Hierarchical Multimodal Metric Learning for Multimodal Classification

Heng Zhang¹, Vishal M. Patel² and Rama Chellappa¹
¹Center for Automation Research, University of Maryland, College Park, MD 20742
²Department of Electrical and Computer Engineering, Rutgers University, NJ, 08854
hzhang98@umiacs.umd.edu, vishal.m.patel@rutgers.edu, rama@umiacs.umd.edu

Abstract

Multimodal classification arises in many computer vision tasks such as object classification and image retrieval. The idea is to utilize multiple sources (modalities) measuring the same instance to improve the overall performance compared to using a single source (modality). The varying characteristics exhibited by multiple modalities make it necessary to simultaneously learn the corresponding metrics. In this paper, we propose a multiple metrics learning algorithm for multimodal data. Metric of each modality is a product of two matrices: one matrix is modality specific, the other is enforced to be shared by all the modalities. The learned metrics can improve multimodal classification accuracy and experimental results on four datasets show that the proposed algorithm outperforms existing learning algorithms based on multiple metrics as well as other approaches tested on these datasets. Specifically, we report 95.0% object instance recognition accuracy, 89.2% object category recognition accuracy on the multi-view RGB-D dataset and 52.3% scene category recognition accuracy on SUN RGB-D dataset.

1. Introduction

Owing to recent developments in sensor technology, researchers and developers are able to collect multimodal data consisting of depth information and RGB images to achieve better performance for tasks such as object detection, classification and scene understanding [20, 7, 18, 30, 38, 32]. Massive image and video data available on the Internet are associated with tags and metadata which are useful for image classification [16] and retrieval [45, 37]. Solutions to these problems can be formulated using multimodal classification frameworks. Multimodal classification has also been studied for other applications such as audio-visual speech classification [27, 33], and multimodal biometrics recognition [29, 44].

How to efficiently and effectively combine different modalities is the key issue in multimodal classification. Feature vectors corresponding to different modalities might be very different even if they essentially represent the same object. Some feature vectors are very discriminative while others are not; some feature vectors are clean while others are noisy; some feature vectors are dense while others are sparse. Many factors like data acquisition, preprocessing and feature extraction can make feature vectors’ behavior quite different. Therefore, direct linear combination of feature vectors or simple linear combination of the result of each modality can not guarantee good performance compared with using certain modality alone.

Metric learning algorithms can learn the Mahalanobis distance from data pairs and side information indicating the relationship between data pairs [40]. The learned distance can be better than Euclidean distance for the original feature space. Extensive research on metric learning in unimodal setting is available in the literature. Typical examples include the one proposed in [40], Large Margin Nearest Neighbor (LMNN) algorithm [36] and Information Theoretical Metric Learning (ITML) algorithm [12].

Extending the unimodal metric learning algorithm to multi-modal metric learning can be a good solution for multimodal classification problems if the learned metrics are appropriate distance measures for corresponding feature spaces. Also, it is important to explore the relationship among multiple metrics and the learning process should take into account the underlying differences among multiple modalities by balancing the contribution of each modality. As will be analyzed in Section 2 and Section 3, existing approaches for multimodal metric learning do not fully capture the relationships among the multiple learned metrics.

Motivated by previous works that consider shared representations in their formulations for multi-modal applications such as [27, 34, 41, 44], we propose a Hierarchical Multimodal Metric Learning (HM3L) algorithm which fully explores the relationships among the different metrics of different modalities. In our formulation, metric of each modality is constructed through the multiplication of modality specific part representing appropriate subspace and a common part (p.s.d matrix) shared by all the met-
Metric learning has been studied in various fields such as machine learning [40, 36], information retrieval [25], computer vision [15] and biometrics [31, 8]. The goal of a metric learning algorithm is to learn a metric so that after data are projected using the learned metric, similar data samples (e.g. from the same class) are clustered together and dissimilar data samples (e.g. samples from different classes) are separated.

