Hierarchical Multimodal Metric Learning for Multimodal Classification

Heng Zhang[1], Vishal M. Patel[2] and Rama Chellappa[1]
[1]Center for Automation Research, University of Maryland, College Park, MD 20742
[2]Department 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 distance metrics. In this paper, we propose a multiple metrics learning algorithm for multimodal data. Metric of each modality is 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 on 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 of data pairs [40]. The learned distance metric can be better than Euclidean distance for the original feature space. Extensive research on metric learning in uni-modal setting is available in the literature. Classical algorithms includes the algorithm proposed in [40], Large Margin Nearest Neighbor (LMNN) algorithm [36] and Information Theoretical Metric Learning (ITML) algorithm [12].

Extending the uni-modal 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 the multiple metrics and the learning process can 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-

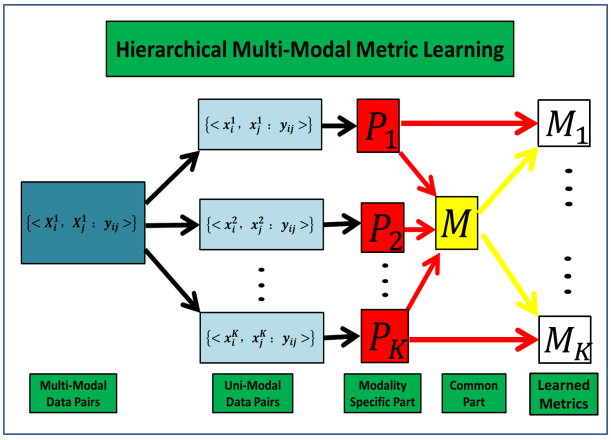

Figure 1. Overview of Hierarchical MultiModal Metric Learning.

rics. Figure 1 gives an overview of the proposed multimodal metric learning algorithm. Given multimodal representations, first we apply modality-specific projections $\mathbf{P}_k$ to each modalities since their representations are very different in nature, then we apply the common metric $\mathbf{M}$ to features after the modality-specific projection assuming the features lie in the same common space.

The rest of this paper is organized as follows. In Section 2, we review different metric learning algorithms. In Section 3, the Hierarchical Multimodal Metric Learning (HM3L) algorithm is proposed and compared with related multiple metrics learning algorithms. In Section 4, an efficient algorithm based on subgradient method is applied to solve the resulting optimization problem. Extensive experimental results on four datasets are presented in Section 5. Finally, Section 6 concludes the paper with a brief summary.

2. Related Work

Metric learning has been studied in various fields from machine learning [40, 36], information retrieval [25] to computer vision [15] and biometrics [31, 8]. The goal of a metric learning algorithm is to learn a distance 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 a recent work, [40] formulated the metric learning problem as a convex optimization problem by utilizing the side information of two data samples being similar or dissimilar. LMNN [36] applies the idea of large margin in Support Vector Machine (SVM) to improve the KNN classification and uses triplet constraints to describe the relative relationship of three samples. In [12], the information theoretical metric learning (ITML) algorithm was proposed which essentially minimizes the differential relative entropy between two multivariate Gaussians under constraints on the distance function.

More recent metric learning algorithms also explore the structure of the metric by enforcing the 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 small number of samples. More comprehensive survey 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] applies multi-wing harmonium (MWH) learning framework to get latent representations from different modalities and learns distance metric under a probabilistic formulation. A Heterogeneous Multi-Metric Learning algorithm proposed in [43] for multi-sensor fusion essentially extends 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 which learns multiple distance 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 multimodal scenario is denoted as

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

which consists of K features from K different modalities, where $x_i^{(1)} \in \mathbb{R}^{l_1}, x_i^{(2)} \in \mathbb{R}^{l_2}, \cdots, x_i^{(K)} \in \mathbb{R}^{l_K}$. Note that the dimension of each feature vector can be different. In multimodal metric learning, the objective is to learn distance metrics for such instances consisting of K feature vectors.

A simple way to learn distance 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 the 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 $\sum_{i=1}^{K} l_i$ distance metric. This problem is even more serious for high-dimensional multimodal data.

