Modality and Component Aware Feature Fusion for RGB-D Scene Classification

Anran Wang[1], Jianfei Cai[1], Jiwen Lu[2], and Tat-Jen Cham[1]
[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] Department of Automation, Tsinghua University, Beijing, China

Abstract

While convolutional neural networks *(CNN) have been excellent for object recognition, the greater spatial variability in scene images typically meant that the standard full-image CNN features are suboptimal for scene classification. In this paper, we investigate a framework allowing greater spatial flexibility, in which the* Fisher vector *(FV) encoded distribution of local CNN features, obtained from a multitude of region proposals per image, is considered instead. The CNN features are computed from an augmented pixel-wise representation comprising multiple modalities of RGB, HHA and surface normals, as extracted from RGB-D data. More significantly, we make two postulates: (1) component sparsity — that only a small variety of region proposals and their corresponding FV GMM components contribute to scene discriminability, and (2) modal non-sparsity — within these discriminative components, all modalities have important contribution. In our framework, these are implemented through regularization terms applying group lasso to GMM components and exclusive group lasso across modalities. By learning and combining regressors for both proposal-based FV features and global CNN features, we were able to achieve state-of-the-art scene classification performance on the SUNRGBD Dataset and NYU Depth Dataset V2.*

1. Introduction

Scene classification is a challenging problem, especially for indoor scenes, due to the large intra-class variation due to vast differences in spatial layouts within each scene class. After a substantial leap in object recognition [13] performance using *convolutional neural networks* (CNN) trained on large-scale object-centric datasets such as ImageNet [4], a scene-centric dataset known as *Places* [36] was introduced for investigating the utility of CNN in scene classification tasks. Although there was a reported performance improvement using a scene-centric CNN, it became obvious that global CNN features extracted from full images were too spatially rigid to be optimal for scene classification.

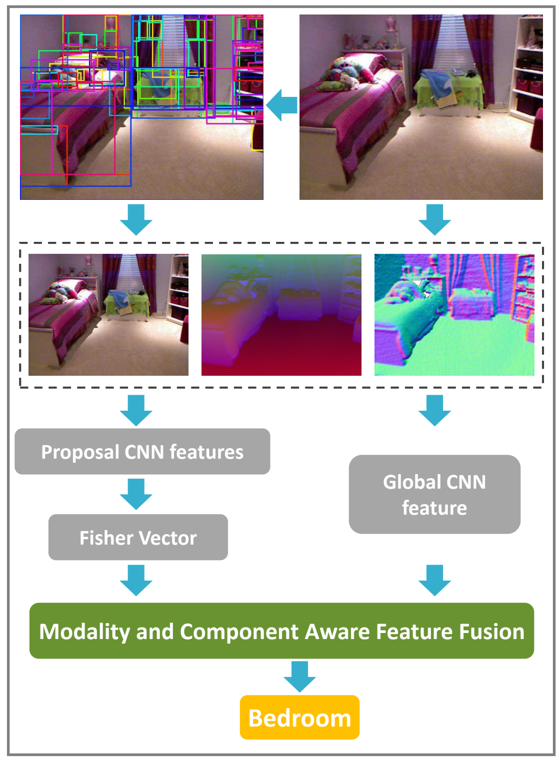

Figure 1. Our framework: we first extract proposals from each RGB-D image. Then for all the proposals and the full image, we derive CNN features from different modalities: RGB, HHA and surface normal (SN). For each image, the proposal based CNN features for each modality are encoded by Fisher Vector (FV), and the resulted multi-modal FV features are regarded as the input to our modality and component aware feature fusion. Finally we combine the regression results of the proposal based FV features and the full-image based CNN features to get the final classification result.

Several methods [5, 33, 38] have been proposed for classifying RGB scene images using local instead of global information. They share a similar pipeline: CNN features were densely extracted at different locations and scales of an image, encoded as a combined feature representation (e.g. via *Fisher vectors* (FV) [18, 20]) and then classified using support vector machines (SVM). Results show that the local features are competitive when compared to full-image based CNN features and provide important complementary information. However, only a small subset of local features are likely to be discriminative in a scene classification task. In many existing works, a task-independent feature representation is used, such as a comprehensive *Gaussian mixture model* (GMM) which models all features for encoding Fisher vectors; this tends to result in overfitting when training regressors.

There are also a few methods proposed for scene classification on RGB-D images [7, 1, 23, 14]. Most of these directly concatenate features from color and depth modalities together prior to classification. Such a direct combination does not adequately exploit the relationship between the different modalities of color and depth.

In our work, we start with the standard pipeline for local feature extraction and feature encoding. In particular, we use an existing object proposal extractor to generate region proposals from each RGB-D image, representing each proposal by the corresponding local CNN features obtained from different modalities. Similar to [8], we extract CNN features from RGB and HHA (horizontal disparity, height above ground, and angle between the local surface normal and direction of inferred gravity). In order to more explicitly capture geometric information, we also extract CNN features from an additional modality of *surface normals* (S-N). For each modality, we use FV to encode the region-proposal-based CNN features.