In [40], metric learning problem was formulated as a convex optimization problem by utilizing the side information of two data samples being similar or dissimilar. LMNN [36] applied the idea of large margin in Support Vector Machine (SVM) to improve the KNN classifier and used triplet constraints to describe the relative relationships among three samples. In [12], the information theoretical metric learning (ITML) algorithm was proposed which essentially minimizes the differential relative entropy between two multivariate Gaussians subject to constraints on the distance function.

More recently proposed metric learning algorithms also explore the structure of the metric by enforcing low-rank constraints [11, 24] or sparse constraints [42, 28, 23] or both sparse and low-rank constraints [22]. For high-dimensional problems, [11] showed that enforcing low-rank constraints on the metric during the learning process is computationally efficient and tractable even with a small number of samples. More comprehensive reviews of various metric learning methods and their applications are summarized in [1, 19].

Several multimodal metric learning algorithms have also been proposed in the literature [39, 13, 43, 17]. For instance, a multimodal metric learning method in [39] applied the multi-wing harmonium (MWH) learning framework to get latent representations from different modalities and learned a metric under a probabilistic formulation. A Heterogeneous Multi-Metric Learning algorithm proposed in [43] for multi-sensor fusion essentially extended the LMNN algorithm [36] for multi-metric learning. Similarly, in [17] a large margin multi-metric learning (LM3L) was proposed for face and kinship verification by learning multiple metrics under which the correlations of different feature representations of each sample are maximized. Some of the other multimodal metric learning algorithms include Pairwise-constrained Multiple Metric Learning (PMML) [10]. Note that these methods can be viewed as multimodal extensions of the classical unimodal metric learning algorithms like ITML and LMNN. One of the limitations of these methods is that they do not explore the relationships among different metrics corresponding to different modalities.

3. Formulation

3.1. Problem Description

Let
\[ S = \{(X_i, X_j)|y_{ij} = 1\} \]
and
\[ D = \{(X_i, X_j)|y_{ij} = -1\} \]
be two sets consisting of similar instance pairs and dissimilar instance pairs, respectively. An instance in the multi-modal scenario is denoted as
\[ X_i = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)}\}, \]
which consists of $K$ features from $K$ different modalities, where $x_i^{(1)} \in \mathbb{R}^{l_1}$, $x_i^{(2)} \in \mathbb{R}^{l_2}$, $\ldots$, $x_i^{(K)} \in \mathbb{R}^{l_K}$. Note that the dimension of each feature vector can be different. In multimodal metric learning, the objective is to learn metrics for such instances consisting of $K$ feature vectors.

A simple way to learn a metric for multimodal data is by concatenating the features of the $K$ modalities into one
feature vector of length $\sum_{i=1}^{K} l_i$ and applying classical metric learning algorithms like LMNN or ITML. The drawback of this approach is the high computational cost incurred by learning an $\sum_{i=1}^{K} l_i$ by $\sum_{i=1}^{K} l_i$ metric. This problem is even more serious for high-dimensional multimodal data.

Existing multimodal metric learning algorithms such as Pairwise-constrained Multiple Metric Learning [16], Large Margin Multi-metric Learning [17], and Heterogeneous Multi-Metric Learning [43], are extensions of the classical unimodal metric learning algorithms in which the distance between any two instances is obtained as

$$d_{M_k}^2(X_i, X_j) = \frac{1}{K} \sum_{k=1}^{K} d_{M_k}^2(x_i^{(k)}, x_j^{(k)})$$

These approaches simultaneously solve $K$ positive semi-definite (p.s.d) matrices $M_k$, $k = 1, \cdots, K$, as metrics in a joint formulation.

3.2. Hierarchical Multimodal Metric Learning (HM3L) Formulation

In order to efficiently learn multiple metrics for multiple modalities as well as to capture the relationship among them, we enforce the different metrics $M_k, k = 1, \cdots, K$ to satisfy the following condition

$$M_k = P_k^T M P_k, \quad k = 1, \cdots, K,$$

where $P_k \in \mathbb{R}^{d \times l_k}$ and $d \leq \min \{l_1, l_2, \cdots, l_K \}$. Also, $M$ is required to be a p.s.d matrix. Using this formulation, one can easily show that if $M \in \mathbb{R}^{d \times d}$ is p.s.d and rank$(M) \leq r$ ($r \leq d$), then for any non-trivial $P_k \in \mathbb{R}^{d \times l_k}$, $M_k = P_k^T M P_k$ is p.s.d and rank$(M_k) \leq r$.

