mathbf{\Lambda}_i}\left|\left|\begin{bmatrix}\mathbf{I}\\ \sqrt{\gamma_1}\mathbf{\Lambda}_c\end{bmatrix}-\begin{bmatrix}\mathbf{D}_i\\ \sqrt{\gamma_1}\mathbf{T}_i\end{bmatrix}\mathbf{\Lambda}_i\right|\right|_F^2+\alpha_1||\mathbf{\Lambda}_i||_1 \quad (13)$$

We can write a similar objective function for computing the sparse coefficients for the depth data. The sparse coefficients of the canonical representation can be computed using the following objective function

$$\begin{aligned}&\min_{\mathbf{\Lambda}_c}||\mathbf{C}-\mathbf{D}_c\mathbf{\Lambda}_c||_F^2+\gamma_1||\mathbf{T}_i\mathbf{\Lambda}_i-\mathbf{\Lambda}_c||_F^2\\ &\quad+\gamma_2||\mathbf{T}_z\mathbf{\Lambda}_z-\mathbf{\Lambda}_c||_F^2+\alpha_3||\mathbf{\Lambda}_c||_1\\ &=\min_{\mathbf{\Lambda}_c}\left|\left|\begin{bmatrix}\mathbf{C}\\ \sqrt{\gamma_1}\mathbf{T}_i\mathbf{\Lambda}_i\\ \sqrt{\gamma_2}\mathbf{T}_z\mathbf{\Lambda}_z\end{bmatrix}-\begin{bmatrix}\mathbf{D}_c\\ \sqrt{\gamma_1}\mathbf{I}_D\\ \sqrt{\gamma_2}\mathbf{I}_D\end{bmatrix}\mathbf{\Lambda}_c\right|\right|_F^2+\alpha_3||\mathbf{\Lambda}_c||_1\end{aligned} \quad (14)$$

where $\mathbf{I}_D$ is the identity matrix. All the three modified objective functions have the same form as that of the standard sparse coding problem which can be solved by using any of the common solvers such as SPAMS [18].

Updating the transformations: To solve for the transformation vectors $\mathbf{T}_i$ and $\mathbf{T}_z$ in (9), the dictionaries and the sparse coefficients are kept fixed at the value of the previous iteration. Thus the objective function reduces to solving the following two equations

$$\begin{aligned}&\min_{\mathbf{T}_i}\gamma_1||\mathbf{T}_i\mathbf{\Lambda}_i-\mathbf{\Lambda}_c||_F^2+\beta_1||\mathbf{T}_i||_F^2\\ &\min_{\mathbf{T}_z}\gamma_2||\mathbf{T}_z\mathbf{\Lambda}_z-\mathbf{\Lambda}_c||_F^2+\beta_2||\mathbf{T}_z||_F^2\end{aligned} \quad (15)$$

The second term is an extra regularization term which is used to avoid over-fitting with respect to the training data. The above equations have closed form solutions given by

$$\begin{aligned}\mathbf{T}_i&=\mathbf{\Lambda}_c\mathbf{\Lambda}_i^T\left(\mathbf{\Lambda}_i\mathbf{\Lambda}_i^T+(\beta_1/\gamma_1)\,\mathbf{I}_D\right)^{-1}\\ \mathbf{T}_z&=\mathbf{\Lambda}_c\mathbf{\Lambda}_z^T\left(\mathbf{\Lambda}_z\mathbf{\Lambda}_z^T+(\beta_2/\gamma_2)\,\mathbf{I}_D\right)^{-1}\end{aligned} \quad (16)$$

The above steps are repeated till convergence.