To address the two issues mentioned previously, we make two important postulates: (1) component sparsity — that we should *not* attempt to utilize all features in the FV GMM components, but rather seek out only a few key components that maximally contribute to scene discriminability. (2) modal non-sparsity — that for these key discriminative components, all modalities will significantly contribute to the discriminability because they provide important complementary information.

To this end, we propose a modality and component aware feature fusion framework for RGB-D scene classification on the extracted multi-modal FV features. In the feature fusion step, we incorporate different levels of structure sparsity regularization that effectively extract discriminative features from different modalities and different GMM components in FV. In order to only consider GMM components in the FV which are discriminative, we first enforce inter-component sparsity to discount unnecessary components. Second, we propose to enhance intra-modal component sparsity with inter-modal non-sparsity. In this way, we encourage discriminative features in different modalities to co-exist. Finally, by learning and combining regressors for both proposal-based FV features and full-image CNN features, we were able to achieve state-of-the-art performance on the SUNRGBD Dataset [23] and NYU Depth Dataset V2 [17].

Fig. 1 shows an overview of the proposed framework.

2. Related Work

Scene Classification: Object recognition performance has recently been boosted through the use of well designed CNN techniques [13] in conjunction with extensive labeled data. To adapt the current CNN techniques for scene classification, Zhou *et al.* [36] introduced a large scene-centric dataset called Places and showed significant performance improvement on scene classification using a CNN trained on this dataset, as compared to directly applying the CNN pretrained on the object-centric dataset ImageNet [4]. Although a scene-centric dataset more appropriately captures the richness and diversity of scene imagery, the typical way of extracting global CNN features from full images may not adequate handle the geometric variability of complex indoor scenes.

Several methods have been proposed to leverage local C-NN features to enhance discriminative capability. Gong *et al.* [5] proposed densely extracting multi-scale CNN activations, aggregating the activations of each scale via *vector of locally aggregated descriptors* (VLAD) [11], and concatenating the multi-scale VLAD features together as the final feature representation. Yoo *et al.* [33] presented a similar framework, except they used *Fisher Vectors* (FV) as the encoding method. In another work [38], Zuo *et al.* showed the importance of the complementary information provided by local features, where they derived local features by learning a discriminative and shareable feature transformation filter bank for local image patches. Among all these methods, few of them take direct care to exclude non-discriminative local features that can lead to overfitting.

There are also several other works that are not developed for scene classification, but related to our method. In particular, Yang *et al.* [30] approached the multi-label image classification problem through multi-view learning, where they derived a feature view by extracting CNN features from object proposals followed by the FV encoding, and constructed a label view using strong labels. Zhang *et al.* [35] dealt with fine-grained image categorization, where they proposed to use feature selection to remove noisy features in FV. Their feature selection is based on the relevances of individual features to class labels, which are calculated independently in different feature dimensions.

With increasing spread of commodity depth cameras that provide depth images along with color images, more RGB-D data are becoming available. A number of methods operating on RGB-D data have been proposed for scene labeling, object recognition and scene classification [19, 21, 2, 8, 16, 25]. There is a recent work in which CNN is used as the feature extraction method for RGB-D data [8], where Gupta *et al.* proposed to encode depth with three channels (HHA). This makes it possible to directly apply the CNN model pre-trained on RGB images, which also have three channels, to HHA to extract CNN features for depth.

On the topic of RGB-D scene classification, Gupta *et al.* [7, 6] described a method to detect contours in RGB-D images and use them for semantic segmentation, further treating the quantized semantic segmentation output as local features for scene classification. Banica *et al.* [1] proposed to apply second-order pooling [3] of hand-crafted features mainly for semantic segmentation as well as on scene classification. Song *et al.* [23] introduced a large scale RGB-D dataset called the SUNRGBD Dataset with ground truth and baselines for different scene understanding tasks. For scene classification, they directly used pre-trained CNN in [36] to extract CNN features from RGB and HHA. Liao *et al.* [14] proposed to include a regularization on semantic segmentation to improve scene classification performance, where their cost function to train CNN contains both the loss of scene classification and the loss of semantic segmentation.

Structure Sparsity: Structure sparsity is an extension of the standard sparsity concept, which aims to facilitate arbitrary structures on the feature set [9]. The effectiveness of structure sparsity for feature learning has been widely proven in different applications such as face recognition [28], web page recognition [37], image super resolution [31], action recognition [32], and object recognition [15].

Here we discuss several representative pieces of research that are relevant to our method. In particular, Tibshirani [24] proposed the idea of "lasso" which minimizes the squared errors with an l_1-norm regularization term. It essentially shrinks some coefficients and sets others to 0. The relationship between the loss function and the regularization term is analyzed in [24].