Existing multimodal metric learning algorithms such as Pairwise-constrained Multiple Metric Learning [10], 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

$$\begin{aligned} d_m^2(X_i, X_j) &= \frac{1}{K}\sum_{i=1}^{K} d_{M_k}^2(x_i^{(k)}, x_j^{(k)}) \\ &= \frac{1}{K}\sum_{i=1}^{K}(x_i^{(k)} - x_j^{(k)})^T \mathbf{M}_k (x_i^{(k)} - x_j^{(k)}). \end{aligned} \tag{1}$$

These approaches simultaneously solve K positive semi-definite (*p.s.d*) matrices $\mathbf{M}_k, k = 1, \cdots, K$ as distance 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 $\mathbf{M}_k, k = 1, \cdots, K$ to satisfy the following condition

$$\mathbf{M}_k = \mathbf{P}_k^T \mathbf{M} \mathbf{P}_k, \quad k = 1, \cdots, K, \tag{2}$$

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

For the given training data, the learned metrics $\mathbf{M}_k$ are obtained by learning modality specific part $\mathbf{P}_k$ and the shared part $\mathbf{M}$ in a hierarchical framework. With the above proposition, as long as $\mathbf{M}$ is *p.s.d*, $\mathbf{M}_k$ is *p.s.d* meaning that $\mathbf{M}_k$ are valid distance metrics.

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

$$\begin{aligned} \min_{\mathbf{M} \in S_d^+} \quad & tr(\mathbf{M}) + \gamma \sum_{k=1}^{K} \|\mathbf{P}_i\|_F^2 \\ s.t. \ & \frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) \leq \mu \quad \text{if } y_{ij} = 1 \\ & \frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) \geq \beta \quad \text{if } y_{ij} = -1. \end{aligned} \tag{3}$$

Here γ controls the relative contribution to the cost function between $\mathbf{P}_k$ and $\mathbf{M}$ and μ and β 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

$$\begin{aligned} \min_{\mathbf{M} \in S_d^+} \quad & tr(\mathbf{M}) + \gamma \sum_{k=1}^{K} \|\mathbf{P}_i\|_F^2 \\ s.t. \ & \frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) \leq \mu + \epsilon_{ij} \quad \text{if } y_{ij} = 1 \\ & \frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) \geq \beta - \epsilon_{ij} \quad \text{if } y_{ij} = -1. \end{aligned} \tag{4}$$

3.3. HM3L-based multimodal classification

Once $\mathbf{P}_k$ and $\mathbf{M}$ are learned, we can easily get $\mathbf{L}$ such that $\mathbf{L}^T\mathbf{L} = \mathbf{M}$ through matrix decomposition. Then the multi-modal data

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

can be projected by $\mathbf{P}_k$ and $\mathbf{L}$ and transformed to

$$\hat{X}_i = \{\mathbf{L}\mathbf{P}_1 x_i^{(1)}, \mathbf{L}\mathbf{P}_2 x_i^{(2)}, \cdots, \mathbf{L}\mathbf{P}_K x_i^{(K)}\}.$$

Concatenation of all the projected features can be used with various classification algorithms like KNN and SVM.

4. Optimization

To solve the proposed optimization problem (4), we apply hinge-loss function to get rid of the constraints which results in an unconstrained optimization problem as follows

$$\begin{aligned} & \min_{\mathbf{M} \in S_d^+} tr(\mathbf{M}) + \gamma \sum_{k=1}^{K} \|\mathbf{P}_i\|_F^2 \\ & + \alpha C \sum_{(X_i, X_j) \in S} \left[\frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) - \mu \right]_+ \\ & + (1-\alpha) C \sum_{X_i, X_j \in D} \left[\beta - \frac{1}{K}\sum_{k=1}^{K} d_M^2(\mathbf{P}_k x_i^{(k)}, \mathbf{P}_k x_j^{(k)}) \right]_+ \end{aligned} \tag{5}$$

where C is a positive number that controls the relative contribution between the constraints on the metric and the constraints on the data samples, α is a constant that balances the relative contribution between the pairs from similar set and pairs from dissimilar set. Let $L(\mathbf{M}; \mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_K)$ denote the above cost function we are trying to minimize. It is a bi-convex optimization problem when we consider $\mathbf{P}_k$ $(k = 1, 2, ..., K)$ together as $\mathbf{P}$. We iteratively solve $\mathbf{M}$ and $\mathbf{P}$ 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 $A^t_{S,P}$, $A^t_{D,P}$, $A^t_{S,M}$ and $A^t_{D,M}$ denote active sets at time t. $A^t_{S,P}$ ($A^t_{D,P}$) means set for similar (dissimilar) pairs that violate the distance constraint when we fix $\mathbf{P}_k$ to update $\mathbf{M}$. Similarly, $A^t_{S,M}$ ($A^t_{D,M}$) means set for similar (dissimilar) pairs that violate the distance constraint when we fix $\mathbf{M}$ to update $\mathbf{P}_k$.