3.4. Discriminative Canonical Sparse Coefficients

Once the sparse coefficients of the canonical representation are obtained, we compare them using Mahalanobis metric. Metric learning methods try to learn a linear transformation so that in the feature space, more importance is given to the relevant dimensions while discarding the non-relevant ones. In this work, we use Large Scale Metric

Learning (LSML) [14] which learns a distance metric by utilizing the concepts of equivalence constraints and provides an optimized solution that is tractable even for large volumes of data. The metric can be learned by utilizing the covariance matrices between the matched pairs (denoted by Σ_1) and non-matched pairs (denoted by Σ_2) as $\mathbf{M} = (\Sigma_1^{-1} - \Sigma_2^{-1})$. If $\mathbf{\Lambda}_c^p$ and $\mathbf{\Lambda}_c^g$ be the sparse coefficients of the canonical representation of the probe and gallery data (can be RGB or depth), then the distance between them is given by

$$d^2(p, g) = (\mathbf{\Lambda}_c^p - \mathbf{\Lambda}_c^g)^T \mathbf{M} (\mathbf{\Lambda}_c^p - \mathbf{\Lambda}_c^g) \tag{17}$$

The training stage of S^2CDL algorithm is given in Algorithm 1.

Algorithm 1 Training Stage of S^2CDL algorithm.

1: **Input** - RGB data $\mathbf{I}$, depth data $\mathbf{Z}$
2: Compute the canonical representation $\mathbf{C}$ from the RGB and depth data.
3: Initialize the dictionaries $\mathbf{D}_x^0$, sparse coefficients $\mathbf{\Lambda}_x^0$ and the transformation matrices $\mathbf{T}_i^0$, $\mathbf{T}_z^0$, $(x = \{i, z, c\})$
4: **while** convergence **do**
5: Update the dictionaries $\mathbf{D}_x$, by using (11).
6: Update the sparse coefficients $\mathbf{\Lambda}_x$ by using (13) and (14).
7: Update the transformation matrices $\mathbf{T}_i$ and $\mathbf{T}_z$ by using (16).
8: **end while**
9: Learn the Mahalanobis metric $\mathbf{M}$ on the sparse coefficient $\mathbf{\Lambda}_c$ of the canonical representation.
10: **Output:** Dictionaries $\mathbf{D}_i$, $\mathbf{D}_z$, $\mathbf{D}_c$, transformation matrices $\mathbf{T}_i$, $\mathbf{T}_z$ and Mahalanobis metric $\mathbf{M}$.

4. Testing

The proposed approach can be used for matching RGBD data when one of the modalities which was available during training is missing during testing. Thus, it can be used for matching same modality data (i.e. RGB with RGB or depth with depth when matching with hidden information) or data across modalities (i.e. RGB with depth data which is the cross-modal scenario). So the testing setup depends upon the scenario in which the algorithm is currently functioning. Next, we will describe the two scenarios in details.

4.1. Matching with hidden information

In this scenario, one of the modality which was present during training is absent during testing, i.e. for testing RGB with RGB or depth with depth. So the extra information that is available during training should be utilized so that it is beneficial during testing. Consider the RGB-RGB matching scenario, where the depth information is available only during training and not during testing. Let the RGB probe and gallery data be denoted by $\mathbf{I}_p$ and $\mathbf{I}_g$. Using the learned dictionary $\mathbf{D}_i$, the corresponding sparse representations $\mathbf{\Lambda}_i^p$ and $\mathbf{\Lambda}_i^g$ are computed. The learned transformation $\mathbf{T}_i$ is used to compute the corresponding canonical representation, $\mathbf{\Lambda}_c^p$ and $\mathbf{\Lambda}_c^g$. Since the canonical data has been generated by using the depth map, this way the transfer of information takes place. The Mahalanobis distance between the canonical representations is taken as the distance between the probe and gallery data. The same procedure is followed for matching depth with depth, when RGB images are only available during training and not during testing. The algorithm for this testing scenario is given in Algorithm 2.

Algorithm 2 S^2CDL: Matching with Hidden Information.

1: **Goal** - Matching RGB probe $\mathbf{I}_p$ with RGB gallery $\mathbf{I}_g$
2: Compute the sparse coefficients for the probe as follows

$$\arg\min_{\mathbf{\Lambda}_i^p} ||\mathbf{I}_p - \mathbf{D}_i \mathbf{\Lambda}_i^p||_F^2 + \alpha_1 ||\mathbf{\Lambda}_i^p||_1$$

Similarly, compute sparse coefficients $\mathbf{\Lambda}_i^g$ for gallery.
3: Compute the canonical representation of probe using the relation:

$$\mathbf{\Lambda}_c^p = \mathbf{T}_i \mathbf{\Lambda}_i^p$$

Similarly, compute the canonical representation $\mathbf{\Lambda}_c^g$ of the gallery.
4: Compute the Mahalanobis distance between the canonical representations of probe and gallery using (17).