Based on lasso, Yuan and Lin [34] further extended it for variable selection with predefined groups, which is usually called "group lasso". Their key assumption is that if a few features in a group are important, then the whole group is regarded as important. For tasks benefiting from the selection of important groups, their method improves the performance of the traditional lasso. Zhou *et al.* [37] further developed a new form of regularization called "exclusive lasso", where they focused on multi-task feature selection. Their assumption is that features that are important for one category become less likely to be important for other categories, and thus their idea is to introduce the competition among different tasks for the same feature. Kong *et al.* [12] shared a similar idea with Zhou *et al.* [37], but they focused on feature selection with multi-group of features. They proposed "exclusive group lasso" to encourage features in different groups to co-exist, which is different from group lasso that enforces inter-group sparsity. Combining with the traditional lasso, exclusive group lasso demonstrates its effectiveness on the spoken letter classification task [37]. In this research, we combine both group lasso and exclusive group lasso in our feature fusion framework to solve the scene classification problem.

3. Multi-modal Proposal-based Global Feature Representation

In our framework, local information is incorporated through the use of region proposals and their corresponding local CNN features. More specifically, we use the publicly available proposal extractor [8] to extract region proposals from each RGB-D image. For each proposal, the local CNN features are then computed from both color and geometry data. In addition to the two modalities (RGB and HHA) used in [8], we further include a third modality of *surface normals* (SN) into our framework, represented as unit 3D vectors. Since all three modalities comprise three channels each, we start with the same 8-layer CNN model pretrained on the Places Dataset [36] for each modality, but then fine-tuned independently. We use the activations of the first fully connected layer (full6, i.e. layer 6 in the 8-layer CNN) in each modality as the CNN features for each proposal. In order to reduce computational complexity, the number of dimensions in the CNN full6 activation vectors is reduced from 4096 to $d = 400$ per modality via PCA. In this way, given an RGB-D image with J extracted object proposals, each proposal in each modality is represented by its corresponding CNN feature vector $\boldsymbol{f}_{ij} \in \mathbb{R}^d$.

The CNN features for all proposals within a single RGB-D image is then encoded with the standard Fisher Vector (FV) [18, 20] approach. The FV encoding consists of a K-component Gaussian Mixture Model (GMM) with parameters of $\lambda = \{w_k, \boldsymbol{\mu}_k, \Sigma_k, k = 1 \dots K\}$, where w_k, $\boldsymbol{\mu}_k$ and Σ_k is respectively the mixing weight, mean and covariance matrix (assumed diagonal) of the k-th Gaussian component. The gradient vectors (w.r.t. mean $\boldsymbol{\mu}_k$ and s.d. $\boldsymbol{\sigma}_k$) are:

$$\begin{aligned} \boldsymbol{g}^i_{\boldsymbol{\mu}_k} &= \frac{1}{\sqrt{w_k}} \sum_{j=1}^{J} \gamma^i_j(k) \left(\frac{\boldsymbol{f}_{ij} - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k} \right) \\ \boldsymbol{g}^i_{\boldsymbol{\sigma}_k} &= \frac{1}{\sqrt{2w_k}} \sum_{j=1}^{J} \gamma^i_j(k) \left(\frac{(\boldsymbol{f}_{ij} - \boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}^2_k} - 1 \right) \end{aligned} \tag{1}$$

where $\gamma^i_j(k)$ is the soft assignment weight of $\boldsymbol{f}_{ij}$ to the k-th component:

$$\gamma^i_j(k) = P(k \,|\, \boldsymbol{f}_{ij}, \lambda) \,. \tag{2}$$

Concatenating the two gradient vectors leads to a $2Kd$-dimensional FV for each modality. By further collating FVs from the three modalities, we obtain a multi-modal feature representation for image i, given by $\boldsymbol{x}_i \in \mathbb{R}^D$, where $D = 6Kd$.

4. Modality and Component Aware Feature Fusion

4.1. Formulation

Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ denote the multi-modal FVs derived from N input RGB-D images, $Y \in \mathbb{R}^{N \times C}$ be the ground truth label matrix with C classes, and $W \in \mathbb{R}^{D \times C}$ be the transformation or weight matrix that maps input features X into the label domain via $X^T W$.

We formulate our method as solving a regression problem with several regularization terms:

$$\begin{aligned} &\min_W F = R + R_1 + R_2 + R_3 \\ &= \frac{1}{2} \left\| X^T W - Y \right\|_F^2 \\ &+ \lambda_1 \left\| W_{(P)} \right\|_2^1 + \lambda_2 \left\| W_{(Q)} \right\|_1^2 + \lambda_3 \| W \|_1 \end{aligned} \tag{3}$$

The first term R is the standard least-squares regression term. It encourages the transformation $X^T W$ to closely reconstruct the labels, biasing towards a W that extracts discriminative information from the features. R_3 is the common l_1-norm regularization term to invoke only a sparse set of feature dimensions, while R_1 and R_2 are explained in detail below. The tradeoffs parameters are λ_1, λ_2 and λ_3.