For the given training data, the learned metrics $M_k$ are obtained by learning the modality specific part $P_k$ and the shared part $M$ in a hierarchical framework. As long as $M$ is p.s.d, $M_k$ is p.s.d meaning that $M_k$ are valid metrics.

By enforcing (2), we establish the relationship among the different modalities. As a result, we can formulate the Hierarchical multimodal metric learning (HM3L) algorithm as the optimization problem specified in (3).

$$\min_{M \in S_d^+} tr(M) + \gamma \sum_{k=1}^{K} \|P_k\|_F^2,$$

s.t. $\frac{1}{K} \sum_{k=1}^{K} d_M^2(P_k x_i^{(k)}, P_k x_j^{(k)}) \leq \mu$ if $y_{ij} = 1$

$$\frac{1}{K} \sum_{k=1}^{K} d_M^2(P_k x_i^{(k)}, P_k x_j^{(k)}) \geq \beta$$ if $y_{ij} = -1.$

Here $\gamma$ controls the relative contribution to the cost function between $P_k$ and $M$ and $\mu$ and $\beta$ are non-negative real numbers which specify the upper bound for distance of two similar instances and lower bound for distance of two dissimilar instances, respectively. We introduce the slack variables $\epsilon_{ij} > 0$ for constraints. Then (3) can be rewritten as

$$\min_{M \in S_d^+} tr(M) + \gamma \sum_{k=1}^{K} \|P_k\|_F^2,$$

s.t. $\frac{1}{K} \sum_{k=1}^{K} d_M^2(P_k x_i^{(k)}, P_k x_j^{(k)}) \leq \mu + \epsilon_{ij}$ if $y_{ij} = 1$

$$\frac{1}{K} \sum_{k=1}^{K} d_M^2(P_k x_i^{(k)}, P_k x_j^{(k)}) \geq \beta - \epsilon_{ij}$$ if $y_{ij} = -1.$

3.3. HM3L-based multimodal classification

Once $P_k$ and $M$ are learned, we can easily get $L$ such that $L^T L = M$ through matrix decomposition. Then the multi-modal data

$$X_i = \{x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(K)}\}$$

can be projected by $P_k$ and $L$ and transformed to

$$\tilde{X}_i = \{LP_1 x_i^{(1)}, LP_2 x_i^{(2)}, \cdots, LP_K x_i^{(K)}\}.$$
M and P by updating one with the other fixed.

The hinge-loss function indicates that only pairs of samples that violate the distance constraints will make contributions to the overall cost function. For notational convenience, let \( A_{S,P}^t, A_{D,P}^t, A_{S,M}^t \) and \( A_{D,M}^t \) denote active sets at time \( t \). \( A_{S,P}^t \) (\( A_{D,P}^t \)) means set for similar (dissimilar) pairs that violate the distance constraint when we fix \( P \) to update \( M \). Similarly, \( A_{S,M}^t \) (\( A_{D,M}^t \)) means set for similar (dissimilar) pairs that violate the distance constraint when we fix \( M \) to update \( P \).

\[
A_{S,P}^t = \{(X_i, X_j) \in S | \frac{1}{K} \sum_{k=1}^{K} d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)}) \geq \mu \}
\]

\[
A_{D,P}^t = \{(X_i, X_j) \in D | \frac{1}{K} \sum_{k=1}^{K} d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)}) \leq \beta \}
\]

\[
A_{S,M}^t = \{(X_i, X_j) \in S | \frac{1}{K} \sum_{k=1}^{K} d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)}) \geq \mu \}
\]

\[
A_{D,M}^t = \{(X_i, X_j) \in D | \frac{1}{K} \sum_{k=1}^{K} d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)}) \leq \beta \}
\]

### 4.1. Updating M

Fixing \( P \), projected sub-gradient method [6] can be applied to solve for \( M \). It involves two key steps.

