Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Recognition

Hongyuan Zhu
I^{2}R, A*Star, Singapore
zhuh@i2r.a-star.edu.sg

Jean-Baptiste Weibel
Georgia Tech, USA
Jean-Baptiste.weibel@gatech.edu

Shijian Lu
I^{2}R, A*Star, Singapore
slu@i2r.a-star.edu.sg

Abstract

RGBD scene recognition has attracted increasingly attention due to the rapid development of depth sensors and their wide application scenarios. While many research has been conducted, most work used hand-crafted features which are difficult to capture high-level semantic structures. Recently, the feature extracted from deep convolutional neural network has produced state-of-the-art results for various computer vision tasks, which inspire researchers to explore incorporating CNN learned features for RGBD scene understanding. On the other hand, most existing work combines rgb and depth features without adequately exploiting the consistency and complementary information between them. Inspired by some recent work on RGBD object recognition using multi-modal feature fusion, we introduce a novel discriminative multi-modal fusion framework for rgbd scene recognition for the first time which simultaneously considers the inter- and intra-modality correlation for all samples and meanwhile regularizing the learned features to be discriminative and compact. The results from the multi-modal layer can be back-propagated to the lower CNN layers, hence the parameters of the CNN layers and multi-modal layers are updated iteratively until convergence. Experiments on the recently proposed large scale SUN RGB-D datasets show that our method achieved the state-of-the-art without any image segmentation.

1. Introduction

Indoor scene recognition based on RGBD camera has attracted increasingly attention due to its wide applications in computer vision and robotics and the popularity of depth sensors. Although much progress has been achieved in the past few years, indoor scene recognition is still challenging due to the large intra-object variation and spatial layout changes, not to mention the challenges caused by the occlusion and low-light condition.

Given an input indoor image, we human can quickly recognize the scene category and generalize the brained learned recognizer to the new place which is not seen before. The key to success of our brain lies in three aspects: **(1)** its exposure to the dense and diversity sampling of our visual world; **(2)** its versatile capacity to abstract compact and discriminative representations of different complexities; **(3)** its high efficiency to fuse information from multi-modalities to perform high-level reasoning.

To improve current scene recognition systems, a large dataset is very helpful to learn meaningful representations and prevent overfitting. Early scene-centric datasets [21, 13] for outdoor and indoor scene recognition either has a small number of scene categories or dataset size compared with the counterparts in object centric datasets, such as ImageNet [14]. Recently, Zhou et al [22] proposed the Places dataset which contains 7 million images from 476 places categories, making it the largest scene and place database, which allows us to train data hungry models. Although depth sensors has become cheaper, the number of influential RGBD dataset is still relative smaller than RGB counterparts. The first generation benchmarks, such as NYU D2 [16] and Berkeley B3DO dataset [6], have bootstrapped the initial progress of RGBD scene understanding. Recently, Song et al [17] proposed the large scale SUN-RGBD dataset, which is the first RGBD dataset that has a similar scale as the PASCAL VOC dataset [3], hence make it possible to borrow the success technique in RGB scene understanding to RGBD research.