4.2. Cross modal Matching

This scenario involves matching RGB with depth data. Consider the scenario, where RGB probe $\mathbf{I}_p$ and depth gallery $\mathbf{Z}_g$ data are available during testing. In this work, the matching is done by converting both the data into the canonical representation. Given the data, the learned dictionaries $\mathbf{D}_i$ and $\mathbf{D}_z$ are used to generate the sparse representation of the probe ($\mathbf{\Lambda}_i^p$) and gallery ($\mathbf{\Lambda}_z^g$) respectively. The corresponding canonical representations are generated using the corresponding transformations $\mathbf{T}_i$ and $\mathbf{T}_z$. The Mahalanobis distance between the canonical representations gives the distance between the probe and gallery. The algorithm for cross modal testing scenario is given in Algorithm 3.

5. Experiments

We evaluate the proposed approach for matching RGBD face data for the two scenarios on three different datasets, the EURECOM database [19], VAP RGB-D-T database [20] and Texas 3D Face Recognition database [10] [9] [8]. Performance comparison with state-of-the-art approaches is also reported.

Algorithm 3 S^2CDL: Cross-Modal Matching.

1: **Goal:** Matching RGB probe $\mathbf{I}_p$ with depth gallery $\mathbf{Z}_g$

2: Compute the sparse coefficients for the RGB probe $\mathbf{\Lambda}_i^p$ and depth gallery $\mathbf{\Lambda}_z^g$ as follows

$$\arg\min_{\mathbf{\Lambda}_i^p} ||\mathbf{I}_p - \mathbf{D}_i\mathbf{\Lambda}_i^p||_F^2 + \alpha_1||\mathbf{\Lambda}_i^p||_1$$

$$\arg\min_{\mathbf{\Lambda}_z^g} ||\mathbf{Z}_g - \mathbf{D}_z\mathbf{\Lambda}_z^g||_F^2 + \alpha_2||\mathbf{\Lambda}_z^g||_1$$

3: Compute the canonical representation of probe and gallery using the relations:

$$\mathbf{\Lambda}_c^p = \mathbf{T}_i\mathbf{\Lambda}_i^p$$
$$\mathbf{\Lambda}_c^g = \mathbf{T}_z\mathbf{\Lambda}_z^g$$

4: Compute the Mahalanobis distance between the canonical representations of probe and gallery using (17).

5.1. Evaluation on EURECOM Dataset

The EURECOM dataset [19] consist of Kinect data from 52 subjects, each having 7 frontal RGB and 7 depth images (top row - Figure (2)). For the experiment, the whole data is divided into a training and testing set, each having data from 26 subjects with no overlapping subjects. SIFT features [17] computed from 9 fiducial locations of the face (corners of mouth, nose, lip) are concatenated together and used as the feature for both RGB as well as depth data. Studies in [19] [4] have shown that features extracted from fiducial landmarks on range images (that use the gray-scale of every pixel to represent the depth of a scan as used in this work), can be used effectively for facial recognition.

Ideally, the canonical representation computed for a single subject under different conditions should be identical since it depends only on the intrinsic characteristics. But due to occlusions, errors in computation of albedo, outliers in depth data, the canonical representations computed from the same subjects are not identical. So for all our experiments, we consider the canonical representation computed from the neutral condition (i.e. without occlusions) as the representation for that subject. The canonical representation of one subject is shown in the middle in Figure 3, and the other images shows the data from the same subject under different conditions.

