A Discriminative Representation of Convolutional Features for Indoor Scene Recognition

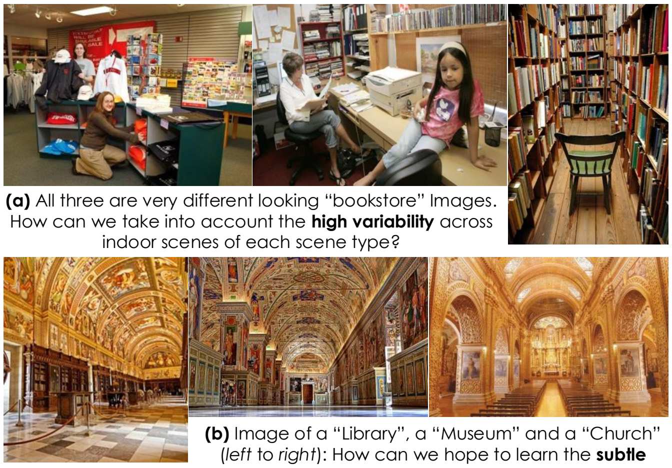
S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. Sohel

***Abstract*—Indoor scene recognition is a multi-faceted and challenging problem due to the diverse intra-class variations and the confusing inter-class similarities that characterize such scenes. This paper presents a novel approach that exploits rich mid-level convolutional features to categorize indoor scenes. Traditional convolutional features retain the global spatial structure, which is a desirable property for general object recognition. We, however, argue that the structure-preserving property of the CNN activations is not of substantial help in the presence of large variations in scene layouts, e.g., in indoor scenes. We propose to transform the structured convolutional activations to another highly discriminative feature space. The representation in the transformed space not only incorporates the discriminative aspects of the target dataset but also encodes the features in terms of the general object categories that are present in indoor scenes. To this end, we introduce a new large-scale dataset of 1300 object categories that are commonly present in indoor scenes. Our proposed approach achieves a significant performance boost over previous state-of-the-art approaches on five major scene classification datasets.**

***Index Terms*—Scene classification, convolutional neural networks, indoor object dataset, feature representations, dictionary learning, sparse coding**

I. Introduction

This paper proposes a novel method that captures the discriminative aspects of an indoor scene to correctly predict its semantic category (e.g., bedroom or kitchen). This categorization can greatly assist in context-aware object and action recognition, object localization, and robotic navigation and manipulation [49], [50]. However, due to the large variabilities between images of the same class and the confusing similarities between images of different classes, the automatic categorization of indoor scenes represents a very challenging problem [35], [50]. Consider, for example, the images shown in Fig. 1. The images in the top row (Fig. 1 a) belong to the same class *'bookstore'* and exhibit a large data variability in the form of object occlusions, cluttered regions, pose changes and varying appearances. The images in the bottom row (Fig. 1 b) are of three different classes and have large visual similarities. A high-performance classification system should therefore be able to address the inherently challenging nature of indoor scenes.


(a) All three are very different looking "bookstore" Images. How can we take into account the **high variability** across indoor scenes of each scene type?

(b) Image of a "Library", a "Museum" and a "Church" (*left to right*): How can we hope to learn the **subtle differences** between different scene types?

Fig. 1: **'Where am I located indoors?'** We want to answer this question by assigning a semantic class label to a given color image. An indoor scene categorization framework must consider high intra-class variability and should be able to address confusing inter-class similarities. This paper introduces a methodology to achieve these challenging requisites. (example images from MIT-67 dataset)

To address the challenges of indoor scenes, previous works [6], [19], [25], [29], [35] proposed to encode either local or global spatial and appearance information. In this paper, we argue that neither of those two representations provide the best answer to effectively addressing indoor scenes. The global representations are unable to model subtle details, and the low-level local representations cannot capture object-to-object relations and global structures [21], [35], [41]. We therefore devise mid-level representations that contain the necessary intermediate level of detail. These mid-level representations neither ignore the local cues nor lose the important scene structure and object category relationships.

Our proposed mid-level representations are derived from densely and uniformly extracted image patches. To encode the discriminative and contextual aspects at multiple scales, our approach uses data augmentation to generate mid-level patches at three different spatial resolutions (Sec. III-A). To extract a rich feature representation from these patches, we use deep Convolutional Neural Networks (CNNs). CNNs provide excellent generic mid-level feature representations and have recently shown great promise for large-scale classification and

S. H. Khan is with the Data61-CSIRO (Commonwealth Scientific and Industrial Research Organisation), 7 London Circuit, Canberra ACT 2601, and the University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia. E-mail: salman.khan@research.uwa.edu.au.

M. Hayat is with the University of Canberra, University Dr, Bruce ACT 2617, Australia. E-mail: munawar.hayat@canberra.edu.au

M. Bennamoun and R. Togneri are with the University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia. E-mail: mohammed.bennamoun, roberto.togneri@uwa.edu.au.

F. Sohel is with the Murdoch University, 90 South St, Murdoch, WA 6150, Australia. E-mail: f.sohel@murdoch.edu.au.

detection tasks [3], [10], [31], [37]. They however tend to preserve the global spatial structure of the images [53], which is not an optimal choice in the presence of large intra-class variations, e.g., in the case of indoor scene categorization (Fig. 1, Sec. IV). We therefore propose a method to discount this global spatial structure while simultaneously retaining the intermediate scene structure, which is necessary to model the mid-level scene elements. For this purpose, we encode the extracted mid-level representations in terms of their association with codebooks[1] of Scene Representative Patches (SRPs). This enhances the robustness of the convolutional feature representations while keeping intact their discriminative power.

It is interesting to note that some previous works hinted at the incorporation of 'wide context' [17], [21], [51] for scene categorization. Such high-level context-aware reasoning has been shown to improve the classification performance. However, in this work, we show that, for the case of highly variant indoor-scenes, mid-level context relationships prove to be the most decisive factor in classification. The intermediate level of the scene details facilitate the learning of subtle differences in the scene composition and its constituent objects. In contrast, global structure patterns can confuse the learning/classification algorithm due to the high inter-class similarities (Sec. IV-C).

As opposed to existing feature encoding schemes, we propose to form multiple codebooks of SRPs. We demonstrate that forming multiple smaller codebooks (instead of one large codebook) proves to be more efficient and produces a better performance (Sec. IV-D). Another key aspect of our feature encoding approach is the combination of supervised and unsupervised SRPs in our codebooks. The unsupervised SRPs are collected from the training data, and the supervised SRPs are extracted from a newly introduced dataset of 'Object Categories in Indoor Scenes' (OCIS). The supervised SRPs provide semantically meaningful information, and the unsupervised SRPs relate more to the discriminative aspects of the different scenes that are present in the target dataset. The efficacy of the proposed approach is demonstrated through extensive experiments on five challenging scene classification datasets. Our experimental results show that the proposed approach consistently achieves state-of-the-art performance.