Besides a large dataset, features are also vital for scene recognition. Many features have been proposed for scene recognition, here we just briefly mention some recent influential and related works due to the space limits. Early scene recognition uses hand-crafted features such as SIFT [11], GIST [12] and CENTRIST [20], which achieved reasonable performance for certain tasks. On the other hand, as the hand-crafted features models low-level activations from V1 cells, they can ignore the more discriminative information in higher hierarchies which are vital for scene understanding. Moreover, the hand-craft features also have a low generalization capability when migrating to new tasks. To avoid the limitations of hand-craft features, the unsupervised feature learning has been proposed, such as deep belief nets [5], deep Boltzmann machines [15], deep autoen-
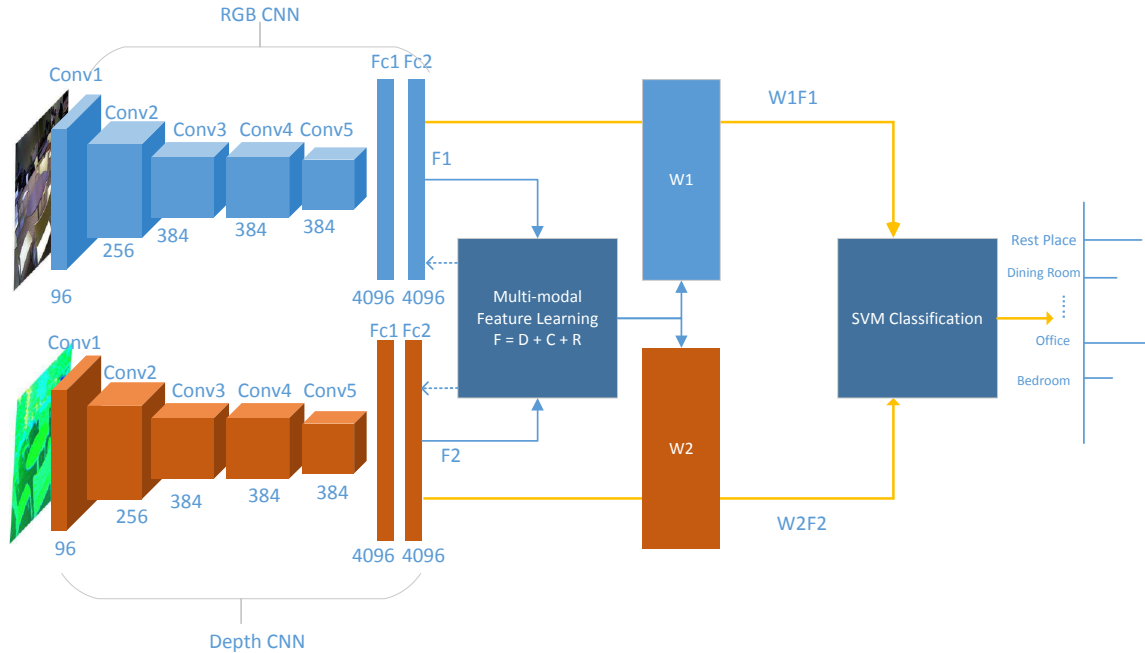
Figure 1. Pipeline of our multi-modal feature fusion framework for RGBD indoor scene recognition. The input to our system are small batches of RGB images and corresponding HHA[4] encoded depth images. The CNNs for RGB and depth modalities are then separately trained and the features from the second fully connected layers are fed into the multi-modal learning layer F for learning two projection matrix W_1 and W_2. F formulates an objective function which maximizes the inter- and intra-modality associations for samples from the same class and vice versa for samples from different classes, and meanwhile regularize the learned features to be discriminative and compact. The blue solid lines denote the forward training of multi-modal layer F. The blue dashed lines denote the parameter back-propagation process. The golden lines denote the final feature fusion between two modalities for final rgbd scene recognition.

coders [8], convolutional deep belief networks [9], hierarchical sparse coding. The recent work of Bo et al [1] on Hierarchical Matching Pursuit and its Multipath variant [2] has achieved good performance for various tasks, including RGBD object recognition. While, the feature hierarchies learned by unsupervised feature learning is still comparatively shallower than the recent popular deep convolutional neural network [7], which first appears in ImageNet classification challenges ILSVRC-2012. Recently, [10] explored to incorporate segmentation in CNN learning framework and achieved the state-of-the-art for rgbd scene recognition.

To learn features for RGBD scene recognition, one can apply existing methods to color and depth modalities separately, or simply treat RGBD as un-differentiated four-channel data. Such separate learning and un-differential handling can ignore the consistency and complementary information between the two modalities and their relative importance for various tasks. Hence, the relationship between different modalities have not been thoroughly investigated for rgbd scene recognition. To resolve the above issue, we proposed a discriminative multi-modal feature fusion framework for RGBD scene recognition. The proposed framework is illustrated in Fig.1. The basic idea is

that we seek to transform the activations from the trained rgb and depth CNNs to a common subspace, such that we can discover the discriminative features for both modalities and simultaneously increase the association between same class' samples and decrease the association between different classes' samples for both intra- and inter-modalities. Our experiments on SUN-RGBD dataset shows that our method out-performed [10] without any image segmentation. More importantly, our research highlights the potential of appropriate feature fusion for RGBD scene recognition, which is worthwhile for further research.