For each run of the experiment, we consider one of the 7 images (RGB or depth depending on the scenario) of a subject of Session 2 in the gallery and all the 7 images (RGB or depth depending on the scenario) from Session 1 as the probe. Thus, in each experiment, there is only one image per subject in the gallery. The experiment is run by choosing each of the 7 images in the gallery and the Rank 1 recognition performance (%) averaged over all the runs is reported. Table 1 shows the accuracy of the proposed algorithm for the task of cross-modal matching on EURECOM dataset. The second and third columns refer to the two cases when RGB is the gallery and depth is the probe and vice versa. Comparison with several recent cross-modal analysis techniques is also reported. Specifically, we compare against (1) Canonical Correlation Analysis (CCA) [12], its recent variant Cluster CCA [22], which reports significant improvement over CCA for cross-modal matching applications; (2) Generalized Multiview Analysis (GMA) [23] which is proposed to handle the cross view classification and retrieval; and (3) Dictionary learn-

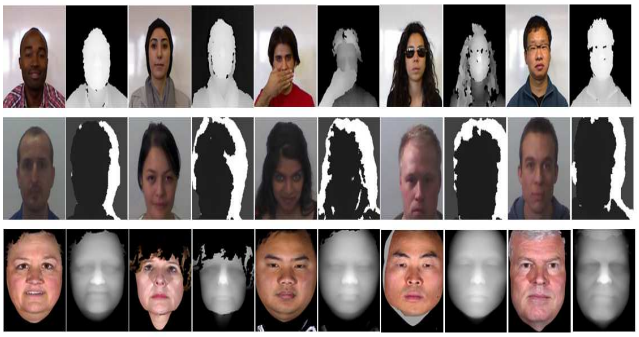

Figure 2. Sample RGB and depth images of few subjects in different conditions from EURECOM [19] dataset (first row), VAP RGB-D-T dataset [20] (middle row) and Texas 3D Face Recognition database [10] [9] [8] (bottom row).

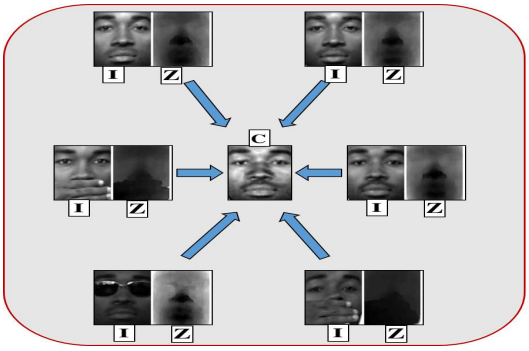

Figure 3. Illustration of canonical representation obtained from the RGB and depth data. Figure shows the canonical representation of a face from the EURECOM data [19] in the middle and the different data having the same canonical representation around it.

Table 1. Rank 1 recognition performance (%) of different algorithms on the EURECOM Dataset [19] for cross modal scenario.

Approach	RGB-Depth	Depth-RGB
CCA [12] [11]	16.40	16.48
Cluster CCA [22]	25.39	24.72
GMA [23]	27.51	31.31
SCDL [28]	31.87	29.67
CDL [13]	32.27	30.22
S^2CDL	**40.21**	**39.01**

Table 2. Rank 1 recognition performance (%) of different algorithms on the EURECOM Face Dataset [19] for the task of RGB-RGB and depth-depth matching. Here, both RGB and depth information are available during training and only RGB images (or depth images) are available during testing, i.e. it is the scenario of matching with hidden depth (RGB) information.

Scenario	**RGB-RGB**								**Depth-Depth**							
Gallery Condition	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Avg.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Avg.**
NN	65.4	73.6	47.3	52.8	43.4	56.6	65.9	57.9	53.9	58.8	46.7	28.0	34.6	47.3	56.6	45.6
LMNN	68.1	76.9	52.2	53.3	47.3	57.1	67.6	60.4	54.4	59.3	48.4	30.2	35.2	49.5	59.9	48.1
LSML	91.8	91.2	75.8	71.4	59.9	81.9	90.7	80.4	69.8	74.7	61.0	38.5	40.1	66.5	73.6	60.6
SVM	77.5	81.9	61.0	58.8	54.4	67.6	71.4	67.5	62.1	63.7	47.3	37.4	27.5	56.6	58.2	50.4
SVM+	78.0	82.4	61.5	59.3	55.0	68.1	72.0	68.1	64.8	64.3	47.8	38.5	26.9	57.7	61.0	51.6
ITML+	92.3	94.0	75.3	72.5	59.9	80.2	91.8	80.9	68.1	75.8	61.5	42.3	41.8	64.3	74.2	61.2
$\mathbf{S^2CDL}$	**93.4**	**96.2**	**76.9**	**74.7**	**60.4**	**84.1**	**92.9**	**82.7**	**70.9**	**77.3**	**65.9**	**43.4**	**42.3**	**67.0**	**75.3**	**63.2**