Component Regularization Term R_1: Since the GMM components of the FV encoding are constructed from all region proposals, which are obtained in a generic fashion, many of these components do not contribute discriminative power for distinguishing between scene classes. Thus, we propose a regularization term based on group lasso [34] which should result in only the expected few discriminative components being associated with large weights, while the remaining components will be associated with zero or small weights. Specifically we define

$$R_1 = \left\| W_{(P)} \right\|_2^1 = \sum_{j=1}^{C} \sum_{p=1}^{P} \left\| (W_{(P)})_p^j \right\|_2 \tag{4}$$

where $(W_{(P)})_p^j \in \mathbb{R}^{2d}$ denotes the weights for the p-th component of the j-th class. There are K components for each modality, resulting in $P = Q \times K$ components in total, where Q is the number of modalities ($Q = 3$). Eq. (4) essentially applies l_2-norm regulation within each component (because the parameters of a component should have similar importance) and l_1-norm regulation across different components. Fig. 2 illustrates the idea of the component-based regularization, where a component is encouraged to have either all zero weights or multiple non-zero weights.

Modality Regularization Term R_2: Although it may be that the discriminative power of different modalities are different, it is expected that for the sparse set of discriminative features, their discriminability comes from a mixture of modalities, rather than due to a single modality in isolation

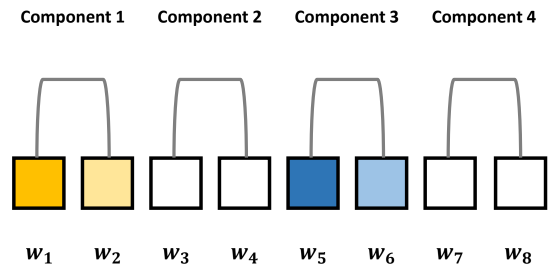

Figure 2. Illustration of the component regularization term, where each component is treated as a group and each group is encouraged to have either all zero weights (white squares) or multiple nonzero weights (colorful squares).

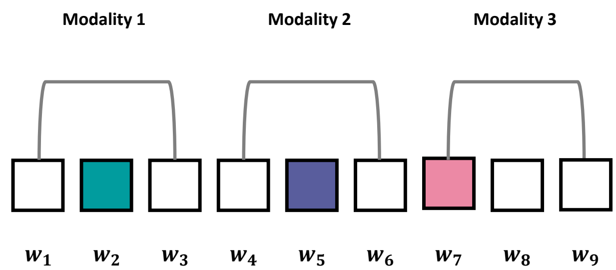

Figure 3. Illustration of the modality regularization term, where each modality is treated as a group and each group is encouraged to have sparse nonzero weights (colorful squares) and many zero weights (white squares).

(i.e. scene classification will not be optimally performed using only data from one modality). Thus, we propose to use the regularization term of exclusive group lasso [37] to encourage discriminative features from different modalities to co-exist, while features within one modality are encouraged to compete with each other. Fig. 3 illustrates the idea of modality regularization, where each modality is encouraged to be associated with sparse non-zero weights within itself, but not so across different modalities. We define the modality regularization term as:

$$R_2 = \left\| W_{(Q)} \right\|_1^2 = \sum_{j=1}^{C} \sum_{q=1}^{Q} \left(\left\| (W_{(Q)})_q^j \right\|_1 \right)^2 \tag{5}$$

where $(W_{(Q)})_i^j \in \mathbb{R}^{2Kd}$ denotes the weights for the i-th modality of the j-th class. Eq. (5) essentially applies l_1-norm regulation within each modality to encourage sparsity and l_2-norm like regulation across different modalities to encourage balance.

4.2. Optimization

To optimize the transformation matrix W in (3), we compute the derivative of the overall cost function w.r.t. $W_j \in \mathbb{R}^D$ for class j, based on existing solutions developed for the lasso, group lasso and exclusive lasso techniques.

Algorithm 1: The optimization pipeline

Input: X: multi-modal FV features;
Y: ground-truth label matrix.
Output: W: transformation matrix.

Step 1 (Initialization):
Initialize W as zero matrix.
Step 2 (Optimization):
For each class j
While not converged do
2.1. Fixing W_j, update $D_j^{(1)}$, $D_j^{(2)}$ and $D_j^{(3)}$ according to (7).
2.2. Fixing $D_j^{(1)}$, $D_j^{(2)}$ and $D_j^{(3)}$, update W_j according to (8).
end while until convergence
end for

Specifically:

$$\frac{\partial F}{\partial W_j} = XX^T W_j - X\boldsymbol{y}_j + \lambda_1 D_j^{(1)} W_j + 2\lambda_2 D_j^{(2)} W_j + \lambda_3 D_j^{(3)} W_j \tag{6}$$

where $\boldsymbol{y}_j \in \mathbb{R}^N$ denotes the label vector for all training images in class j, while $D_j^{(1)}$, $D_j^{(2)}$ and $D_j^{(3)}$ are all diagonal $D \times D$ matrices dependent on W_j. The i-th diagonal elements of $D_j^{(1)}$, $D_j^{(2)}$ and $D_j^{(3)}$ are calculated as

$$\begin{aligned} D_{ij}^{(1)} &= \frac{1}{\left\|(W_{(P)})_p^j\right\|_2} \\ D_{ij}^{(2)} &= \frac{\left\|(W_{(Q)})_q^j\right\|_1}{|W_{ij}|} \\ D_{ij}^{(3)} &= \frac{1}{2|W_{ij}|} \end{aligned} \tag{7}$$

Detailed derivations can be found in [26, 27, 12].