Our work is inspired by two recent works [19] and [18] for RGBD object recognition, which applied multi-modal feature learning to fuse the response from the CNNs trained with RGB image and surface normal images. There are several differences between our work and these two works. Firstly, the task domain is quite different. [19] and [18] targets the RGBD object recognition for the prop-like images in controlled environment, while for scene recognition, the image is much more cluttered. In terms of theoretical side, [19] enforced the intra- and inter-modalities correlations between pairwise samples and [18], enforce the correlations between the features of each sample individually. While, we
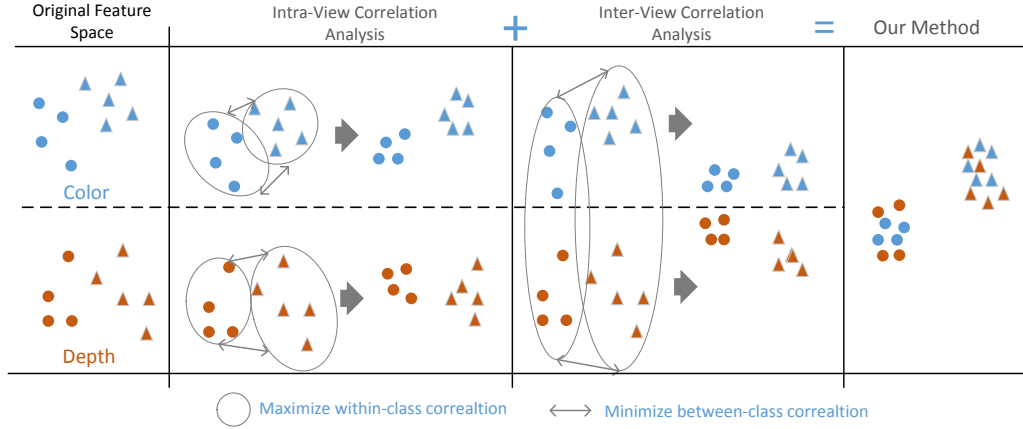
Figure 2. Conceptual illustration of our method, (⃝ and △) denote the samples from different classes, and the blue and orange colors denote the samples from rgb and depth modalities separately.

consider both information in our formulation. In addition, our formulation includes the regularization for the within-class and between-class inter-view associations, which is ignored in [19]. We prove the necessity of this regularization by testing our implementation of [19] on the SUN-RGBD dataset. The results demonstrate our methods' theoretical and performance advantages. Moreover, in terms of initial feature formulation, we include CNN features fine tuned on the HHA encoded depth layer [4], which has proved to be more discriminative than raw depth image and surface normal image.

2. Proposed method

Let $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{d_1 \times N}$ and $Y = [y_1, y_2, ..., y_N] \in \mathbb{R}^{d_2 \times N}$ denote the d_1- and d_2-dimensions of the activations from the second fully connected (fc2) layer of color and depth CNN in one data batch of N images. Let $W_i \in \mathbb{R}^{M_i \times d_i}$ be the transformation matrix for the modality $i, (i = 1, 2)$, and $F_1 \in \mathbb{R}^{M_1 \times N}$ and $F_2 \in \mathbb{R}^{M_2 \times N}$ be the learned features for the rgb and depth modalities, respectively. M_i is the projected feature dimension for color ($i = 1$) and depth features ($i = 2$).

Our task is to learn a new representation $F_1 = W_1 X$ and $F_2 = W_2 Y$ such that the correlation for same class' samples are maximized for both inter- and intra-modalities, and vice versa for different classes' samples, and require the learned features to be compact and discriminative according to the input samples. Finally, the learned features $F_s = [F_1 \quad F_2]$ are fed to the SVM to train the final scene classifier. Noted, our framework can incorporate any state-of-the-art features as input and any state-of-the-art classifiers for final prediction, e.g. we can incorporate the sparse logistic regression to from an end-to-end learning system and back-propagate the parameters to the lower layers as in

[18]. We choose SVM in our work due to its robustness to outliers. An illustration of our method is shown in Fig. 2.

2.1. Formulation

To learn features for both modalities, our objective function is formulated as

$$\min_{\{W_1, W_2, F_1, F_2, \alpha_1, \alpha_2\}} F = \sum_{i=1}^{2} \alpha_i^p [t_1 D_i + t_2 R_i] + \beta(C_1 - \lambda_2 C_2)$$

$$\text{subject to } \alpha_1 + \alpha_2 = 1, \alpha_1 \geq 0, \alpha_2 \geq 0$$

$$(1)$$

where $D_i := D_i(W_i)$ is the intra-modality discriminative term and $R_i := R_i(W_i, F_i)$ is the intra-modality reconstruction term for modality $i, i = \{1, 2\}$, which are to be elaborated in following parts. These two terms are balanced by the trade-off factor t_1 and t_2 respectively. α_1 and α_2 are self-adjusted weights for each modality. The hyperparameter p is introduced to avoid trivial sub-optimal solution when only one modality is selected before such nonlinear handling [19]. C_1 and C_2 are the inter-modality correlation terms for within-class and between-class samples.