$$A^t_{S,P} = \{(X_i, X_j) \in S | \frac{1}{K} \sum_{k=1}^{K} d^2_{\mathbf{M}_{t-1}}(\mathbf{P}_{k,t-1} x_i^{(k)}, \mathbf{P}_{k,t-1} x_j^{(k)}) \geq \mu\}$$

$$A^t_{D,P} = \{(X_i, X_j) \in D | \frac{1}{K} \sum_{k=1}^{K} d^2_{\mathbf{M}_{t-1}}(\mathbf{P}_{k,t-1} x_i^{(k)}, \mathbf{P}_{k,t-1} x_j^{(k)}) \leq \beta\}$$

$$A^t_{S,M} = \{(X_i, X_j) \in S | \frac{1}{K} \sum_{k=1}^{K} d^2_{\mathbf{M}_t}(\mathbf{P}_{k,t-1} x_i^{(k)}, \mathbf{P}_{k,t-1} x_j^{(k)}) \geq \mu\}$$

$$A^t_{D,M} = \{(X_i, X_j) \in D | \frac{1}{K} \sum_{k=1}^{K} d^2_{\mathbf{M}_t}(\mathbf{P}_{k,t-1} x_i^{(k)}, \mathbf{P}_{k,t-1} x_j^{(k)}) \leq \beta\}.$$

4.1. Updating $\mathbf{M}$

Fixing $\mathbf{P}_k$, projected sub-gradient method [6] can be applied to solve $\mathbf{M}$. It involves two key steps.

Step 1:

$$\mathbf{M}_{tmp} = \mathbf{M}_t - \eta g_t(\mathbf{M}), \tag{6}$$

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

$$g_t(\mathbf{M}) = \mathbf{I}_{d\times d} + C\alpha \sum_{(X_i,X_j)\in A^t_{S,P}} \left[\frac{1}{K}\sum_{k=1}^{K} \mathbf{P}_{k,t-1} B^{(k)}_{i,j} \mathbf{P}^T_{k,t-1}\right] +$$

$$C(1-\alpha) \sum_{(X_i,X_j)\in A^t_{D,P}} \left[-\frac{1}{K}\sum_{k=1}^{K} \mathbf{P}_{k,t-1} B^{(k)}_{i,j} \mathbf{P}^T_{k,t-1}\right] \tag{7}$$

Where $B^{(k)}_{i,j} = (x_i^{(k)} - x_j^{(k)})(x_i^{(k)} - x_j^{(k)})^T$ is a rank 1 matrix.

Step 2:

$$\mathbf{M}_{t+1} = \mathbf{V}^T [\mathbf{\Sigma}]_+ \mathbf{V}, \tag{8}$$

where $\mathbf{V}^T\mathbf{\Sigma}\mathbf{V}$ is the eigenvalue decomposition of $\mathbf{M}_{tmp}$. Projecting $\mathbf{M}_{tmp}$ 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 $\mathbf{P}$

Fixing $\mathbf{M}$, each $\mathbf{P}_k$ can be updated separately through gradient descent as

$$\mathbf{P}_{k,t} = \mathbf{P}_{k,t-1} - \eta g_t(\mathbf{P}_k), \quad k = 1, 2, ..., K, \tag{9}$$

where $g_t(\mathbf{P}_k)$ is the gradient of $L(\mathbf{P}_k)$ at time t and it is derived as

$$g_t(\mathbf{P}_k) = 2\gamma \mathbf{P}_{k,t-1} + C\alpha \sum_{(X_i,X_j)\in A^t_{S,M}} \left[\frac{2}{K}\mathbf{M}_t \mathbf{P}_{k,t-1} B^{(k)}_{i,j}\right] +$$

$$C(1-\alpha) \sum_{(X_i,X_j)\in A^t_{D,M}} \left[-\frac{2}{K}\mathbf{M}_t \mathbf{P}_{k,t-1} B^{(k)}_{i,j}\right] \tag{10}$$

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 γ, α, η, μ, β, C and maximum iteration T.