The **major contributions** of this paper are as follows **1).** We propose a new mid-level feature representation for indoor scene categorization using large-scale deep neural nets (Sec. III). **2)** Our feature description incorporates not only the discriminative patches of the target dataset but also the general object categories that are semantically meaningful (Sec. III-C). **3).** We collect the first large-scale dataset of object categories that are commonly present in indoor scenes. This dataset contains more than 1300 indoor object classes (Sec. IV-A). **4).** To improve the efficiency and performance of our approach, we propose to generate multiple smaller codebooks and a feasible feature encoding (Sec. III-C). **5).** Finally, we introduce a novel method to encode feature associations using max-margin hyper-planes (Sec. III-D).

II. Related Work

According to the level of image description, existing scene classification techniques can be categorized into three types: **1).** techniques that capture low-level appearance cues, **2).** techniques that capture the high-level spatial structure of the scene and **3).** techniques that capture mid-level relationships. The techniques that capture low-level appearance cues [6], [19] perform poorly on the majority of indoor scene types because they fail to incorporate the high-level spatial information. The techniques that model the human perceptible global spatial envelope [29] also fail to address the high variability of indoor scenes. The main reason for the low performance of these approaches is their neglect of the fine-grained objects, which are important to the task of scene classification.

Considering the need to extract global features as well as the characteristics of the constituent objects, Quattoni et al. [35] and Pandey et al. [33] represented a scene as a combination of root nodes (which capture the global characteristics of the scene) and a set of regions of interest (which capture the local object characteristics). However, the manual or automatic identification of these regions of interest makes their approach indirect and thus complicates the scene classification task. Another example of an indirect approach to scene recognition is the approach proposed by Gupta et al. [9], where the grouping, segmentation and labeling outcomes are combined to recognize scenes. Learned mid-level patches were employed for scene categorization by Juneja et al. [14], Doersch et al. [4] and Sun et al. [42]. However, these works involved substantial effort in learning the distinctive primitives, which includes a discriminative patch ranking and selection. In contrast, our mid-level representation does not require any learning. Instead, we uniformly and densely extract the mid-level patches from the images and show that these perform best when combined with supervised object representations.

Deep Convolutional Neural Networks have recently shown great promise in large-scale visual recognition and classification [3], [16], [32], [37], [54]. Although CNN features have demonstrated their discriminative power for images with one or multiple instances of the same object, they preserve the spatial structure of the image, which is not desirable when addressing the variability of indoor scenes [10]. CNN architectures utilize max-pooling operations to address the local spatial variability in the form of rotation and translation [16]. However, these operations are insufficient when addressing the large-scale deformations of objects and parts that are commonly present in indoor scenes [10], [54]. In this work, we propose a novel representation that is robust to variations in the spatial structure of indoor scenes. Our technique represents an image in terms of the association of its mid-level patches with the codebooks of the SRPs.

III. The Proposed Method

The block diagram of our proposed pipeline, called 'Deep Un-structured Convolutional Activations (DUCA)', is shown in Fig 2. Our proposed method first densely and uniformly extracts mid-level patches (Sec III-A), represents them by their convolutional activations (Sec III-B) and then encodes

[1] A codebook is a collection of distinctive mid-level patches.

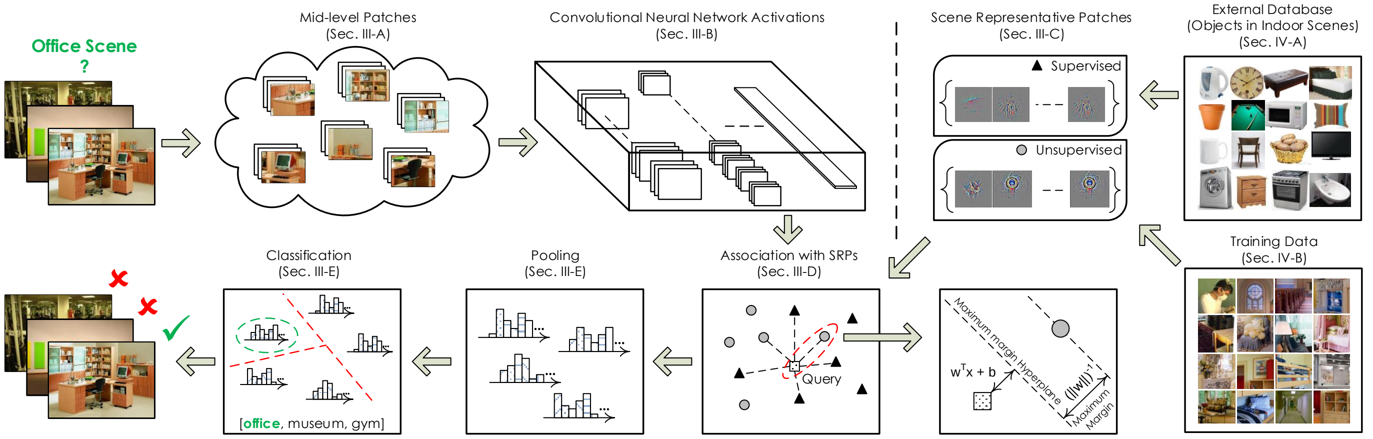


Fig. 2: Deep Un-structured Convolutional Activations: We extract dense mid-level patches from a given input image, represent the extracted patches by their convolutional activations and encode them in terms of their association with the codebooks of Scene Representative Patches (SRPs). The designed codebooks have both supervised and unsupervised SRPs. The resulting associations are then pooled, and the class-belonging decisions are predicted using a linear classifier.

them in terms of their association with the codebooks of SRPs (Sec III-D), which are generated in supervised and unsupervised manners (Sec III-C). A detailed description of each component of the proposed pipeline is presented next.

A. *Dense Patch Extraction*

To address the high variability of indoor scenes, we propose to extract mid-level feature representations instead of global [29] or local [6], [19] feature representations. Mid-level representations do not ignore object level relationships and discriminative-appearance-based local cues (unlike the high-level global descriptors) and do not ignore the holistic shape and scene structure information (unlike the low-level local descriptors). For each image, we extract dense mid-level patches using a sliding window of 224×224 pixels with a fixed step size of 32. To extract a reasonable number of patches, the smaller dimension of the image is re-scaled to an appropriate length (700 pixels in our case). Note that the idea of dense patch extraction is analogous to dense key-point extraction [28], which has shown very promising performance over well-designed key-point extraction methods in a number of tasks (e.g., action recognition [47]).

Prior to dense patch extraction, we augment the images of the dataset with their flipped, cropped and rotated versions to enhance the generalizability of our feature representation. First, five cropped images (four from the corners and one from the center) of $\frac{2}{3}$ size are extracted from the original image. Each original image is also subjected to CW and CCW rotations of $\frac{\pi}{6}$ radians, and the resulting images are included in the augmented set. The horizontally flipped versions of all eight images (1 original + 5 cropped + 2 rotated) are also included. The proposed data augmentation results in a reasonable performance boost (see Sec. IV-D).