Intra-modality discriminative term: The discriminative term $D_1(W_1)$ for RGB modality is intended to find a W_1 to project the RGB CNN activations X to a space in which the distance between x_i and x_j is small if they are of the same class, otherwise large if they are from different classes. $D_2(W_2)$ is similarly defined for the depth CNN activations Y. The constraint can be defined as: if two objects are from the same class ($y_{ij} = 1$), their relative feature distance should be smaller than a given threshold $\mu_1 - \tau_1$, otherwise ($y_{ij} = -1$) the distance should be larger than $\mu_1 + \tau_1$, which is similarly defined as in [19]. Mathemati-

cally, this can be expressed as

$$y_{ij}(\mu_1 - d_{W_1}(x_i, x_j)) > \tau_1 \quad (2)$$

where the distance between a pair of the CNN activations x_i and x_j is computed as

$$d_{W_1}(x_i, x_j) = (W_1 x_i - W_1 x_j)^T \cdot (W_1 x_i - W_1 x_j) \quad (3)$$

and the discriminative term is defined as follows

$$D_1(W_1) = \sum_{\forall i,j} h(\tau_1 - y_{ij}(\mu_1 - d_{W_1}(x_i, x_j))) \quad (4)$$

where h is a smoothed hinge loss function $h(x) = max(0, x)$, where $D_2(W_2)$ is similarly defined for the depth modality.

Intra-modality reconstruction term: The reconstruction term $R_1(W_1, F_1)$ for color modality is defined as

$$R_1(W_1, F_1) = \\ ||W_1 X - F_1||_F^2 + ||W_1^T F_1 - X||_F^2 + \lambda_1 g(F_1) \quad (5)$$

this term enforces the learned feature F_1 to be similar to the W_1 transformed X, while the second term encourages F_1 to reconstruct X when back-transformed via W_1^T, and the third term g is the smooth L_1 penalty function [18]. With the reconstruction term, the supervised information is introduced to allow W_1 better fit the observed training data individually. Likewise, $R_2(W_2, F_2)$ corresponds to the reconstruction error for the depth modality Y.

Inter-modality correlation term: With the discriminative term and reconstruction term, the distances between same class samples are decreased and the distances between different classes samples are increased for each modality. However, the data captured from different modalities may suffer from missing information or noise pollution, hence we seek to exploit the correlation between different modalities to reduce misclassification, such that the association of within-class samples are maximized, while the association between different classes are minimized for inter-views. Such regularization is introduced by adding two inter-view correlation terms C_1 and C_2 which minimize the pairwise distances between the color and depth modalities of the same class and vice versa for the samples from different classes.

$$C_1(W_1, W_2) = \sum_c \sum_{\forall i,j \in c} [\sqrt{(d_{W_1}(x_i, x_j))} - \sqrt{(d_{W_2}(y_i, y_j))}]^2 \quad (6)$$

and

$$C_2(W_1, W_2) = \sum_{c,d}^{c \neq d} \sum_{\forall i \in c, j \in d} [\sqrt{(d_{W_1}(x_i, x_j))} - \sqrt{(d_{W_2}(y_i, y_j))}]^2 \quad (7)$$

2.2. Alternating optimization

To our knowledge, there is no closed-form solution to Eq.1 because we need to solve W_k, F_k and a_k jointly. To address this, we adopt the alternating approach to optimize W_k, F_k, a_k. The pseudocode of our algorithm is illustrated as in Algorithm 1:

Algorithm 1 Optimizing our proposed feature fusion framework

Input: Training set with two modalities: X, Y, the corresponding label
Output: Feature projection matrix: W_1, W_2

  Step 1 (Initialization):
  Initializae $W_1, W_2, F_1, F_2, a_1, a_2$
  Step 2 (Optimization)
  for k =1, 2, ..., K do
    2.1. Fix W_1, W_2, a_1, a_2 **Update** F_1, F_2 **with (8)**

    2.2. Fix F_1, F_2, a_1, a_2, **Update** W_1, W_2 **with (10)**

    2.3. Fix W_1, W_2, F_1, F_2, **Update** a_1, a_2 **with (12)**

  end for

In Step 2.1 and 2.2, we update the other variables using the gradient descent algorithm, where the same learning rate γ is used. In Step 2.1, we update F_1, F_2, the derivate of F respect to F_1 are shown below:

$$\frac{\partial F}{\partial F_1} = 2a_1^p t_2 [(F_1 - W_1 X) \\ + W_1 (W_1^T F_1 - X) + \lambda_1 g'(F1)] \quad (8)$$

then F_1 is updated as

$$F_1 \leftarrow F_1 - \gamma \frac{\partial F}{\partial F_1} \quad (9)$$

In step 2.2, when F_1, F_2, a_1, a_2 are fixed, W_i is updated, e.g.