Once the derivative $\frac{\partial F}{\partial W_j}$ is available, W_j is updated as

$$W_j \leftarrow W_j - \gamma \frac{\partial F}{\partial W_j} \tag{8}$$

where γ is the learning rate. As $D_j^{(1)}$, $D_j^{(2)}$ and $D_j^{(3)}$ depend on W_j, we update D_j and W_j in an iterative way. The optimization pipeline is shown in Algorithm 1.

Using this optimization procedure, we learn an optimal transformation matrix W. In the testing stage, once the multi-modal FV features X of a test RGB-D image have been extracted, the regression values are computed simply using $X^T W$, with the maximum regression value regarded as the classification result.

To further leverage global features, we also adapt the proposed feature fusion framework to the multi-modal C-NN features applied on full images. Compared with the proposal-based feature fusion framework, the only difference is that the full-image based framework does not have components, because it is a single measurement rather than modeled as a distribution. In other words, the cost function of the full-image based framework only contains R, R_2 and R_3 terms of (3). Finally the regression values from both the proposal-based and the full-image based frameworks are added to obtain the final classification.

5. Experiments

To evaluate the effectiveness of our proposed modality and component aware feature fusion framework, we perform scene classification experiments on the SUNRGBD Dataset [23] and the NYU Depth Dataset V2 [22]. The details of the experiments and the results are described in the following sections.

5.1. Datasets and Experimental Setup

SUNRGBD Dataset: This dataset has 19 scene categories. It consists of 10,335 RGB-D scene images, including 3,784 Kinect v2 images, 1,159 Intel RealSense images captured by Song *et al.* [23], 1,449 Kinect v1 images taken from the NYU Depth Dataset V2 [22], 554 Kinect v1 images selected from the Berkeley B3DO Dataset [10], and 3,389 Asus Xtion images selected from SUN3D videos [29]. We follow the experiment settings stated in [23] and only keep categories with more than 80 images. Using the publicly available split, there are in total 4,845 images for training and 4,659 images for testing.

NYU Depth Dataset V2: This dataset consists of 1,449 images. It has 27 scene categories but only a few of them are well represented. Following the procedures stated in [7], the original 27 categories are reorganized into 10 scene categories, including the 9 most common categories and an 'other' category for images in the remaining categories. We use the publicly available split, which has 795 images for training and 654 images for testing.

Metrics: For both datasets, we report the means of diagonal values of the confusion matrices, which are the average precisions over all scene classes. Another metric we considered is the overall accuracy, which is the precision over all test images. Since we found these two metrics to be strongly correlated, only the former is listed for presentation conciseness.

Fine-tuning: Our starting point is the current state-of-the-art CNN model (Places-CNN) for scene classification, pre-trained on the Places Dataset [36] (2.5 million RGB images with 205 scene categories). To better adapt the pre-trained CNN network for RGB-D data, especially for the HHA and surface normal modalities, we fine-tuned the