B. *Convolutional Feature Representations*

We must map the raw image patches to a discriminative feature space where scene categories are easily separable. For this purpose, instead of using shallow or local feature representations, we use the convolutional activations from a trained deep CNN architecture. Learned representations based on CNNs have significantly outperformed hand-crafted representations in nearly all major computer vision tasks [3], [11]. Our CNN architecture is similar to 'AlexNet' [16] (trained on ILSVRC 2012) and consists of 5 convolutional and 3 fully-connected layers. The main difference compared to AlexNet is the dense connections between each pair of consecutive layers in the 8-layered network (in our case). The densely and uniformly extracted patches from the images are fed into the network's input layer after mean normalization. The processed output from the network is taken from an intermediate fully connected layer (7^{th} layer). The resulting feature representation of each mid-level patch has a dimension of 4096.

Although CNN activations capture rich discriminative information, they are inherently highly structured. This is mainly because of the sequence of operations involved in the hierarchical layers of CNNs, which preserve the global spatial structure of the image. This constraining structure is a limitation when addressing highly variable indoor scene images. To address this, we propose to encode our patches (represented by their convolutional activations) into an alternate feature space that is found to be even more discriminative (Sec. III-D). Specifically, an image is encoded in terms of the association of its extracted patches with the codebooks of the Scene Representative Patches (SRPs).

We utilize multiple-sized patches through data augmentation; therefore, contextual cues at multiple spatial resolutions can be extracted and included in our feature representation. Specifically, the extracted patches encode low-level local information, intermediate-level spatial relationships and high-level global information. This is illustrated in Fig. 3. The densely extracted patches (shown in the top row) from the rescaled image capture the different objects of the indoor scenes. The patches shown in the middle row cover the spatial relationships between these objects and are extracted from cropped versions of the original image (Sec. III-A). The patches in the bottom row capture more of the holistic and global information in the scene image, and they are extracted from the (original, flipped and rotated versions of the) original sized image. As indicated in our experiments (Sec. IV-E), the mid-level dense patch extraction strategy proves to be very useful because it captures information at multiple spatial levels.

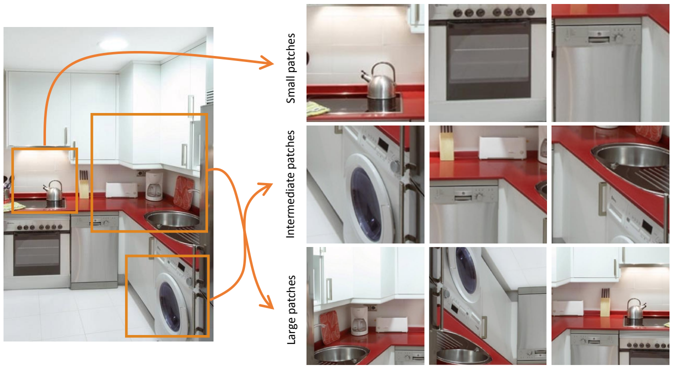


Fig. 3: Patches of three different sizes are extracted from the original image, which contains different levels of details about the indoor scene.

C. Scene Representative Patches (SRPs)

An indoor scene is a collection of several distinct objects and concepts. We are interested in extracting a set of image patches of these objects and concepts, which we call 'Scene Representative Patches' (SRPs). The SRPs can then be used as elements of a codebook to characterize any instance of an indoor scene. Examples of these patches for a bedroom scene include a bed, wardrobe, sofa or a table. Designing a comprehensive codebook of these patches is a very challenging task. There are two possible solutions: **1)** automatically learn to discover a number of discriminative patches from the training data or **2)** manually prepare an exhaustive vocabulary of all objects that can be found in indoor scenes. These solutions are quite demanding because of the possibility of a very large number of objects and because this may require automatic object detection, localization or distinctive patch selection, which in of themselves are very challenging and computationally expensive.

In this work, we propose a novel approach to compile a comprehensive set of SRPs. Our proposed approach avoids the drawbacks of the above-mentioned strategies and successfully combines their strengths, i.e., it is computationally very efficient while being highly discriminative and semantically meaningful. Our set of SRPs has two main components, compiled in a *supervised* and an *unsupervised* manner. These components are described next.

1) Supervised SRPs: A codebook of supervised SRPs is generated from images of well-known object categories expected to be present in a particular indoor scene (e.g., a microwave in a kitchen or a chair in a classroom). The codebook contains human-understandable elements that carry well-defined semantic meanings (similar to attributes [5] or object banks [21]). In this regard, we introduce the first large-scale database of object categories in indoor scenes (Sec. IV-A). The introduced database includes an extensive set of indoor objects (more than 1300). The codebook of supervised SRPs is generated from images of the database by extracting dense mid-level patches after re-sizing the smallest dimension of each image to 256 pixels. The number of SRPs in the compiled codebook is equal to the object categories in the OCIS database. For this purpose, in the feature space, each SRP is a max-pooled version of convolutional activations (Sec III-B) of all the mid-level patches extracted from that object category. The supervised codebook is then used in Sec. III-D to characterize a given scene image in terms of its constituent objects.

2) Unsupervised SRPs: The codebook of unsupervised SRPs is generated from the patches extracted from the training data. First, we densely and uniformly extract patches from training images by following the procedure described in Sec. III-A. The SRPs can then be generated from these patches using any unsupervised clustering technique. However, in our case, we randomly sample the patches as our unsupervised SRPs. This is because we are addressing a very large number of extracted patches, and an unsupervised clustering can be computationally prohibitive. We demonstrate in our experiments (Sec. IV-C, IV-D) that random sampling does not cause any noticeable performance degradation while achieving significant computational advantages.

Ideally, the codebook of SRPs should be all inclusive and cover all discriminative aspects of indoor scenes. One might therefore expect a large number of SRPs to cover all possible aspects of various scene categories. Although this is indeed the case, feature encoding from a single large codebook would be computationally burdensome (Sec. IV-D). We therefore propose to generate multiple codebooks of relatively small sizes. The association vectors from each of these codebooks can then be concatenated to generate a high-dimensional feature vector. This guarantees the incorporation of a large number of SRPs at a low computational cost. To this end, we generate three unsupervised codebooks, each with 3000 SRPs. The codebook size was selected empirically on a small validation set.

The SRPs in the supervised codebook are semantically meaningful; however, they do not include all possible aspects of the different scene categories. The unsupervised codebook compensates this shortcoming and complements the supervised codebook. The combinations of both supervised and unsupervised codebooks results in an improved discrimination and accuracy (see Sec. IV-D).

D. Feature Encoding from SRPs

Given an RGB image $I \in \mathbb{R}^{H \times W \times 3}$, our task is to find its feature representation in terms of the previously generated codebooks of SRPs (Sec. III-C1 and III-C2). For this purpose, we first densely extract patches $\{p^{(i)} \in \mathbb{R}^{224 \times 224 \times 3}\}_{i=1}^{N}$ from the image using the procedure explained in Sec. III-A. Next, the patches are represented by their convolutional activations (denoted by $x^{(i)}$), as discussed in Sec. III-B. The patches are then encoded in terms of their association with the SRPs of the codebooks. The following two strategies are devised for this purpose.