$$\frac{\partial F}{\partial W_1} = 2\alpha_1^p t_1 [(W_1 X - F_1) X^T + W_1 (W_1^T F_1 - X)] \\ + 2t_2 W_1 [\alpha_1^p \sum_{\forall i,j} y_{ij} h'(\tau_1 - y_{ij}(\mu_1 - dW_1(x_i, x_j)) A_{i,j}) \\ + \beta(\sum_c \sum_{\forall i,j \in c} (1 - \sqrt{\frac{dW_2(y_i, y_j)}{dW_1(x_i, x_j)}}) A_{i,j} \\ - \lambda_2 \sum_{c,d}^{c \neq d} \sum_{\forall i \in c, j \in d} (1 - \sqrt{\frac{dW_2(y_i, y_j)}{dW_1(x_i, x_j)}}) A_{i,j}))] \quad (10)$$

where $A_{i,j} = (x_i - x_j)^T (x_i - x_j)$

$$W_1 \leftarrow W_1 - \gamma \frac{\partial F}{\partial W_1} \quad (11)$$

Then by fixing W_1, W_2, F_1, F_2, we can update α_1, α_2 accordingly by attaching Lagrange multiplier based on :

$$
\begin{aligned}
L(\alpha, \eta) =& \alpha_1^p T_1 + \alpha_2^p T_2 - \eta(\alpha_1 + \alpha_2 - 1) \\
=& \alpha_1^p (t_1 D_1(W_1) + t_2 R_1(W_1, F_1)) \\
&+ \alpha_2^p (t_1 D_2(W_2) + t_2 R_2(W_2, F_2)) \\
&+ -\eta(\alpha_1 + \alpha_2 - 1)
\end{aligned}
\tag{12}
$$

By setting $\frac{\partial L(\alpha, \eta)}{\alpha}$ and $\frac{\partial L(\alpha, \eta)}{\eta}$ to 0, α_i can be updated as:

$$
\alpha_i = \frac{(1/T_i)^{1/(p-1)}}{\sum_{i=1}^{2} (1/T_i)^{1/(p-1)}}
\tag{13}
$$

Finally, the learned weight can be back-propagated to the lower layer of CNN by

$$
\frac{\partial F}{\partial x_i} = \alpha_1^p t_1 \frac{\partial D_1(W_1)}{\partial x_i} + \beta \frac{\partial C}{\partial x_i} + \alpha_1^p t_2 \frac{\partial R_1(W_1, F_1)}{\partial x_i} \tag{14}
$$

where

$$
\frac{\partial D_1(W_1)}{\partial x_i} = \sum_{\forall j} 2 y_{ij} W_1^T W_1 (x_i - x_j) \\
h'(\tau_1 - y_{ij}(\mu_1 - d_{W_1}(x_i, x_j)))
\tag{15}
$$

$$
\frac{\partial C(W_1, W_2)}{\partial x_i} =
$$

$$
2 W_1^T W_1 \sum_c \sum_{\forall i,j \in c} (x_i - x_j) \sqrt{\frac{d_{W_1}(x_i, x_j) - d_{W_2}(y_i, y_j)}{d_{W_1}(x_i, x_j)}}
$$

$$
- \lambda_2 \sum_{c,d}^{c \neq d} \sum_{\forall i \in c, j \in d} (x_i - x_j) \sqrt{\frac{d_{W_1}(x_i, x_j) - d_{W_2}(y_i, y_j)}{d_{W_1}(x_i, x_j)}}
\tag{16}
$$

$$
\frac{\partial R_1(W_1, F_1)}{\partial x_i} = W_1^T(W_1 x_i - f_i) - (W_1^T f_i - x_i) \tag{17}
$$

for color modality X and is similarly defined for depth modality Y.

3. Experiments

To evaluate the effectiveness of our proposed method, we perform experiments on recently proposed SUN-RGBD dataset [17]. The details of the experiments and the results are discussed in the following sections.

3.1. Datasets and experiment setup

SUN-RGBD Dataset: The SUN-RGBD dataset is the first large scale dataset which has a similar scale as PASCAL VOC [3]. The dataset was captured by four different sensors, and contains 10,335 RGB-D images in 47 scene categories. For scene categorization, the benchmark of

scene classification is conducted on 19 subsets of the dataset with more than 80 samples.