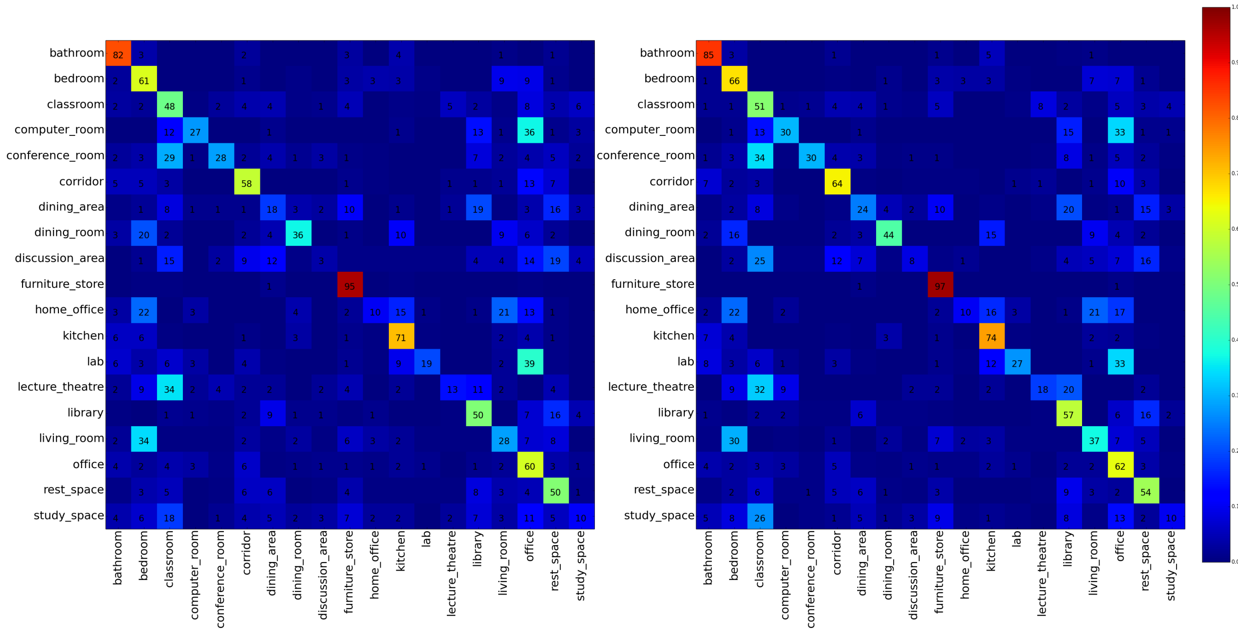

Figure 4. Confusion matrices of 'FV (L1)' (left) and 'FV (Modality+Component+L1)' (right) on SUNRGBD Dataset. It shows that by adding the modality and component regularization terms, the performance is improved for almost all the classes.

Places-CNN with our relevant data. For the SUNRGBD Dataset, we fine-tuned the Places-CNN with each of the three modalities (RGB, HHA and surface normals) from training images, utilizing image-level labels. For the NYU Depth Dataset V2, the fine-tuning was carried out in two stages: first with images from the SUNRGBD Dataset (but excluding the NYU V2 images), then using training images from the NYU Depth Dataset V2.

After one of the fine-tuned CNNs has operated on a region proposal in an image, the CNN activation vector of the first fully connected layer (full6) is extracted. As stated previously, based on a collection of such vectors in training images, PCA is then used to reduce these 4096-dimensional vectors to 400 dimensions, and further encoded as GMM-based Fisher vectors.

Parameters: The parameter settings for the two datasets are identical. The number of GMM components K for each modality is 64. The PCA-reduced dimensionality of the CNN activation vector is $d = 400$. The parameters λ_1, λ_2 and λ_3 in (3) for proposal-based feature fusion are set at 0.005, 0.01 and 0.001 respectively with standard 5-fold crossvalidation. For full-image-based feature fusion ($\lambda_1 = 0$), we empirically set λ_2 and λ_3 to be 0.001 and 0.0001 respectively. The learning rates γ in (8) are set at 10^{-4} and 10^{-8} for proposal-based and full-image-based feature fusion respectively. When optimizing W_j for each class, the number of iterations is fixed at 100.

Table 1. Comparing the classification results of the proposal based FV features and the full-image based CNN features under different modalities with linear SVM classifier on SUNRGBD Dataset.

Accuracy (%)	Full (SVM)	FV (SVM)	FV+Full (SVM)
RGB	40.4	36.2	-
HHA	36.3	34.6	-
SN	34.3	30.6	-
RGB+HHA	44.9	39.7	-
RGB+HHA+SN	45.7	41.2	45.9

5.2. Results on SUNRGBD Dataset

We first compare the linear SVM classification results of the proposal-based FV features and the full-image-based CNN features obtained from different combinations of modalities, without including our proposed regularization terms. We considered three baselines: 1) 'Full (SVM)': the full-image-based CNN features with SVM; 2) 'FV (SVM)': the proposal-based FV features with SVM; and 3) 'FV+Full (SVM)': concatenating the full-image features and the FV features prior to linear SVM classification. Table 1 shows the comparison results. Among the three individual modalities, RGB features achieve the best performance; however, it is clear that combination of the three modalities substantially improves performance. The com-

Table 2. Comparison of different baselines of our proposed feature fusion framework on SUNRGBD Dataset.

Method	Accuracy (%)
FV (SVM)	41.2
FV (L1)	41.0
FV (Modality + L1)	43.9
FV (Component + L1)	42.7
FV (Modality+Component+L1)	45.1
Full (SVM)	45.7
Full (L1)	44.9
Full (Modality + L1)	45.4
Combine FV and Full	48.1

Table 3. Comparison with state-of-the-art methods on SUNRGBD Dataset.

Method	Accuracy (%)
Song *et al.* [23]	39.0
Liao *et al.* [14]	41.3
Ours	48.1

parisons between 'RGB+HHA' and 'RGB+HHA+SN' indicate that expressing surface normals (SN) as explicitly separate from HHA leads to improved performance, although both are indirectly extracted from depth images. More importantly, we can see that 'FV (SVM)' performs poorly compared with 'Full (SVM)', despite 'FV (SVM)' features having 51,200 dimensions ($D = 6Kd = 6 \times 64 \times 400$) while 'Full (SVM)' features only have 4096×3 dimensions. Even the combination of 'FV+Full (SVM)' only slightly improves the performance. This is mainly due to many dimensions of the FV features not having discriminative power but which cause regressor overfitting, unless better regularization is used (as implemented in our proposed feature fusion framework).