**Step 1:**

\[
M_{t+1} = M_t - \eta g_t(M),
\]

where \( g_t(M) \) is the gradient of \( L(M) \) at time \( t \) and it is derived as,

\[
g_t(M) = \frac{\partial L(M)}{\partial M} = \alpha \sum_{(X_i, X_j) \in A_{S,P}^t} \left[ \frac{1}{K} \sum_{k=1}^{K} \text{sgn}(\mu - d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)})) \right] +
\]

\[
C(1-\alpha) \sum_{(X_i, X_j) \in A_{D,P}^t} \left[ -\frac{1}{K} \sum_{k=1}^{K} \text{sgn}(\beta - d_{M_{t-1}}^2(P_{k,t-1}x_i^{(k)}, P_{k,t-1}x_j^{(k)})) \right]
\]

Where \( \text{sgn}(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0 \end{cases} \) is a rank 1 matrix.

**Step 2:**

\[
M_{t+1} = V^T \Sigma + V,
\]

where \( V^T \Sigma V \) is the eigenvalue decomposition of \( M_{t+1} \). Projecting \( M_{t+1} \) onto the \( p.s.d \) cone can be done by thresholding the eigenvalues by keeping the positive eigenvalues and setting the negative ones to be 0.

### 4.2. Updating P

Fixing \( M \), each \( P \) can be updated separately through gradient descent as

\[
P_{k,t} = P_{k,t-1} - \eta g_t(P_k), \quad k = 1, 2, \ldots, K,
\]

where \( g_t(P_k) \) is the gradient of \( L(P_k) \) at time \( t \) and it is derived as

\[
g_t(P_k) = 2\gamma P_{k,t} + C \alpha \sum_{(X_i, X_j) \in A_{S,M}^t} \left[ \frac{2}{K} M_{t-1} P_{k,t-1} B_{i,j}^{(k)} \right] +
\]

\[
C(1-\alpha) \sum_{(X_i, X_j) \in A_{D,M}^t} \left[ -\frac{2}{K} M_{t} P_{k,t-1} B_{i,j}^{(k)} \right]
\]

The overall Hierarchical Multimodal Metric Learning (HM3L) algorithm is summarized in Algorithm 1.

### Algorithm 1: Hierarchical Multimodal Metric Learning (HM3L)

**Inputs:**

\[ S = \{(X_i, X_j)|y_{ij} = 1\}, \]

\[ D = \{(X_i, X_j)|y_{ij} = -1\}, \] positive integer \( \gamma, \alpha, \eta, \mu, \beta, C \) and maximum iteration \( T \).

**Initialization:**

To initialize \( P_k \) (\( k = 1, 2, \ldots, K \)):

- construct \( X^k \in \mathbb{R}^{k \times N} \) of \( x_i^{(k)} \) from \( S \) and \( D \);
- perform PCA on \( X^k \) to obtain \( P_{k,0} \in \mathbb{R}^{k \times k} \).

To initialize \( M \):

- set \( M_0 = I_{d \times d} \).

**Main loop:**

for \( t = 1 : T \) do

- calculate \( A_{S,P}^t \) and \( A_{D,P}^t \) to update \( M \) through (7), (6) and (8);
- calculate \( A_{S,M}^t \) and \( A_{D,M}^t \) to update \( P \) through (10) and (9),

end

**Outputs:**

\( P_k \) \((k = 1, 2, \ldots, K)\) and \( M \).

### 5. Experiments

To illustrate the effectiveness of our method, we present experimental results on four publicly available multimodal datasets: NUS-WIDE dataset [9], RGB-D Object dataset [20], CIN 2D3D object dataset [7] and SUN RGB-D dataset [32]. The details of these datasets, experimental setups and experimental results are given in the following subsections.

For experiments on each dataset, we include (1) the baseline result (without metric learning) obtained by certain features plus either NN or SVM classifiers depending on which was used to report the baseline result, (2) results from the proposed HM3L method as well as other publicly available multiple metrics learning methods [10, 43] by first transforming the features used in the baseline result, then applying either NN or SVM classifier, (3) results from other methods which reported the best results on that experiment.