1) Sparse Linear Coding: Let $X \in \mathbb{R}^{4096\times m}$ be a codebook of m SRPs, and a mid-level patch feature representation ($x^{(i)}$) is sparsely reconstructed from the SRPs of the codebook using

$$\min_{f^{(i)}} \left\| X f^{(i)} - x^{(i)} \right\|_2 + \lambda \left\| f^{(i)} \right\|_1. \qquad (1)$$

where λ is the regularization constant. The sparse coefficient vector $f^{(i)}$ is then used as the final feature representation of the patch.

2) Proposed Classifier Similarity Metric Coding: We propose a new soft encoding method that uses the maximum margin hyper-planes to measure feature associations. Given a codebook of m SRPs, we train m linear binary one-vs-all SVMs. An SVM finds the maximum margin hyperplane that optimally discriminates an SRP from all others. Let $W \in \mathbb{R}^{4096\times m}$ be the learnt weight matrix of all learnt SVMs. A patch feature representation $x^{(i)}$ can then be encoded in terms of the trained SVMs using $f^{(i)} = W^T x^{(i)}$.

Because we have multiple codebooks (K in total), the patch feature representation $x^{(i)}$ is separately encoded from all the codebooks. The final representation of $x^{(i)}$ is then achieved by concatenating the encoded feature representation from all codebooks into a single feature vector $f^{(i)} = \left[f_1^{(i)} f_2^{(i)} \cdots f_K^{(i)} \right]$. It is important to note that, although the number of patches (N) extracted from each image can be different, the dimensions of the feature vector $f^{(i)}$ remain the same because the number of dictionary atoms are fixed.

E. Classification

The encoded feature representations from all mid-level patches of the image are stacked together as rows of a matrix F_n:

$$F_n = [(f^{(1)})^T \, (f^{(2)})^T \cdots (f^{(N)})^T]^T. \qquad (2)$$

Afterward, the columns of the matrix F_n are pooled to produce the overall feature representation (v_n) corresponding to the n^{th} image:

$$v_n = \mathcal{P}(F^{(i)}). \qquad (3)$$

where $\mathcal{P}(\cdot)$ is the pooling function. Two commonly used pooling operations (mean pooling and max pooling) are explored in our experiments (see Sec. IV-D). Finally, to perform classification, we use one-vs-one linear SVMs:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}\mathbf{w}^T + C\sum_n \left(\max(0, 1 - y^{(t)}\mathbf{w}^T v_n) \right)^2. \qquad (4)$$

where $\mathbf{w}$ is the normal vector to the learned max-margin hyperplane, C is the regularization parameter, and $y^{(t)}$ is the binary class label of the feature vector v_n.

IV. Experiments and Evaluation

We evaluate our approach on three indoor scene classification datasets. These include the MIT-67 dataset, 15 Category Scene datasets and the NYU indoor scene dataset. Confusing inter-class similarities and high within-class variabilities make these datasets very challenging. Specifically, MIT-67 is the largest dataset of indoor scene images and contains 67 classes. The images of many of these classes are very similar looking, e.g., *inside-subway* and *inside-bus* (see Fig 7 for example confusing and challenging images). Moreover, we also report results on two event and object classification datasets (the Graz-02 dataset and the 8-Sports event dataset) to demonstrate that the proposed technique is applicable to other related tasks. A detailed description of each of these datasets, followed by our experimental setups and the corresponding results, is presented in Secs. IV-B and IV-C. First, we provide a description of our introduced OCIS dataset below.

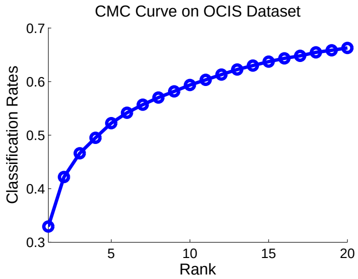


Fig. 4: CMC Curve for the benchmark evaluation on the OCIS dataset. The curve illustrates the challenging nature of the dataset.

A. A Dataset of Object Categories in Indoor Scenes

There is an exhaustive list of scene elements (including objects, structures and materials) that can be present in indoor scenes. Any information about these scene elements can prove crucial to the scene categorization task (and even beyond - e.g., for semantic labeling or attribute identification). However, to the best of our knowledge, there is no publicly available dataset of these indoor scene elements. In this paper, we introduce the first large-scale OCIS (*Object Categories in Indoor Scenes*) database. The database contains a total of $15,324$ images spanning more than 1300 frequently occurring indoor object categories. The number of images in each category is approximately 11. The database can potentially be used for fine-grained scene categorization, high-level scene understanding and attribute-based reasoning. To collect the data, a comprehensive list of 1325 indoor objects was manually chosen from the labelings provided with the MIT-67 [35] dataset. This taxonomy includes a diverse set of object classes

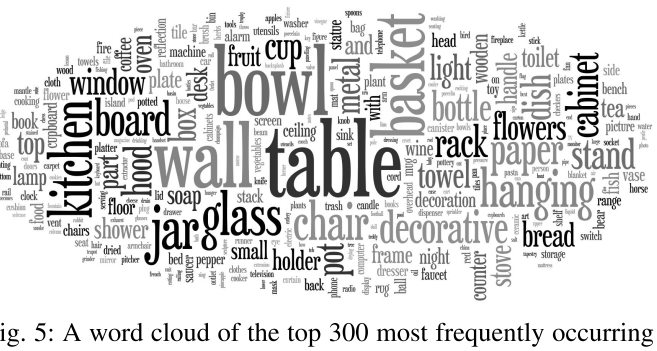


Fig. 5: A word cloud of the top 300 most frequently occurring classes in our introduced Object Categories in Indoor Scenes (OCIS) database.

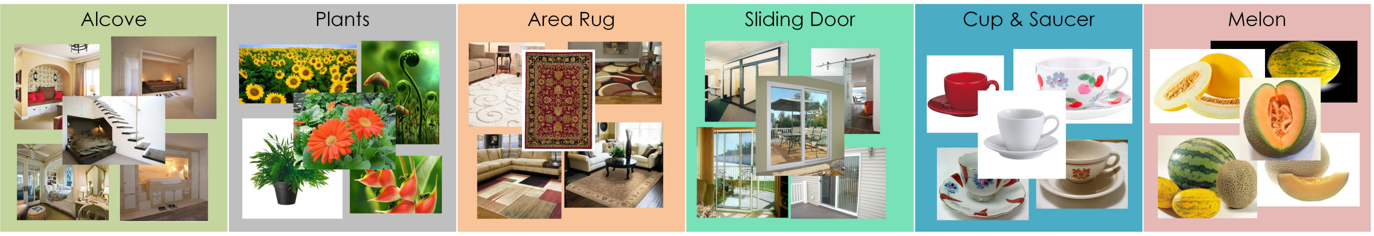


Fig. 6: Example images from the 'Object Categories in Indoor Scenes' dataset. This dataset contains a diverse set of object classes with different sizes and scales (e.g, *Alcove* and *Melon*). Each category includes a rich set of images with differences in terms of appearance, shape, viewpoint and background.