Architecture of CNNs The architecture of the CNN for scene classification are exactly the same as the AlexNet [7]. The network contains eight layers with weights, with five convolutional layers and the three fully-connected layers. The network has about 60 million parameters. For scene classification, the network was initialized by using the network of PlacesCNN [22], which was trained from 205 categories of places with minimum 5,000 and maximum 15,000 images per-category. The last fully-connected layer is removed and the second fully-connected layer (fc2) is used for feature extraction. Then we fine-tune the network for the RGB images in SUNRGBD dataset, which form the indoor scene centric CNN. To train the depth CNN, we first encode the depth images with the HHA encoding method proposed by Gupta et al. [4], which generates three channels at each pixel with the information of horizontal disparity, height above ground and the angle the pixel's local surface normal makes with the inferred gravity direction. HHA encodes the properties of geometric pose that emphasize complementary discontinuities in the image. After the HHA encoding, we fine-tune the RGB CNN on the encoded depth dataset. We use the recent popular deep learning platform Caffe, the network was trained on a Titan X.

3.2. Parameter setting and analysis

For our multi-modal feature fusion framework, the dimension M_i of the projected features are set to the same $M_i = 2048$ for both modality, the effects of different number of M_i can be observed in Fig.3, one can see that both excessive small or large M_i lead to performance drop. The parameters $\beta = 1e^{-10}, \lambda_1 = 1, \lambda_2 = 1e^{-11}, \mu_1 = 100, \mu_2 = 1000, \tau_1 = 1, \tau_2 = 1, K = 300$ are tuned empirically on the validation set and then fixed during testing.

We also evaluate the contribution of each term by incrementally including them in our objective function, which is achieved by turning the related trade-off factors λ_2 and t_2 from 0 to our validated parameters. We also self-implemented Wang et al.'s method [19] and tested its performance on the SUN RGB-D dataset with the parameters tuned on the validation set. From Table.1, one can observe that by considering within-class and between-class correlation for inter-modality and intra-modality explicitly, we achieved 6.3% improvement over Wang et al. [19]. Our full model achieved around 8.7% improvement over the model with just discriminative term and correlation term. The underlying reason is that the discriminative term and correlation term learns discriminative feature for pairwise samples, however final classification is evaluated on each sample individually. Hence the introduction of reconstruction term allow the supervised information to regularize on the learned transformation matrix, such that it can better fit the

observed samples, therefore achieves large performance improvement.

Model	Accuracy(%)
Wang et al [19]	26.5
Discriminative term + Correlation term	32.8
Full model	**41.5**

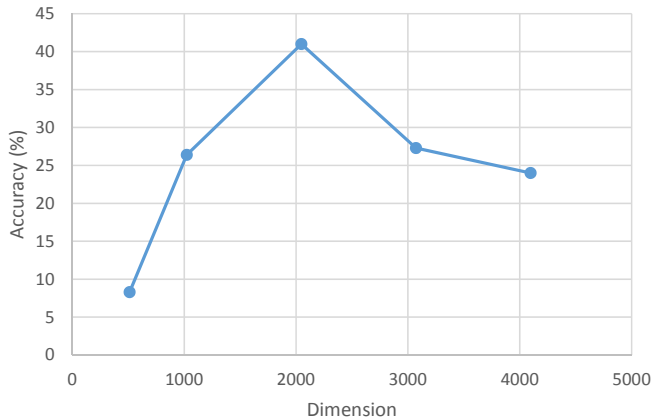Table 1. The illustration of the contribution of each term.

Figure 3. The effect of choosing different M_i.

3.3. Convergence analysis

To show the convergence of our methods, we plot the value of our objective function for the 300 iterations with all the parameters fixed. From the graph, we can see that the objective function converge quickly with around first 200 iterations and converge slowly for next 100 iterations, as shown in Fig.4.
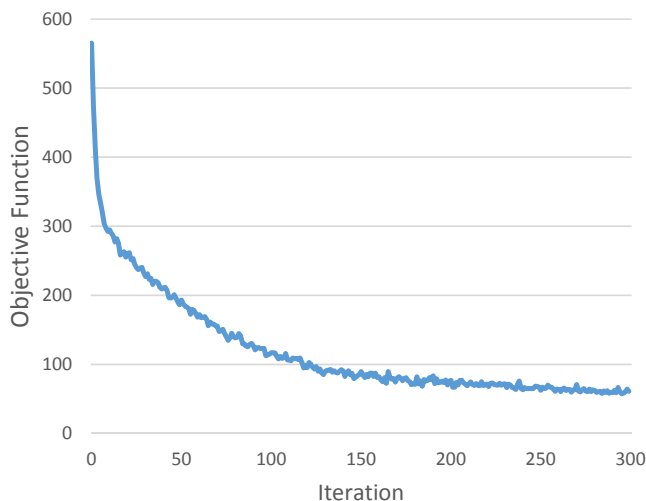

Figure 4. Convergence of the objective function.