Table 2 shows the impact of our modality and component aware feature fusion frameworks with the added regularization terms. Here we consider seven other settings: 1) 'FV (L1)': using our framework only on proposal-based CNN features and with only the R_3 (L1-norm) regularization term active; 2) 'FV (Modality+L1)': proposal-based features only with R_2 and R_3 active; 3) 'FV (Component+L1)': proposal-based features only with R_1 and R_3 active; 4) 'FV (Modality+Component+L1)': proposal-based features only with R_1, R_2 and R_3 active; 5) 'Full (L1)': using our framework only on full-image based CNN features, with only R_3 active; 6) 'Full (Modality+L1)': full-image features only with R_2 and R_3 active; 7) 'Combine FV and Full': combined regression using both 'FV (Modality+Component+L1)' and 'Full (Modality+L1)', which is our final result.

Table 4. Comparing the classification results of the proposal based FV features and the full-image based CNN features under different modalities with linear SVM classifier on NYU Depth Dataset V2.

Accuracy (%)	Full (SVM)	FV (SVM)	FV+Full (SVM)
RGB	53.5	49.2	-
HHA	51.5	52.2	-
SN	51.7	44.8	-
RGB+HHA+SN	58.5	55.8	58.7

From Table 2, we can see that our feature fusion framework is very effective for the multi-modal FV features, greatly improving the performance from 41.0% under the setting of 'FV (L1)' to 45.1% under the setting of 'FV (Modality+Component+L1)'. It demonstrates that the discriminative information of the high-dimensional multimodal FV features can be better extracted with the developed structure sparsity regularization. Although 'Full (Modality+L1)' does not outperform 'Full (SVM)' (mainly because the full-image based CNN features are not of high dimensions), the combination of the regression results of the FV features and the full-image based features ('Combine FV and Full') achieves the best performance. This suggests that the proposal-based features contain pertinent local information not represented in full-image-based features.

Table 3 shows comparison with state-of-the-art methods. We compared with: 1) Song *et al.* [23], which directly uses pre-trained Places-CNN to extract features from RGB and HHA followed by RBF kernel SVM for classification; and 2) Liao *et al.* [14], which incorporates features extracted from semantic segmentation to improve scene classification. It can be seen that our proposed method significantly outperforms the two state-of-the-art methods.

In Fig. 4, we visualize the confusion matrix to give the performance comparison between 'FV (L1)' and 'FV (Modality+Component+L1)'. It can be seen that there is a performance improvement for almost every class. We can also spot some misclassification cases, *e.g.* many 'lab' images are misclassified as 'office', and some 'lecture_theatre' images are misclassified as 'classroom'. These are due to both visual and semantic similarity between such classes.

5.3. Results on NYU Depth Dataset V2

We also obtained results on NYU Depth Dataset V2, where we can make similar observations to those for the SUNRGBD Dataset. Table 4 compares the classification results of the proposal-based FV features and the full-image-based CNN features with linear SVM classifier. Table 5 compares the results under different baseline settings of our modality and component aware feature fusion framework. In this dataset, we can also see that the 'FV (Modality+Component+L1)' baseline significantly outper-

Table 5. Comparison of different baselines of our proposed feature fusion framework on NYU Depth Dataset V2.

Method	Accuracy (%)
FV (SVM)	55.8
FV (L1)	53.5
FV (Modality + L1)	56.7
FV (Component + L1)	55.5
FV (Modality+Component+L1)	59.8
Full (SVM)	58.5
Full (L1)	58.8
Full (Modality + L1)	59.1
Combine FV and Full	63.9

Table 6. Comparison with state-of-the-art methods on NYU Depth Dataset V2. We reimplemented the second-order pooling method [1] and show our reproduced results as 'O2P'.

Method	Accuracy (%)
Gupta *et al.* [6]	45.4
SPM on SIFT [6]	38.9
SPM on G. Textons [6]	33.8
SPM on SIFT+G. Textons [6]	44.9
O2P on color SIFT and LBP	41.0
O2P on depth SIFT and LBP	48.5
O2P on color+depth	50.9
Ours	63.9

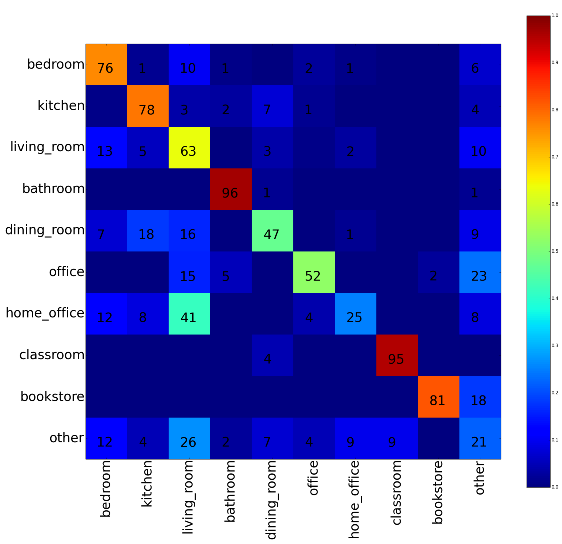

Figure 5. Confusion matrix of our final results ('Combine FV and Full') on NYU Depth Dataset V2.

forms 'FV (L1)' by over 6%, which further proves the effectiveness of the structured sparsity promoted by our feature fusion method. By combining the regression results of 'FV (Modality+Component+L1)' and 'Full (Modality+L1)', we further improve the performance significantly.