ing approaches, namely Semi-Coupled Dictionary Learning (SCDL) [28] and Coupled Dictionary Learning (CDL) [13] which have shown their usefulness for both synthesis applications as well as classification tasks. For fairness, SCDL [28] and CDL [13] are used in conjunction with LSML [14] while reporting the results. We observe that the proposed approach performs favorably compared to the state-of-the-art cross-modal techniques.

Now we evaluate the performance of the proposed algorithm for the task of matching with hidden information. The training consists of both RGB and depth data. During testing, either the RGB images or the depth data are available. The results for RGB-RGB matching (with depth as the hidden information) and depth-depth matching (with RGB image as the hidden information) are reported in Table 2. For this scenario, we compare the proposed approach with other approaches for matching data in the same domain, i.e. (1) Nearest neighbor: where the features are directly matched; (2) Classifier based approaches like SVM [26]; (3) Metric learning approaches like LMNN [31] and LSML [14] and (4) SVM+ [26] and ITML+ [32] which are specially designed for matching with hidden information. We observe that for this scenario also, the proposed algorithm performs favorably as compared with the other state-of-the-art approaches.

5.2. Evaluation on VAP RGB-D-T Dataset

The VAP RGB-D-T dataset [20] consists of images (RGB, depth, thermal) of 51 subjects in three different scenarios - rotation, expression and illumination. For our experiments, we have considered the RGB and depth faces corresponding to the illumination setup, which consists of 15300 images (middle row - Figure (2)).

We evaluate our algorithm for the cross-modal scenario by splitting the dataset into training and testing sets consisting of images of 26 and 25 subjects respectively. Experimental results are given in Table 3 where comparison with other methods shows the effectiveness of the proposed approach. For matching with hidden information, we consider images from 10 subjects to constitute the training set. One image from each of the remaining 41 subjects are taken as the gallery and the rest of the images as probe. We observe from the results in Table 4 that the proposed approach performs favourably as compared with the state-of-the-art approaches.

Table 3. Rank 1 recognition performance (%) of different algorithms on the VAP RGB-D-T Dataset [20] for cross-modal scenario.

Approach	**RGB-Depth**	**Depth-RGB**
CCA [12] [11]	15.89	19.86
Cluster CCA [22]	26.23	23.23
GMA [23]	41.69	32.16
SCDL [28]	39.21	36.64
CDL [13]	40.45	34.36
$\mathbf{S^2CDL}$	**47.13**	**37.69**

Table 4. Rank 1 recognition performance (%) of different algorithms on the VAP RGB-D-T Dataset [20] for hidden information scenario.

Approach	**RGB-RGB**	**Depth-Depth**
NN	58.37	55.53
LMNN [31]	82.93	59.60
LSML [14]	93.43	68.56
SVM [26]	65.96	56.46
SVM+ [26]	65.99	56.49
ITML+ [32]	94.10	68.58
$\mathbf{S^2CDL}$	**96.08**	**71.60**

5.3. Evaluation on Texas 3D Face Recognition Dataset

The Texas 3D Face Recognition dataset [10] [9] [8] consists of 3D facial images of 118 subjects with variable number of samples from 1 to 89 (giving a total of 1149 image pairs) captured using a stereo imaging system. The faces are in neutral pose, emotionless and without hats and eyeglasses (bottom row - Figure (2)). The images are registered and the location of the fiducial points are also provided. For

Table 5. Rank 1 recognition performance (%) of different algorithms on the Texas 3D Face Recognition database [10] [9] [8] for cross-modal scenario RGB-Depth (R-D) and Depth-RGB (D-R) using both SIFT [17] and LBP [21] features.