#### 5.1. Tagged image classification on NUS-WIDE dataset

The NUS-WIDE dataset [9] consists of 269,648 web images and tags from Flickr. For a fair comparison with previous results
Table 2 shows the KNN classification accuracies of different methods. As can be seen from the table, the HM3L method performs the best and it outperforms all the other methods. This experiment clearly shows that our method can provide better distance measures which can enhance the performance of a classification algorithm.

To show whether the proposed algorithm converges, we empirically show the convergence of our algorithm by plotting the normalized cost function values versus iterations. From Figure 2, we can observe that the proposed algorithm converges in a few iterations.

5.2. Object recognition on RGB-D Object dataset

RGB-D Object dataset [20] is a large scale multi-view dataset for 3D object recognition, segmentation, scene labeling and so on. It consists of video recordings of 300 everyday objects organized into 51 different categories. The video recordings were captured by cameras mounted at 3 different elevation angles of 30°, 45° and 60°. A single RGB-D frame consists of both an RGB image and a depth image. Evaluation protocols for various computer vision tasks such as instance recognition and category recognition were set in [20]. RGB-D Images were sampled every 5th frame of the videos and in total about 45,000 RGB-D images were collected.

Kernel descriptors [3] [4] were extracted as features for RGB images and depth image. For RGB images, the LBP kernel descriptor, Gradient kernel descriptor and normalized color kernel descriptor were extracted. For depth images, the gradient kernel descriptor and the LBP kernel descriptor were extracted from depth images; normal kernel descriptor and size kernel descriptor were extracted from point clouds which were converted from the depth images. For each kernel descriptor, object-level features were obtained from 1000 dimensional basis vector for (1 + 4 + 9) × 1000 = 14000; principal component analysis was used to reduce the the dimensionality to 1000. After feature extraction, each RGB-D image was represented by seven kernel descriptors and each kernel descriptor by a 1000 dimensional vector.

5.2.1 Experimental Setup

For the instance recognition experiment, images corresponding to videos captured at angles 30° and 60° were used for training, and images corresponding to videos captured at angle 45° were used for testing. For the category recognition experiment, one object was randomly chosen and left out from each category for testing and all views of the remaining objects were used for training. Ten trials were repeated for category recognition.

For instance and category recognition tasks, we first learned multiple metrics for Seven kernel descriptors using the similar and dissimilar set of the RGB-D images generated from the training data. We then performed linear SVM classification [14] based on the learned metrics. We also compared the performance of our method with the results reported in [34] which are based on deep learning-based methods for RGB-D image classification.

5.2.2 Experimental Results

Classification results for instance recognition and category recognition are shown in Table 2 and Table 3 respectively. From these tables, we make the following observations. (1) the proposed HM3L-based classification method outperforms the best results obtained from MMSS [34] which applies deep architectures on...
the RGB-D images for both instance recognition testing on over 13800 instances and category recognition overall ten trials. (2) The proposed HM3L algorithm can boost classification accuracy compared to the case where metrics learning was not performed. (3) HM3L-based multimodal classification outperforms other multiple metrics learning-based classification and this shows that the idea of capturing the relationship for different multiple metrics can help to learn more appropriate distance measures.

Confusion matrices of classification results based on the proposed algorithm are shown in Figure 3 for instance recognition experiment and in Figure 4 for the 8th trial of category recognition experiment. The testing data of recognition experiment are placed such that testing samples of the same objects are put together and objects from the same category are grouped together. As we can see from Figure 3, for each of 300 objects, most samples are classified correctly (diagonal) and many errors are made due to the misclassification of certain samples to other objects from the same category. Examples of misclassification in category recognition is shown in Figure 5. For each column, the objects on top was misclassified to the category represented by certain object in the bottom. We can see that errors occur due to similar color and shape.