Table 6 shows comparisons with state-of-the-art methods. Gupta *et al.* [7, 6] used the semantic segmentation output (i.e. the probabilities of belonging to different semantic classes) as local features and applied spatial pyramid (SPM) on them. We also show the results of the three baselines in [6]: 1) histograms of vector quantized color SIFT as features with SPM; 2) histograms of geocentric textons with SPM; 3) combination of 1) and 2) with SPM.

Recently, Banica *et al.* [1] made use of second-order pooling (O2P) [3] of hand-crafted features mainly for the RGB-D semantic segmentation problem, but they also directly apply O2P features for scene classification as an additional application. For RGB-D scene classification on NYU Depth Dataset V2, they reported a very high classification results of 83.81%. Despite our careful reimplementation of their method in detailed consultation with one of the authors, we were unable to reproduce and verify their published percentages; hence we only list the results obtained from our implementation of the O2P method in Table 6[1]. Specifically, we conduct second-order pooling on SIFT and Local Binary Patterns (LBP) for both color and depth images. The pooling was done in subregions of a 1, 2×2 and 4×4 SPM. Fig. 5 shows the confusion matrix of our final results ('Combine FV and Full'). We can see that the results of 'home_office' and 'other' classes were not as good as other classes, since 'others' is not well defined, while 'home_office' is significantly confused with 'living_room'.

6. Conclusion

In this paper, we proposed a modality and component aware feature fusion framework that effectively makes use of high-dimensional FV features from RGB, HHA and surface normal modalities. We formulate our method as a regression problem with regularization terms corresponding to modality and component related structure sparsity. By combining the regression results of the proposal based multi-modal FV features and the full-image based multi-modal CNN features, we achieved state-of-the-art scene classification performance on the SUNRGBD Dataset and the NYU Depth Dataset V2.

Acknowledgment

This research, which is carried out at BeingThere Centre, is supported by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. The research is also in part supported by MOE Tier 1 RG 138/14.

[1] We attempted a careful reimplementation in consultation with an author of [1] based on codes of [3], who was also unable to figure out the reason for the discrepancy in results.

References

[1] D. Banica and C. Sminchisescu. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *CVPR*, pages 3517–3526, 2015. 2, 3, 8

[2] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS*, pages 2115–2123, 2011. 3

[3] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, pages 430–443. 2012. 3, 8

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 2

[5] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, pages 392–407. 2014. 1, 2

[6] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *IJCV*, 112(2):133–149, 2014. 3, 8

[7] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, pages 564–571, 2013. 2, 3, 5, 8

[8] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. 2014. 2, 3

[9] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *The Journal of Machine Learning Research*, 12:3371–3412, 2011. 3

[10] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013. 5

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010. 2

[12] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding. Exclusive feature learning on arbitrary structures via $l_{1,2}$-norm. In *NIPS*, pages 1655–1663, 2014. 3, 5

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1, 2

[14] 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. *arXiv preprint arXiv:1509.06470*, 2015. 2, 3, 7

[15] N. Naikal, A. Y. Yang, and S. S. Sastry. Informative feature selection for object recognition via sparse pca. In *ICCV*, pages 818–825, 2011. 3

[16] D. S. Nathan Silberman and R. Fergus. Instance segmentation of indoor scenes using a coverage loss. In *ECCV*, pages 616–631, 2014. 3

[17] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[18] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156. 2010. 2, 3

[19] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766, 2012. 3

[20] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013. 2, 3

[21] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *ICCV Workshops*, pages 601–608, 2011. 3

[22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 5

[23] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 2, 3, 5, 7

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 3

[25] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multi-modal unsupervised feature learning for rgb-d scene labeling. In *ECCV*, pages 453–467. 2014. 3

[26] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, and L. Shen. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *ICCV*, pages 557–562, 2011. 5

[27] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012. 5

[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009. 3

[29] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632, 2013. 5

[30] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai. Can partial strong labels boost multi-label object recognition? *arXiv preprint arXiv:1504.05843*, 2015. 2

[31] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *TIP*, 19(11):2861–2873, 2010. 3

[32] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, pages 1331–1338, 2011. 3

[33] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon. Fisher kernel for deep neural activations. *arXiv preprint arXiv:1412.1628*, 2014. 1, 2

[34] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 3, 4

[35] Y. Zhang, X.-s. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do. Weakly supervised fine-grained image categorization. *arXiv preprint arXiv:1504.04943*, 2015. 2

[36] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014. 1, 2, 3, 5

[37] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010. 3, 4

[38] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang. Learning discriminative and shareable features for scene classification. In *ECCV*, pages 552–568. 2014. 1, 2