3.4. Comparison with other methods

We compare our method with different baselines. Specifically, we compare with the hand-craft descriptor GIST [12], unsupervised feature learning descriptor HMP [1] and supervised AlexNet [7], PlaceCNN [22] and the state-of-the-art SSCNN [10] on SUNRGBD dataset. The accuracy figures of GIST, AlexNet, PlaceNN, SSCNN on RGB or RGB+D are cited from [10]. For HMP, we compute the descriptor on both RGB and HHA encoded depth, then the features are concatenated to train the classifier.

The introduction of the our comparing methods is given as follow:

- GIST [12]. GIST is a famous hand-craft scene descriptor, which computes the spectral information in an image through Discrete Fourier Transform (DFT). The spectral signals are then compressed by the Karhunen-Loeve Transform (KLT).

- Hierarchical Matching Pursuit (HMP) [1] is an recent proposed unsupervised feature learning method. It builds feature hierarchy layer-by-layer using matching pursuit encoder. We use the original code provided by [1] for best performance.

- AlexNet [7]: Since the performance of our method, PlacesCNN and SSCNN are based on AlexNet, hence the performance of original AlexNet is included. AlexNet was trained on the object centric ImageNet dataset, while our method [14], PlacesCNN and SS-CNN on the scene centric dataset, such as Places [22] and SUN-RGBD [17].

- Places-CNN [22]. Place-CNN is pre-trained on 2.5 million scene images using Alexnet. In [22], both Linear SVM and RBF Kernel SVM are considered to train and classify the Place-CNN extracted features on RGB and RGBD.

- SSCNN [10] makes use of a slightly modified AlexNet that trained with the SUN-RGBD dataset. The network is divided into two branches, one for semantic segmentation, the other for image classification. The image classification branch is regularized by the semantic segmentation.

- Our model. We also use our projected features for RGB and RGBD to train linear SVM.

The experiment result is shown in Table.2 and Table.3, where the accuracy is the mean accuracy of 19 scene classes. The confusion matrix is also shown in Fig.5, where the diagonal represent the recognition accuracy of each scene. From the tables, it can be seen that the CNN trained with scene centric databases, such as Place-CNN, SSCNN
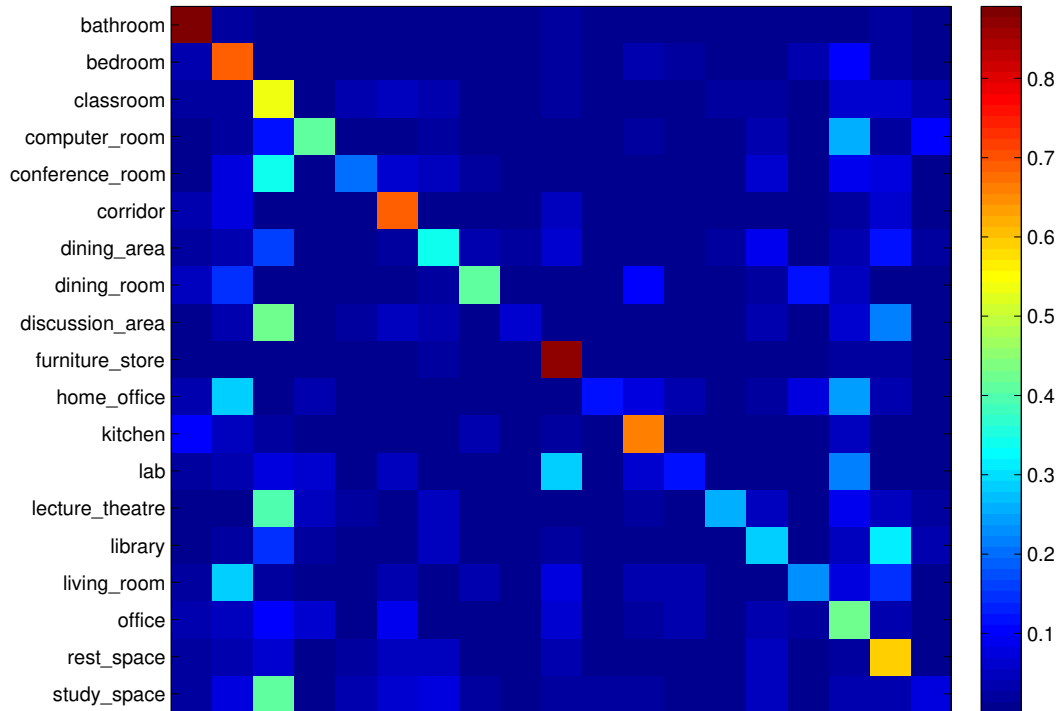
Figure 5. The confusion matrix of our method's scene recognition performance on SUN-RGBD datasset. The vertical axis shows the ground-truth classes, the horizontal axis show the predicted classes.