### 5.3. Object recognition on CIN 2D3D dataset

CIN 2D3D object classification dataset [7] contains segmented color and depth images of 154 objects from eighteen categories of common household and office objects. Each category contains between three to fourteen objects. Each object was recorded using a high-resolution color camera and a time-of-flight rang sensor. Objects were rotated using a turn table and snapshots taken every ten degrees and yields 36 views per object. Each view is one data

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.8995</td>
<td>0.8995</td>
<td>0.8995</td>
<td>0.9286</td>
<td>0.8056</td>
<td>0.9352</td>
<td>0.9233</td>
<td>0.9140</td>
<td>0.9224</td>
</tr>
<tr>
<td>3-NN</td>
<td>0.8108</td>
<td>0.6653</td>
<td>0.8849</td>
<td>0.8929</td>
<td>0.6944</td>
<td>0.9021</td>
<td>0.9220</td>
<td>0.9246</td>
<td>0.9431</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.6971</td>
<td>0.4868</td>
<td>0.8426</td>
<td>0.8519</td>
<td>0.5860</td>
<td>0.8849</td>
<td>0.9299</td>
<td>0.9114</td>
<td>0.9418</td>
</tr>
<tr>
<td>10-NN</td>
<td>0.4775</td>
<td>0.2394</td>
<td>0.7646</td>
<td>0.7394</td>
<td>0.4405</td>
<td>0.8333</td>
<td>0.9139</td>
<td>0.9008</td>
<td>0.9339</td>
</tr>
<tr>
<td>20-NN</td>
<td>0.1348</td>
<td>0.0450</td>
<td>0.6230</td>
<td>0.4841</td>
<td>0.1746</td>
<td>0.7130</td>
<td>0.9074</td>
<td>0.8876</td>
<td>0.9223</td>
</tr>
</tbody>
</table>

Table 1. KNN Classification Accuracy under learned metrics for tagged images.

<table>
<thead>
<tr>
<th>Methods</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lai [20]</td>
<td>90.0</td>
<td>54.7</td>
<td>91.2</td>
</tr>
<tr>
<td>Bo [4]</td>
<td>82.9</td>
<td>-</td>
<td>90.4</td>
</tr>
<tr>
<td>Blum [4]</td>
<td>92.1</td>
<td>51.7</td>
<td>92.8</td>
</tr>
<tr>
<td>HM3L</td>
<td>93.4</td>
<td>55.6</td>
<td>96.0</td>
</tr>
</tbody>
</table>

Table 2. Instance recognition accuracy on RGB-D Object dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lai 20</td>
<td>63.7±2.2</td>
<td>74.5±2.1</td>
<td>83.8±3.5</td>
</tr>
<tr>
<td>Bo 4</td>
<td>80.2±2.1</td>
<td>80.3±2.9</td>
<td>86.5±2.1</td>
</tr>
<tr>
<td>Blum 4</td>
<td>82.9±3.1</td>
<td>-</td>
<td>86.4±2.3</td>
</tr>
<tr>
<td>HM3L</td>
<td>82.4±3.1</td>
<td>81.2±2.8</td>
<td>87.5±2.9</td>
</tr>
<tr>
<td>MMSS [10]</td>
<td>79.2±2.4</td>
<td>83.5±2.6</td>
<td>88.5±2.2</td>
</tr>
<tr>
<td>PMML [10] + linear SVM</td>
<td>75.8±3.2</td>
<td>77.4±2.4</td>
<td>87.3±1.8</td>
</tr>
<tr>
<td>HMML [43] + linear SVM</td>
<td>81.0±2.1</td>
<td>81.9±2.4</td>
<td>89.2±1.6</td>
</tr>
</tbody>
</table>

Table 3. Category recognition accuracy on RGB-D Object dataset.
sample consisting of RGB image and Depth image. Following the procedures used to extract kernel descriptors for samples in RGB-D object dataset, we also extract kernel descriptors for data samples in 2D3D dataset.

5.3.1 Experiment Results
The evaluation protocol for category classification was set in the original paper [7]. Six objects per category were used for training and remaining objects were used for testing. For each object, eighteen views were selected for training and eighteen for testing. The training set consists of 82 objects with a total of 1476 views. The test set consists of 74 objects with 1332 views. Same methods as included in the RGB-D dataset were evaluated. Classification results for category recognition are shown in Table 4. As can be seen from this table, the proposed HM3L-based multimodal classification gives the best performance on average.