Initialization:

To initialize $\mathbf{P}_k$ (k = 1,2,...,K):
construct $\mathbf{X}^k \in \mathbb{R}^{l_k \times N}$ of $x_i^{(k)}$ from S and D;
perform PCA on $\mathbf{X}^k$ to obtain $\mathbf{P}_{k,0} \in \mathbb{R}^{d\times l_k}$.
To initialize $\mathbf{M}$:
set $\mathbf{M}_0 = \mathbf{I}_{d\times d}$.

Main loop:

for $t = 1 : T$ **do**
calculate $A^t_{S,P}$ and $A^t_{D,P}$ to update $\mathbf{M}$ through (7), (6) and (8);
calculate $A^t_{S,M}$ and $A^t_{D,M}$ to update $\mathbf{P}_k$ through (10) and (9).
end

Outputs:

$\mathbf{P}_k$ $(k = 1, 2, \ldots, K)$ and $\mathbf{M}$.

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) the proposed HM3L method as well as other publicly available multiple metrics learning methods [10, 43] to first transform the features used in the baseline result, then apply NN or SVM classifier, (3) 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 reported in [39], same subset of tagged images, same train/test splitting, same sets of similar (dissimilar) pairs of instances and same feature extraction procedures are applied. A subset of 1521 tagged images are used. These tagged images consist of 30 classes (actor, airplane, bicycle, bridge, buddha, building, butterfly, camels, car, cathedral, cliff, clouds, coast, computers, desert, flag, flowers, food, forest, glacier, hills, lake, leaf, monks, moon, motorcycle, mushrooms, ocean, police, pyramid) and roughly 50 tagged images per class are randomly selected. By randomly splitting the dataset, 765 tagged images are used as training data and the remaining are used as testing data. From the training data, 9613 pairs of similar instances and 10067 pairs of dissimilar instances are selected to learn distance metrics. For images, 1024-D bag of visual words based on SIFT descriptors is extracted to represent the image modality; for tags, 1000-D bag of words is extracted to represent the associated tag modality. Therefore, one instance of tagged image is represented by feature vectors of two modalities.

5.1.1 Experiment Setup

For every approach considered, distance metrics are first learned. Then, KNN classification under the learned distance metrics is performed using the training and testing data. The value of K is chosen to be 1, 3, 5, 10 and 20. We compare the performance of our method with that of "Xing + Original", "ITML+Original", "Xing + MWH", "ITML + MWH", "MKE" [26], Heterogeneous Multi-Metric Learning (HMML) [43] and PMML [10]. "Xing+Original" and "ITML+Original" methods essentially apply algorithms proposed in [40] and [12] on the concatenated feature vectors from different modalities. Similarly, "Xing+MWH" and "ITML+MWH" correspond to the algorithms combined with the MWH model proposed in [39]. All parameters are tuned using cross-validation on training data.

5.1.2 Experimental Results

Table 1 shows the KNN classification accuracies of different methods. As can be seen from the table, the HM3L method performed the best and it outperforms all the other methods. This experiment clearly show that our method can provide better distance measures which can enhance the performance of a classification algorithm.

To show whether the proposed algorithm converge, 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 30^0, 45^0 and 60^0. A single RGB-D frame which consists of both an RGB image and a depth image. Evaluations 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.

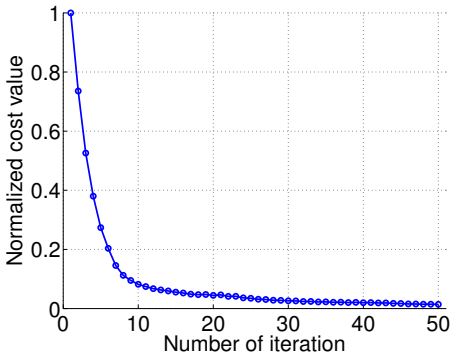

Figure 2. Normalized cost function over iterations.

Kernel descriptors [3] [4] were extracted as features for RGB images and depth image. For RGB images, LBP kernel descriptor, Gradient kernel descriptor and normalized color kernel descriptor were extracted. For depth images, gradient kernel descriptor, 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 was obtained from 1000 dimensional basis vector for 1×1, 2×2, 3×3 pyramid sub-regions. The basis vector was learned by K-means on about 400,000 sample kernel descriptors from training data. The dimensionality of each kernel descriptor is $(1 + 4 + 9) \times 1000 = 14000$ and further apply principal component analysis to reduce the the dimensionality to 1000. After feature extraction, each RGB-D image is represented by 7 kernel descriptors and each kernel descriptor is 1000 dimensional vector.