	SIFT		LBP	
Approach	**R-D**	**D-R**	**R-D**	**D-R**
CCA [12] [11]	46.08	37.48	86.86	82.51
Cluster CCA [22]	52.27	50.05	87.25	86.37
GMA [23]	65.60	62.41	94.01	95.46
SCDL [28]	53.72	60.78	88.98	91.59
CDL [13]	59.90	56.32	92.56	93.14
S^2CDL	**66.67**	**64.73**	**96.23**	**97.10**

Table 6. Rank 1 recognition performance (%) of different algorithms on the Texas 3D Face Recognition database [10] [9] [8] for hidden information scenario RGB-RGB (R-R) and Depth-Depth (D-D) using both SIFT [17] and LBP [21] features.

	SIFT		LBP	
Approach	**R-R**	**D-D**	**R-R**	**D-D**
NN	90.61	76.48	83.21	82.54
LMNN [31]	92.98	81.73	87.91	89.59
LSML [14]	94.21	92.16	95.06	95.36
SVM [26]	91.05	78.59	84.48	84.03
SVM+ [26]	91.16	78.72	84.59	84.17
ITML+ [32]	95.85	89.24	92.89	93.59
S^2CDL	**96.50**	**93.27**	**98.10**	**97.73**

our experiments, we have used 14 fiducial points per face image to extract the SIFT [17] features.

We first evaluate our algorithm for the cross-modal scenario of matching RGB to depth images and vice versa. We have used images from 60 subjects for training and 58 subjects for testing. The results are given in Table 5 which shows that our proposed algorithm outperforms other state-of-the-art methods. For matching with hidden information, i.e. matching RGB-RGB or depth-depth with the other as the hidden data, firstly, all the subjects having only a single image in the dataset are removed as they cannot be used in this scenario. Out of the remaining 103 subjects, 30 subjects are taken for training. One image each from the rest of the subjects is taken as gallery and the rest as probe data. The results are shown in Table 6 where comparison with the state-of-the-art methods shows the effectiveness of the proposed approach.

Implementation Details - The scale and orientation for SIFT was taken as 3 and 1 respectively. It was observed that the proposed method gives the best results with the following parameter settings : $\{\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \gamma_1, \gamma_2\} = \{10, 50, 10, 2.5, 2.5, 5, 10\} \times 10^{-4}$ in (9). The optimal size of dictionary based on validation data was found to be 50 for the EURECOM [19] and the same was used for the other datasets also. It was also observed that the algorithm generally converges (reconstruction error difference between consecutive iterations is less than some threshold) after around 30 iterations. Experiments for the RGB-RGB scenario on the EURECOM [19] dataset using separate KSVD initialization in (10) gives an average performance of 82.5% compared to 82.7% (Table 2). So we observe that the initialization does not have any significant effect on the recognition performance. We also did an experiment on the same dataset to test the usefulness of the metric learning part. We observe that the average result improves from 78.0% to 82.7% on using the learned metric $\mathbf{M}$. This shows that though the sparse codes are effective in grouping similar objects, the metric learning increases the discriminatory aspects for even better performance.

Comparison with other features - To test the usefulness of the proposed approach on other features, we repeat the experiments using LBP features [21] for Texas 3D Face Recognition dataset [10] [9] [8], the results of which are included in Tables 5 and 6. We observe that for the hidden information scenario, the results are similar to that of SIFT, but there is significant improvement for the cross-modal scenario. In this case also, **S^2CDL** outperforms the other state-of-the-art approaches.