Method	Accuracy (%)	Method	Accuracy (%)
ROI + GIST [CVPR'09] [35]	26.1	OTC [ECCV'14] [25]	47.3
MM-Scene [NIPS'10] [56]	28.3	Discriminative Patches [ECCV'12] [41]	49.4
SPM [CVPR'06] [19]	34.4	ISPR [CVPR'14] [24]	50.1
Object Bank [NIPS'10] [21]	37.6	D-Parts [ICCV'13] [42]	51.4
RBoW [CVPR'12] [34]	37.9	VC + VQ [CVPR'13] [23]	52.3
Weakly Supervised DPM [ICCV'11] [33]	43.1	IFV [CVPR'13] [14]	60.8
SPMSM [ECCV'12] [18]	44.0	MLRep [NIPS'13] [4]	64.0
LPR-LIN [ECCV'12] [38]	44.8	CNN-MOP [ECCV'14] [8]	68.9
BoP [CVPR'13] [14]	46.1	CNNaug-SVM [CVPRw'14] [36]	69.0
Hybrid Parts + GIST + SP [ECCV'12] [55]	47.2	**This Paper**	**71.8**

MIT-67 Indoor Scene Dataset

TABLE I: Mean accuracy on the MIT-67 Indoor Scene Dataset. Comparisons with previous state-of-the-art methods are also shown. Our approach performs best in comparison to techniques that use a single or multiple feature representations.

ranging from a *'house'* to a *'handkerchief'*. A word cloud of the top 300 most frequently occurring classes is shown in Fig. 5. The images for each class are then collected using an online image search (Google API). Each image contains one or more instances of a specific object category. To illustrate the diverse intra-class variability of this database, we show some example images in Fig. 6. Our in-house annotated database will be made freely available to the research community.

For the benchmark evaluation, we represent the images of the database by their convolutional features and feed them to a linear classifier (SVM). A train-test split of 66%-33% is defined for each class. The classification results in terms of the Cumulative Match Curve (CMC) are shown in Fig. 4. The rank-1 and rank-20 identification rates are found to be only 32% and 67%, respectively. These modest classification rates suggest that indoor object categorization is a very challenging task.

B. Evaluated Datasets

The performance of our proposed method is evaluated on three indoor scene datasets, namely, the MIT-67 dataset, the 15-Category Scene dataset and the NYU Indoor Scene dataset. We also performed experiments on other related scene datasets, i.e., the Graz-02 dataset and the 8-Sports event dataset. These datasets contain multiple objects in each scene and have high spatial layout variations. We present below a brief description of each of these datasets followed by an analysis of our achieved performance.

1) Indoor Scene Datasets:

a) MIT-67 Dataset: This dataset contains 15,620 images of 67 indoor categories. For the performance evaluation and comparison, we followed the standard evaluation protocol in [35], in which a subset of data are used (100 images per class) and a train-test split is defined as $80\% - 20\%$ for each class.

b) 15 Category Scene Dataset: This dataset contains images of 15 urban and natural scene categories. The number of images in each category ranges from 200-400. For our experiments, we use the same evaluation setup as in [19], where 100 images per class are used for training and the remainder are used for testing.

c) NYU v1 Indoor Scene Dataset: This dataset consists of 7 indoor scene categories with a total of 2347 images. Following the standard experimental protocol [40], we used a $60\% - 40\%$ train-test split for evaluation. Care has been taken when splitting the data to ensure that a minimal or no overlap of the consecutive frames exists between the training and testing sets.

2) Other Scene Datasets:

a) Inria Graz-02 Dataset: This dataset consists of 1096 images belonging to 3 classes (bikes, cars and people) in the presence of heavy clutter, occlusions and pose variations. For the performance evaluation, we used the protocol defined in [26]. Specifically, for each class, the first 150 odd images are used for training, and the 150 even images are used for testing.

b) UIUC 8-Sports Event Dataset: This dataset contains 1574 images of 8 sports categories. Following the protocol defined in [20], we used 70 randomly sampled images for training and 60 for testing.

15-Category Scene Dataset			
Method	Accuracy(%)	Method	Accuracy (%)
GIST-color [IJCV'01] [29]	69.5	ISPR [CVPR'14] [24]	85.1
RBoW [CVPR'12] [34]	78.6	VC + VQ [CVPR'13] [23]	85.4
Classemes [ECCV'10] [44]	80.6	LMLF [CVPR'10] [2]	85.6
Object Bank [NIPS'10] [21]	80.9	LPR-RBF [ECCV'12] [38]	85.8
SPM [CVPR'06] [19]	81.4	Hybrid Parts + GIST + SP [ECCV'12] [55]	86.3
SPMSM [ECCV'12] [18]	82.3	CENTRIST+LCC+Boosting [CVPR'11] [52]	87.8
LCSR [CVPR'12] [39]	82.7	RSP [ECCV'12] [12]	88.1
SP-pLSA [PAMI'08] [1]	83.7	IFV [46]	89.2
CENTRIST [PAMI'11] [49]	83.9	LScSPM [CVPR'10] [7]	89.7
HIK [ICCV'09] [48]	84.1		
OTC [ECCV'14] [25]	84.4	**This paper**	**94.5**

TABLE II: Mean accuracy on the 15-Category Scene Dataset. Comparisons with the previous best techniques are also shown.

UIUC 8-Sports Dataset	
Method	Accuracy (%)
GIST-color [IJCV'01] [29]	70.7
MM-Scene [NIPS'10] [56]	71.7
Graphical Model [ICCV'07] [20]	73.4
Object Bank [NIPS'10] [21]	76.3
Object Attributes [ECCV'12] [22]	77.9
CENTRIST [PAMI'11] [49]	78.2
RSP [ECCV'12] [12]	79.6
SPM [CVPR'06] [19]	81.8
SPMSM [ECCV'12] [18]	83.0
Classemes [ECCV'10] [44]	84.2
HIK [ICCV'09] [48]	84.2
LScSPM [CVPR'10] [7]	85.3
LPR-RBF [ECCV'12] [38]	86.2
Hybrid Parts + GIST + SP [ECCV'12] [55]	87.2
LCSR [CVPR'12] [39]	87.2
VC + VQ [CVPR'13] [23]	88.4
IFV [46]	90.8
ISPR [CVPR'14] [24]	89.5
This paper	**98.7**

TABLE III: Mean accuracy on the UIUC 8-Sports Dataset.

NYU Indoor Scene Dataset	
Method	Accuracy (%)
BoW-SIFT [ICCVw'11] [40]	55.2
RGB-LLC [TC'13] [43]	78.1
RGB-LLC-RPSL [TC'13] [43]	79.5
This paper	**80.6**

TABLE IV: Mean Accuracy for the NYU v1 Dataset.