<table>
<thead>
<tr>
<th>Methods</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browatzki [7]</td>
<td>66.6</td>
<td>74.6</td>
<td>82.8</td>
</tr>
<tr>
<td>HMIF [32]</td>
<td>86.3</td>
<td>87.6</td>
<td>91.0</td>
</tr>
<tr>
<td>MMSS [32]</td>
<td></td>
<td></td>
<td>91.3</td>
</tr>
<tr>
<td>PMML [32] + linear SVM</td>
<td>90.6</td>
<td>82.7</td>
<td>91.8</td>
</tr>
<tr>
<td>HML + linear SVM</td>
<td>86.8</td>
<td>83.4</td>
<td>90.8</td>
</tr>
</tbody>
</table>

Table 4. Category recognition accuracy (in %) on CIN 2D3D dataset.

5.4. Scene Categorization on SUN RGB-D dataset
The SUN RGB-D dataset [32] consists of 10355 RGB-D scene images including 3784 Kinect v2 images, 1159 Intel RealSense images as well as 1449 images taken from the NYU Depth Dataset V2 [30], 554 scene images from the Berkeley B3DO Dataset [18], and 3389 Asus Xtion images from SUN3D videos [38]. We choose the same Places-CNN [46] scene features of dimension 4096 for both RGB image and depth image which were used to report the baseline results in [32].

5.4.1 Experimental Results
We followed the standard experimental setup for scene categorization task according to [32]. Specifically, nineteen scene categories with more than eighty images were used. These scene categories are bathroom, bedroom, classroom, computer room, conference room, corridor, dining area, dining room, discussion area, furniture store, home office, kitchen, lab, lecture theatre, library, living room, office, rest space, study space.

The train and test split is available in [32]. In total, 4845 samples were used for training and 4659 samples were used for testing. The standard average categorization accuracy was used for evaluation. We applied the proposed HM3L method to the Places-CNN features, transformed the original features with the learned matrices, and then applied one-vs-all rbf SVM for classification. The scene category recognition results are shown in Table 5.

<table>
<thead>
<tr>
<th>Methods</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place-CNN + linear SVM</td>
<td>35.6</td>
<td>25.5</td>
<td>37.2</td>
</tr>
<tr>
<td>Place-CNN + rbf SVM</td>
<td>38.1</td>
<td>27.7</td>
<td>39.0</td>
</tr>
<tr>
<td>Liao [32]</td>
<td>36.1</td>
<td>-</td>
<td>41.3</td>
</tr>
<tr>
<td>Zhu [47]</td>
<td>-</td>
<td>-</td>
<td>41.5</td>
</tr>
<tr>
<td>Wang [35]</td>
<td>-</td>
<td>-</td>
<td>48.1</td>
</tr>
<tr>
<td>PMML [32] + rbf SVM</td>
<td>40.7</td>
<td>30.5</td>
<td>44.2</td>
</tr>
<tr>
<td>HML + rbf SVM</td>
<td>47.9</td>
<td>32.6</td>
<td>51.7</td>
</tr>
<tr>
<td>HML + rbf SVM</td>
<td>48.6</td>
<td>33.2</td>
<td>52.3</td>
</tr>
</tbody>
</table>

Table 5. Scene categorization accuracy (in %) on SUN RGB-D dataset.

6. Conclusions
In this paper, we proposed a hierarchical multimodal metric learning algorithm which can efficiently learn multiple metrics for multi-modal data while fully exploiting the relationships among these metrics. The proposed approach makes no assumption about the feature type or applications. We view feature learning as a different problem and only focus on learning discriminative metrics for multimodal data in order to improve the multimodal classification accuracy. As we separate the feature learning process from the metric learning process, the proposed approach can be applied to many different applications with many different feature types. Experimental results on four datasets show that the proposed metric learning algorithm outperforms other metric learning algorithms dealing with multi-modal data and provide the best performance for all the experiments considered. As the concept of modality is quite general and many computer vision problems can be considered in multi-modal settings, the proposed HM3L algorithm can be applied where appropriate metrics are required and can boost the performance of related computer vision tasks.
Acknowledgment

This work was supported by DARPA Active Authentication Project under cooperative agreement FA8750-13-2-0279 and by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

References