and our model, out-perform the object-centric AlexNet and those hand-craft GIST and HMP based on unsupervised feature learning, which proves the advantage of scene specific modeling. The other notable phenomenon is that the performance of the classifiers trained by combining the rgb and depth features always out-perform the rgb counterparts, which confirms the common intuition that there are some consistency and complementary information between rgb and depth modalities.

Moreover, the SVM trained with our projected features, outperform the PlaceCNN RBF baseline by around 2.5% in terms of accuracy, which means the features learned by our method has preserved most of the discriminative feature from the original RGB and depth CNN, while eliminating a lot of noisy and redundant features, which proves the validity of feature fusion. Our method's slightly outperformed SSCNN by 0.2%. Our projected color features out-performed the SSCNN's color version by 0.9%. Actually, our method and SSCNN present different tools toward RGB-D scene recognition problem. SSCNN goes along feature engineering, while we go along feature fusion, and these two methods are both effective and they could be combined together, which is left for future study.

In terms of learned feature dimension, SSCNN is the most compact, which has only 512 dimensions. The dimen-

sion of our learned feature (4096 dimensions) lies in between SSCNN and HMP (28000 dimensions). But regarding the generalization capability, our method out-perform SSCNN as our framework only needs image-wise ground-truth, while SSCNN need to prepare pixel-wise ground-truth which is restrictive when applied to a new task. Moreover, SSCNN requires image segmentation, which would introduce extra computational overhead.

4. Conclusion

In this work, we propose a novel discriminative feature fusion framework for RGBD scene recognition. Our framework considers the inter- and intra-modalities correlation for all class samples and meanwhile regularize the learned feature to be discriminative and compact. Our method outperforms other state-of-the-arts over the recently proposed SUN-RGBD dataset in terms of the accuracy, feature length and learning overhead. Overall, our work shows the potential of feature fusion for RGBD scene recognition, which is worthwhile for further research.

References

[1] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS*, 2011.

Class	bathroom	bedroom	classroom	computer room	conference room	corridor
Accuracy (%)	89.1	69.0	54.1	41.5	20.2	69.0
Class	dining area	dining room	discussion area	furniture store	home office	kitchen
Accuracy (%)	33.7	41.3	6.7	87.4	11.5	66.5
Class	lab	lecture theatre	library	living room	office	rest space
Accuracy (%)	12.4	25.6	28.8	22.4	43.1	59.3
Class			study space	Average Accuracy		
Accuracy (%)			7.5	41.5		

Table 3. Scene recognition accuracy for different classes.

Model	Input	Accuracy (%)
GIST [12] +	RGB	19.7
RBF Kernel SVM	RGB+D	23.0
HMP [1] +	RGB	21.7
Linear SVM	RGB+D	25.7
AlexNet [7]	RGB	24.3
	RGB+D	30.7
Place-CNN [22] +	RGB	35.6
Linear SVM	RGB+D	37.2
Place-CNN [22] +	RGB	38.1
RBF Kernel SVM	RGB+D	39.0
SSCNN [10]	RGB	36.1
	RGBD	41.3
Our Model +	RGB	37.0
Linear SVM	RGB+D	**41.5**

Table 2. Scene recognition accuracy of different methods.

[2] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *CVPR*, 2013.

[3] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] S. Gupta, R. B. Girshick, P. A. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014.

[5] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[6] 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 *ICCV Workshop*, 2011.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[8] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[9] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[10] 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. *CoRR*, abs/1509.06470, 2015.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[15] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *AISTATS*, 2009.

[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[17] S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015.

[18] A. Wang, J. Cai, J. Lu, and T. C. and. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *ICCV*, 2015.

[19] A. Wang, J. Lu, J. Cai, T. Cham, and G. Wang. Large-margin multi-modal deep learning for RGB-D object recognition. *TMM*, 17(11):1887–1898, 2015.

[20] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *TPAMI*, 33(8):1489–1501, 2011.

[21] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[22] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.