5.2.1 Experimental Setup

For the instance recognition experiment, images corresponding to the videos captured at angles 30^0 and 60^0 are used for training, and images corresponding to the videos captured at angle 45^0 are 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. 10 trials were repeated for category recognition.

For the instance and category recognition tasks, we first learn multiple metrics for 7 kernel descriptors using the similar and dissimilar set of the RGB-D images generated from the training data. We then perform linear SVM classification [14] based on the learned metrics. We also compare 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 Experiment Results

Classification results for instance recognition and category recognition are shown in Table 2 and Table 3 respectively. From

Methods	Xing+Original	ITML+Original	Xing+MWH	ITML+MWH	MKE[26]	Xie[39]	PMML[10]	HMML[43]	HM3L
1-NN	0.8995	0.8995	0.8995	0.9286	0.8056	0.9352	0.9233	0.9140	**0.9524**
3-NN	0.8108	0.6653	0.8849	0.8929	0.6944	0.9021	0.9220	0.9246	**0.9431**
5-NN	0.6971	0.4868	0.8426	0.8519	0.5860	0.8849	0.9299	0.9114	**0.9418**
10-NN	0.4775	0.2394	0.7646	0.7394	0.4405	0.8333	0.9139	0.9008	**0.9339**
20-NN	0.1548	0.0450	0.6230	0.4841	0.1746	0.7130	0.9074	0.8876	**0.9223**

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

Methods	RGB	Depth	RGB-D
Lai [20]	60.7	46.2	74.8
Bo [4]	90.8	54.7	91.2
Blum [2]	82.9	-	90.4
HMP [5]	92.1	51.7	92.8
MMSS [34]	-	-	94.0
PMML [10] + linear SVM	92.7	53.4	92.9
HMML [43] + linear SVM	90.0	51.9	92.1
HM3L + linear SVM	**93.34**	**55.6**	**95.0**

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

Methods	RGB	Depth	RGB-D
Lai [20]	64.7±2.2	74.5±3.1	83.8 ± 3.5
Bo [4]	80.7±2.1	80.3±2.9	86.5 ±2.1
Blum [2]	-	-	86.4 ±2.3
HMP [5]	**82.4 ± 3.1**	**81.2 ± 2.3**	87.5 ±2.9
MMSS [34]	-	-	88.5 ± 2.2
PMML [10] + linear SVM	80.2	77.7 ± 2.4	88.5 ± 1.4
HMML [43] + linear SVM	75.8± 3.2	77.4 ± 2.4	87.3 ± 1.8
HM3L + linear SVM	81.0 ± 2.7	79.1 ± 2.4	**89.2 ± 1.6**

Table 3. Category recognition accuracy on RGB-D Object dataset.

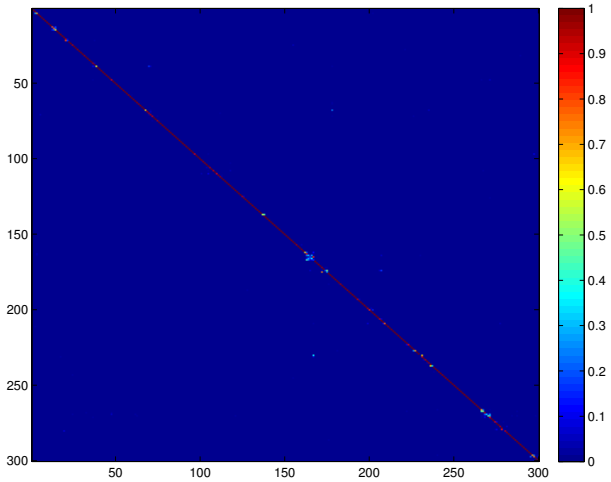

Figure 3. Confusion matrix for Instance recognition result.

these tables, we have following observations. (1) the proposed HM3L-based classification outperform 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 10 trials. (2) The proposed HM3L algorithm can boost the classification accuracy compared to the case where metrics learning were 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

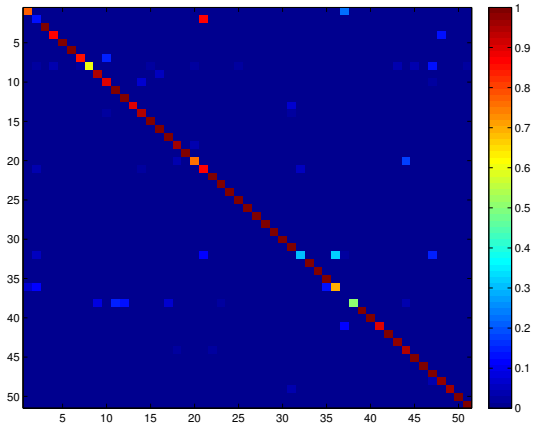

Figure 4. Confusion matrix for 8th trial category recognition result.