Graz-02 Dataset				
	Cars	People	Bikes	Overall
OLB [SCIA'05] [30]	70.7	81.0	76.5	76.1
VQ [ICCV'07] [45]	80.2	85.2	89.5	85.0
ERC-F [PAMI'08] [27]	79.9	-	84.4	82.1
TSD-IB [BMVC'11] [15]	87.5	85.3	91.2	88.0
TSD-k [BMVC'11] [15]	84.8	87.3	90.7	87.6
This paper	98.7	98.0	99.0	**98.6**

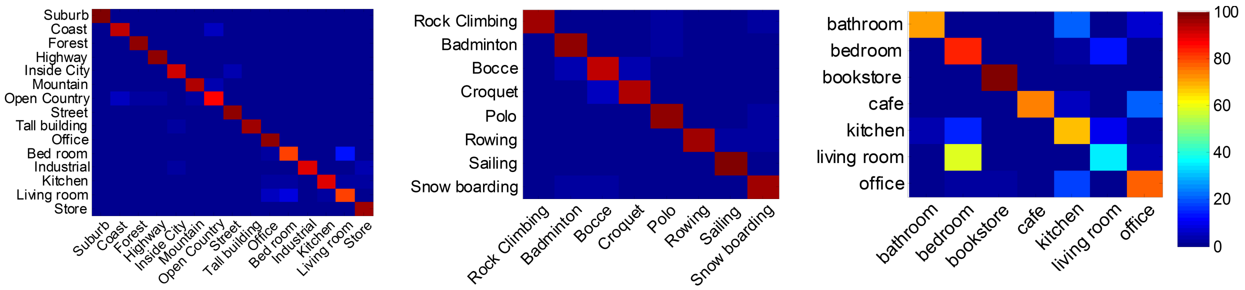
TABLE V: Equal Error Rates (EER) for the Graz-02 dataset.

C. Experimental Results

The quantitative results of the proposed method applied to the task of indoor scene categorization are presented in Tables I, II and IV. The proposed method achieves the highest classification rate on all three datasets. Note that the reported performances are achieved with the sparse linear coding method described in Sec. III-D. Compared with the state of the art, a relative performance increment of 4.1%, 5.4% and 1.3% is achieved on the MIT-67, Scene-15 and NYU datasets, respectively. Among the compared methods, the mid-level feature representation-based methods [4], [41], [42] perform better than the other methods. Our proposed mid-level feature-based method not only achieves a higher accuracy but also is computationally efficient (e.g., [4] takes weeks to train several part detectors). Furthermore, compared with existing methods, our proposed method uses a lower dimensional feature representation for classification (e.g., the Juneja et al. [14] Improved Fisher Vector (IFV) has dimensionality > 200K, and the the Gong et al. [8] MOP representation has dimensionality > 12K).

In addition to indoor scene classification, we also evaluate our approach on other scene classification tasks where large variations and deformations are present. To this end, we report the classification results on the UIUC 8-Sports dataset and on the Graz-02 dataset (see Tables III and V). It is interesting to note that the Graz-02 dataset contains heavy clutter and pose and scale variations (e.g., for certain 'car' images, only 5% of the pixels are covered by the *car* in a scene). Our approach achieved high accuracies of 98.7% and 98.6% on the UIUC 8-Sports and Graz-02 datasets, respectively. These performances are 10.2% and 12.6% higher than the previous best methods on the UIUC 8-Sports and Graz-02 datasets, respectively.

The class-wise classification accuracies of the MIT-67, UIUC 8-Sports, Scene-15 and NYU datasets are shown in the form of confusion matrices in Figs. 8 and 10. Note the very strong diagonal in all confusion matrices. The majority



(a) 15-Category Scenes Dataset

(b) UIUC 8-Sports Dataset

(c) NYU Indoor Scene Dataset

Fig. 8: Confusion matrices for three scene classification datasets. *(Best viewed in color)*

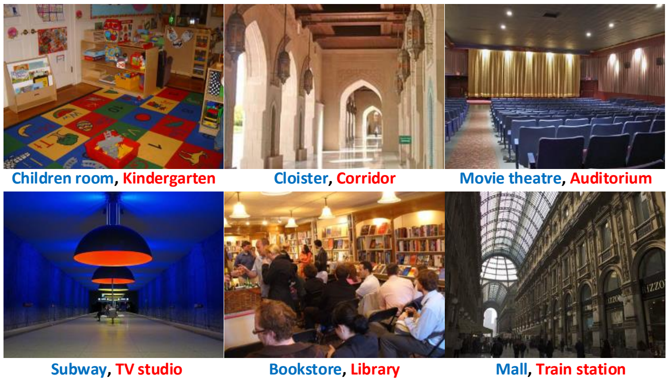


Fig. 7: Example mistakes and the limitations of our method. Most of the incorrect predictions are due to ambiguous cases. The actual and predicted class names are shown in *'blue'* and *'red'*, respectively. *(Best viewed in color)*

Variants of Our Approach	Accuracy (%)
Supervised codebook	68.5
Unsupervised codebook	69.9
Supervised + Unsupervised	71.8
K-means clustering	72.0
Random sampling	71.8
Single large codebook	71.4
Multiple smaller codebooks	71.8
Sparse linear coding	71.8
Classifier similarity metric coding	69.9
Mean-poling	69.7
Max-pooling	71.8
Original data	69.1
Data augmentation	71.8

TABLE VI: Ablative Analysis on MIT-67 Scene Dataset.

($>$ 90%) of the mistakes are made for the closely related classes, e.g., *coast-opencountry* (Fig. 8a), *croquet-bocce* (Fig. 8b), *bedroom-livingroom* (Fig. 8c), *dentaloffice-operatingroom* (Fig. 10) and *library-bookstore* (Fig. 10). We also show examples of miss-classified images in Fig. 7. The results show that the classes with significant visual and semantic similarities are confused among each others, e.g., *childrenroom-kindergarten* and *movietheatre-auditorium* (Fig. 7).

To visualize which patches contributed most to a correct classification, we find the classifier scores of each patch for the correct prediction of a scene type. We plot the heat map of the patch contribution scores in Fig. 9. Note that, to calculate patch contributions, we do not pool the feature vectors to generate a single feature representation for the entire image (Sec. III-E). Rather, we use the already trained classifier to predict scores for each patch instead of the entire image.

We find that the most distinctive patches, which carry valuable information, have a higher contribution to the correct prediction of a scene class. Moreover, mid-level patches carry an intermediate level of scene details and contextual relationships between objects, which facilitates the scene classification process.

D. Ablative Analysis

To analyze the effect of the different components of the proposed scheme on the final performance, we conduct an ablation study. Table VI summarizes the results achieved on the MIT-67 Scene dataset when different components are replaced or removed from the final framework.

We find that the supervised and unsupervised codebooks individually perform reasonably well. However, their combination produces a state-of-the-art performance. For the unsupervised codebook, k-means clustering performs slightly better, however, at the cost of a considerable amount of computational resources ($\sim$ 40 GB RAM for the MIT-67 dataset) and processing time ($\sim$ 1 day for the MIT-67 dataset). In contrast, the random sampling of MRPs provides a comparable performance, with a large boost in computational efficiency. Feature encoding from a single large codebook not only produces a lower performance but also requires more computational time and memory. In our experiments, feature encoding from a single large codebook requires almost four times the amount of time ($\sim$ 15.16 sec/image) taken by multiple smaller codebooks ($\sim$ 3.44 sec/image). The resulting features performed best when the max-pooling operation was applied to combine them.