6. Summary

In this paper, we have proposed a novel framework which can be used for matching with hidden information as well as for cross-modal matching, with applications to RGBD data of Lambertian objects. The algorithm, termed as S^2CDL constructs a canonical representation which is on one hand robust to illumination variations, and also captures the information of both the RGB image and the depth data. The relation between the sparse coefficients of the RGB data and the depth data with that of the canonical representation is learnt simultaneously which is finally used for matching using a Mahalanobis metric. Extensive experiments show the usefulness of S^2CDL for both the scenarios for RGBD data of faces. Though this paper deals with RGBD data of Lambertian objects, i.e., faces, the framework can be extended for other applications like object classification and action recognition with RGBD data, with a proper choice of the canonical representation.

References

[1] Microsoft kinect. http://www.xbox.com/en-us/kinect.

[2] Primesense. http://www.primesense.com/.

[3] M. Aharon, M. Elad, and A. Bruckstein. The K-SVD : An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[4] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi. A set of selected SIFT features for 3D facial ex-

pression recognition. In *International Conference on Pattern Recognition (ICPR), 2010 20th*, pages 4125–4128. IEEE, 2010.

[5] S. Biswas, G. Aggarwal, and R. Chellappa. Robust estimation of albedo for illumination-invariant matching and shape recovery. *IEEE Transactions on Pattern Analysis and Machine Learning*, 31(5):884–899, 2009.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. *International Conference on Machine Learning*, pages 209–216, 2007.

[7] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider. Incorporating privileged information through metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(7):1086–1098, 2013.

[8] S. Gupta, K. R. Castkeman, M. K. Markey, and A. C. Bovik. Texas 3D face recognition database. *http://live.ece.utexas.edu/research/texas3dfr/index.htm*.

[9] S. Gupta, K. R. Castkeman, M. K. Markey, and A. C. Bovik. Texas 3d face recognition database. *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 97–100, 2010.

[10] S. Gupta, M. K. Markey, and A. C. Bovik. Anthropometric 3D face recognition. *International Journal of Computer Vision*, 90(3):331–349, 2010.

[11] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis : An overview with application to learning methods. *Neural Computing*, 16(12):2639–2664, 2004.

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

[13] D. A. Huang and Y. C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. *IEEE International Conference on Computer Vision*, pages 2496–2503, 2013.

[14] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *International Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[15] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems*, pages 801–808s, 2006.

[16] K. Lee, J. Ho, and D. Kriegman. Nine points of light: acquiring subspaces for face recognition under variable lighting. *International Conference on Computer Vision and Pattern Recognition*, 1:519–526, 2001.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International journal of computer vision*, volume 60, pages 91–110. Springer, 2004.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. *International Conference on Machine Learning*, pages 689–696, 2009.

[19] R. Min, N. Kose, and J. L. Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 44(11):1534–1548, 2014.

[20] O. Nikisins, K. Nasrollahi, M. Greitans, and T. B. Moeslund. RGB-DT based face recognition. *IEEE International Conference on Pattern Recognition*, pages 1716–1721, 2014.

[21] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, pages 971–987. IEEE, 2002.

[22] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal. Cluster canonical correlation analysis. *International Conference on Artificial Intelligence and Statistics*, 2014.

[23] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. *International Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[24] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. *IEEE International Conference on Computer Vision*, pages 825–832, 2013.

[25] V. Vapnik and R. Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

[26] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009.

[27] S. Wang, D. Tao, and J. Yang. Relative attribute SVM+ learning for age estimation. *IEEE Transactions on Cybernetics*, 16(3):827–839, 2015.

[28] S. Wang, D. Zhang, L. Yan, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. *International Conference on Computer Vision and Pattern Recognition*, pages 593–600, 2012.

[29] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB Journal*, pages 1–23, 2015.

[30] Z. Wang and Q. Ji. Classifier learning using hidden information. *International Conference on Computer Vision and Pattern Recognition*, pages 4969–4977, 2015.

[31] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, pages 1473–1480, 2005.

[32] X. Xu, W. Li, and D. Xu. Distance metric learning using privileged information for face verification and person reidentification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(99):3150–3162, 2015.

[33] W. Zhang, H. Ji, G. Liao, and Y. Zhang. A novel extreme learning machine using privileged information. *Neurocomputing*, 168(9):823–828, 2015.