Figure 5. Examples of prediction errors in category recognition experiment.

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 large portion of the 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 18 categories of common household and office objects. Each category contains between 3 to 14 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 10 degrees and yields 36 views per object. Each view is one data sample consisting of RGB image and Depth image. Follow the similar procedures 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]. 6 objects per category were used for training and remaining objects were used for testing. For each object, 18 views are selected for training and 18 views for testing. The training set consisted of 82 objects with a total of 1476 views. The test set contained 74 objects with 1332 views. Same methods as included in RGB-D dataset are 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.

Methods	RGB	Depth	RGB-D
Browatzki [7]	66.6	74.6	82.8
HMP [5]	86.3	**87.6**	91.0
MMSS [34]	-	-	91.3
PMML [10] + linear SVM	**90.6**	82.7	91.8
HMML [43] + linear SVM	86.8	83.4	90.8
HM3L + linear SVM	89.9	86.4	**92.9**

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

5.4. Scene Categorization on SUN RGB-D dataset

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, 19 scene categories with more than 80 images are 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 are used for training and 4659 samples are used for testing. The standard average categorization accuracy is used for evaluation. We apply the proposed HM3L method to the Places-CNN features, transform the original features with the learned matrices, and then apply one-vs-all rbf SVM for classification. The scene category recognition results are shown in Table 5.

From results, we have following observations. (1) the proposed HM3L-based classification outperform the best results obtained from [35, 47]. (2) The proposed HM3L algorithm as well other two multiple metrics learning algorithms can significantly boost the classification accuracy compared to the baseline case in which metrics learning were not performed. (3) HM3L-based multimodal classification outperforms other multiple metrics learning-based classification and this again shows that the importance of capturing the relationship for different multiple metrics in the learning process.

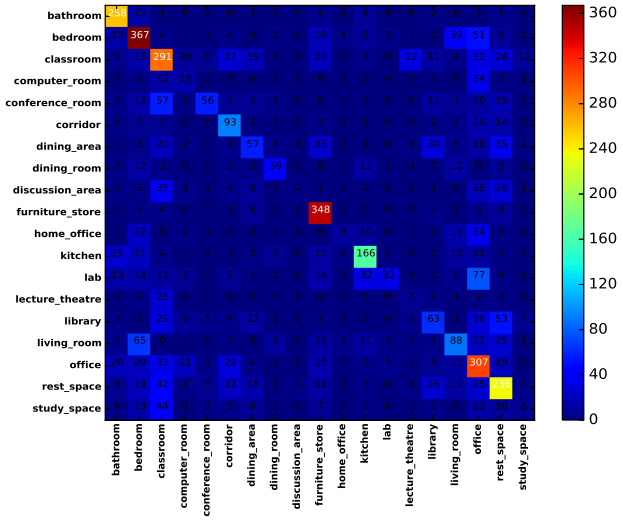

Figure 6. Confusion matrix for scene recognition result.

Methods	RGB	Depth	RGB-D
Place-CNN + linear SVM [32]	35.6	25.5	37.2
Place-CNN + rbf SVM [32]	38.1	27.7	39.0
Liao [21]	36.1	-	41.3
Zhu [47]	-	-	41.5
Wang [35]	-	-	48.1
PMML [10] + rbf SVM	40.7	30.5	44.2
HMML [43] + rbf SVM	47.9	32.6	51.1
HM3L + rbf SVM	**48.6**	**33.2**	**52.3**

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

6. Conclusions

In this paper, we proposed hierarchical multimodal metric learning algorithm which can efficiently learn multiple metrics for multi-modal data while fully exploring the relationship 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 distance 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

[1] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2013. 2

[2] M. Blum, J. T. Springenberg, J. Wlfing, and M. Riedmiller. A learned feature descriptor for object recognition in rgb-d data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1298–1303, May 2012. 6