E. Effectiveness of Mid-level Information

An interesting aspect of our approach relates to the use of the mid-level patch encoding as a feature representation for indoor scenes. Because several previous object detection and

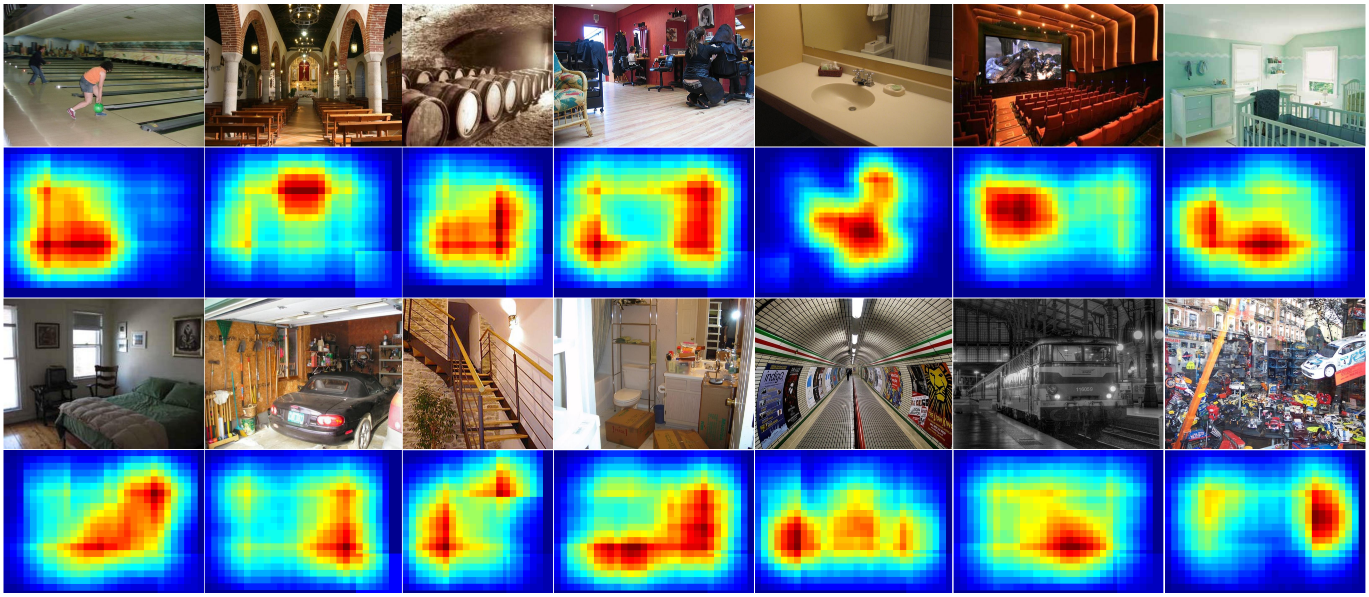


Fig. 9: The contributions of distinctive patches for the correct class prediction of a scene are shown in the form of a heat map ('*red*' means greater contribution). These examples show that our approach captures the discriminative properties of distinctive mid-level patches and uses them to predict the correct class. *(Best viewed in color)*

outdoor scene classification works (e.g., [29]) proposed to use global feature representations, we compare our performance with a global feature description strategy. For this purpose, we learn the parameters of the last three fully connected layers of the CNN model (used in Sec. III-B for the convolutional feature encoding of the mid-level patches) using indoor scenes from the MIT-67 dataset. Specifically, instead of using the mid-level patches, the network is trained with the full images. The trained network model when evaluated on the testing images of the MIT-67 dataset yields a classification accuracy of 63.7%. We attribute this low performance to two main causes. *First*, the number of training samples (only 5,360 images of the MIT-67 dataset) is insufficient to learn a set of optimal parameters. *Second*, because the network was trained with full images (instead of the mid-level patches), it learns to preserve the spatial structure of an image. This is not desirable for many indoor scene categories because the constituent objects in images of the same scene type can also be present in various other possible spatial layouts. Therefore, the proposed mid-level patch extraction strategy helps to achieve spatial layout invariance and as a result produces a higher accuracy.

F. Dimensionality Analysis

We also analyze the dimensions of the features used in our approach and those of the state-of-the-art approaches for scene classification on the MIT-67 dataset. Table VII shows that our approach was able to perform significantly better in terms of classification accuracy with a relatively low-dimensional feature representation. The Spatial Pyramid Matching (SPM) [19] obtains the lowest dimensional feature representation (half the dimensions of our approach). Our approach, however, achieves a 37.4% better accuracy compared to the results in [19]. Compared to two variants of a recent CNN-based approach [13], namely, the cross-level LLC coding (CNN-CLLP) and the multi-level max-pooling (CNN-MLMP), our method not only performs better (by a decent margin of $3-4\%$) but also uses a lower dimensional feature representation (10K compared to 16K and 20K for CNN-CLLP and CNN-MLMP, respectively).

Approach	Dimensions	Accuracy(%)
SPM [19]	$5,000$	34.4
FV + BoP [14]	$2,21,550$	63.2
MLRep [4]	$60,000$	64.0
CNN-MLMP [13]	$20,480$	67.9
CNN-MOP [8]	$12,288$	68.9
CNN-CLLP [13]	$16,384$	69.0
This paper	$10,300$	71.8

TABLE VII: Analysis of feature dimensions and their corresponding accuracies.

G. Timing Analysis

To study the efficiency of our approach, we present a detailed timing analysis. Our experiments were conducted using a standard machine with an Intel® Core™ i7-4770 CPU (3.4 GHz) on the largest publicly available indoor scene data set (MIT-67). The training procedure consists of four main steps (Sec. III), and the procedure takes (on average) 3.49 seconds/image. More specifically, the technique requires 44.1 ± 5.1 milliseconds to extract the mid-level patches and compute their convolutional features, $5.6 \pm .3$ micro-seconds for the uniform random sampling of the encoded patches to generate codebooks, 3.44 ± 0.4 seconds to represent a training image in terms of the association of its patches with the elements of the generated codebooks, and $4.8 \pm .5$ milliseconds to train a multi-class Support Vector Machine