[3] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 244–252, 2010. 5

[4] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 821–826, Sept 2011. 5, 6

[5] L. Bo, X. Ren, and D. Fox. Unsupervised feature learning for RGB-D based object recognition. In *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER 2012, June 18-21, 2012, Québec City, Canada*, pages 387–402, 2012. 6, 7

[6] S. Boyd and A. Mutapcic. Stochastic subgradient methods, 2007. 4

[7] B. Browatzki, J. Fischer, B. Graf, H. H. Blthoff, and C. Wallraven. Going into depth: Evaluating 2d and 3d cues for object classification on a new, large-scale object dataset. In *IEEE International Conference on Computer Vision Workshops*, pages 1189–1195, Nov 2011. 1, 4, 6, 7

[8] S. Chopra, R. Hadsell, and Y. Lecun. Learning a similarity metric discriminatively, with application to face verification. In *In Proc. of Computer Vision and Pattern Recognition Conference*, pages 539–546. IEEE Press, 2005. 2

[9] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*, 2009. 4, 5

[10] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3554–3561, Washington, DC, USA, 2013. 2, 3, 4, 5, 6, 7

[11] J. V. Davis and I. S. Dhillon. Structured metric learning for high dimensional problems. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 195–203, New York, NY, USA, 2008. ACM. 2

[12] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, Corvalis, Oregon, USA, 2007. 1, 2, 5

[13] X. Di and V. M. Patel. Large margin multi-modal triplet metric learning. In *IEEE International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–8, 2017. 2

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008. 5

[15] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. 2

[16] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 902–909, June 2010. 1

[17] J. Hu, J. Lu, J. Yuan, and Y. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Computer Vision 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014*, pages 252–267, 2014. 2, 3

[18] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *IEEE International Conference on Computer Vision Workshops*, pages 1168–1174, Nov 2011. 1, 7

[19] B. Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012. 2

[20] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824, May 2011. 1, 4, 5, 6

[21] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 2318–2325, 2016. 7

[22] D. K. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In *International Conference on Machine Learning*, 2013. 2

[23] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu. Semi-supervised sparse metric learning using alternating linearization optimization. In *International Conference on Knowledge Discovery and Data Mining*, pages 1139–1148, 2010. 2

[24] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S. Chang. Low-rank similarity metric learning in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2792–2799, 2015. 2

[25] B. McFee and G. Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, June 2010. 2

[26] B. McFee and G. Lanckriet. Learning multi-modal similarity. *Journal of Machine Learning Research*, 12:491–523, February 2011. 5, 6

[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696, 2011. 1

[28] R. Rosales and G. Fung. Learning sparse metrics via linear programming. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 367–373, 2006. 2

[29] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):113–126, Jan 2014. 1

[30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, pages 746–760, 2012. 1, 7

[31] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013*, 2013. 2

[32] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, June 2015. 1, 4, 7

[33] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014. 1

[34] A. Wang, J. Cai, J. Lu, and T. J. Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1125–1133, Dec 2015. 1, 5, 6, 7

[35] A. Wang, J. Cai, J. Lu, and T.-J. Cham. Modality and component aware feature fusion for rgb-d scene classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[36] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006. 1, 2

[37] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 153–162, New York, NY, USA, 2013. ACM. 1

[38] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *2013 IEEE International Conference on Computer Vision*, pages 1625–1632, Dec 2013. 1, 7

[39] P. Xie and E. P. Xing. Multi-modal distance metric learning. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1806–1812. AAAI Press, 2013. 2, 5, 6

[40] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*. 1, 2, 5

[41] P. Yang, K. Huang, and C. Liu. Multi-task low-rank metric learning based on common subspace. In *Neural Information Processing - 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part II*, pages 151–159, 2011. 1

[42] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2214–2222. Curran Associates, Inc., 2009. 2

[43] H. Zhang, T. Huang, N. Nasrabadi, and Y. Zhang. Heterogeneous multi-metric learning for multi-sensor fusion. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8, July 2011. 2, 3, 4, 5, 6, 7

[44] H. Zhang, V. M. Patel, and R. Chellappa. Robust multimodal recognition via multitask multivariate low-rank representations. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pages 1–8, 2015. 1

[45] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multimodal image retrieval. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 846–851 Vol. 1, Oct 2005. 1

[46] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems 27*, pages 487–495. 2014. 7

[47] H. Zhu, J.-B. Weibel, and S. Lu. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7