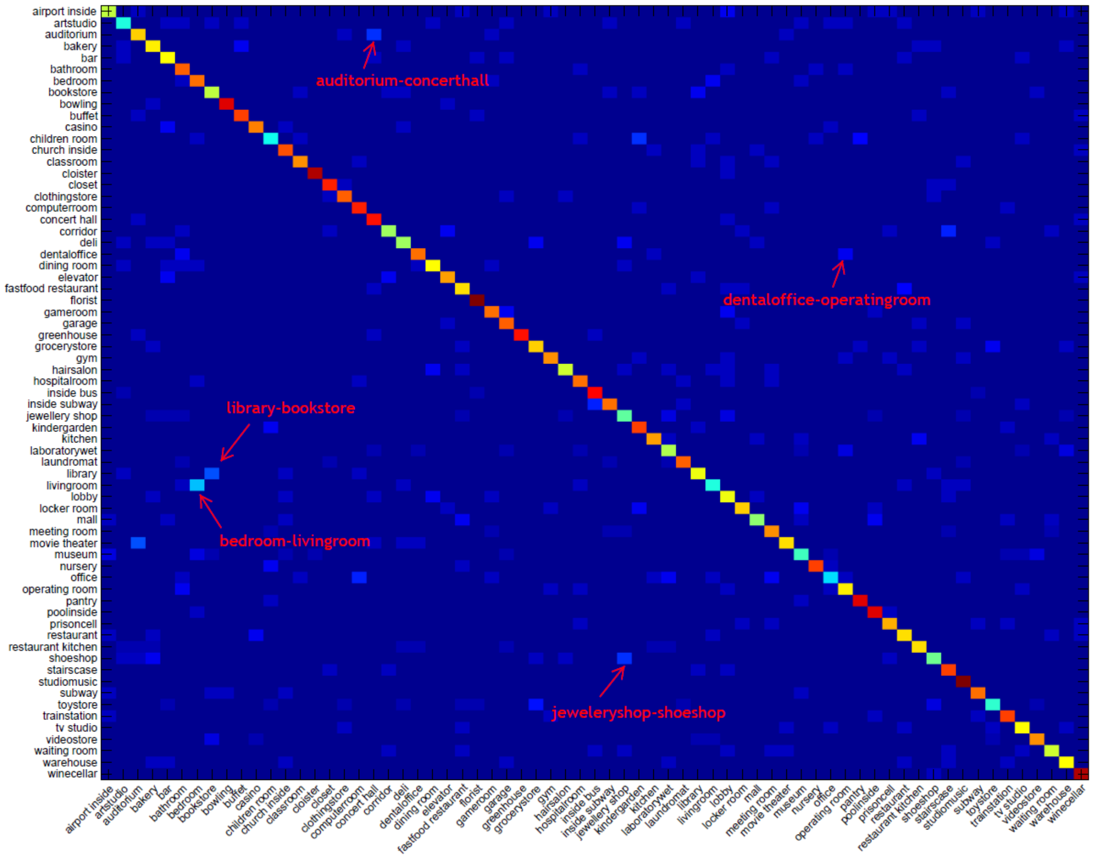


Fig. 10: Confusion Matrix for the MIT-67 dataset. The proposed method confuses similar-looking classes with each other, e.g., library with bookstore, bedroom with living room and dental office with operating room. *(Best viewed in color)*

(SVM) Classifier. All these timings are per image. The total training time on the largest publicly available indoor scene dataset (MIT-67 dataset with 5,360 training images) is ~ 5 hours.

For a given test image, the technique requires 44.3 ± 5.0 milliseconds to extract the mid-level patches and compute their convolutional features, 3.47 ± 0.4 seconds to associate them with the codebooks and 0.18 ± 0.3 milliseconds to classify the image using the trained SVM model. The technique therefore requires a total of 3.4 seconds to classify a given image using our proposed approach.

H. Sensitivity Analysis

We perform a sensitivity analysis of the proposed method with respect to the codebook size, which is one of the most important parameters of our proposed approach. To study the relationship between the number of codebook elements and the time consumed to find patch associations with these elements, we perform experiments with different sizes of the codebooks. Specifically, experiments are conducted by fixing the total number of codebooks to 3 and gradually increasing the number of elements in each codebook from 1000 to 5000. Two dominant trends can be noticed from our experimental results (shown in Fig. 11). *First*, the association time increases when the number of codebook elements are increased. *Second*, a parallel implementation with multiple small codebooks (shown in '*blue*') is substantially quicker compared to the use of a single large codebook (shown in '*red*'). More specifically, the association computation time is significantly reduced by a factor of ~ 5. Considering the computational advantages and the achieved performance, we therefore make an intermediate choice of a total of 9000 elements, which are divided equally among the three codebooks (shown with a '*gray*' line in Fig. 11).

V. Conclusion

This paper proposed a robust feature representation based on discriminative mid-level convolutional activations for highly variable indoor scenes. To suitably contrive the convolutional activations for indoor scenes, the paper proposed to break their inherently preserved global spatial structure by encoding them in a number of codebooks. These codebooks are composed of distinctive patches and of the semantically labeled elements. For the labeled elements, we introduced the first large-scale dataset of object categories of indoor scenes. Our approach achieves state-of-the-art performance on five very challenging datasets.

Acknowledgments

This research was supported by the SIRF and IPRS scholarships from the University of Western Australia (UWA) and the Australian Research Council (ARC) grants DP110102166, DP150100294 and DE120102960.

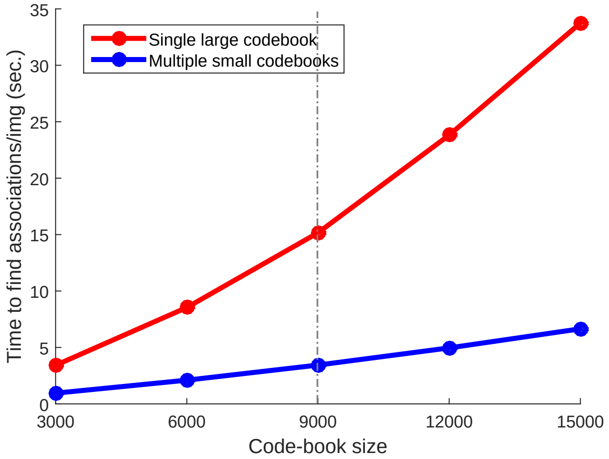


Fig. 11: Trend of the time consumed to associate extracted patches (per image) with the codebook elements with respect to the codebook size.

References

[1] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2559–2566.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[4] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Advances in Neural Information Processing Systems*, 2013, pp. 494–502.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1778–1785.

[6] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *International Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2005, pp. 524–531.

[7] S. Gao, I. W. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3555–3561.

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European Conference on Computer Vision*. Springer International Publishing, 2014, pp. 392–407.

[9] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 564–571.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[12] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 730–743.

[13] Z. Jie and S. Yan, "Robust scene classification with cross-level llc coding on cnn features," in *Computer Vision–ACCV 2014*. Springer, 2014, pp. 376–390.

[14] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 923–930.

[15] J. Krapac, J. Verbeek, F. Jurie *et al.*, "Learning tree-structured descriptor quantizers for image categorization," in *British Machine Vision Conference*, 2011.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[17] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1284–1291.

[18] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *European Conference on Computer Vision*. Springer, 2012, pp. 359–372.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *International Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 2169–2178.

[20] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[21] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in Neural Information Processing Systems*, 2010, pp. 1378–1386.

[22] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *Trends and Topics in Computer Vision*. Springer, 2012, pp. 57–69.

[23] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 851–858.

[24] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," 2014.

[25] R. Margolin, L. Zelnik-Manor, and A. Tal, "Otc: A novel local descriptor for scene classification," in *European Conference on Computer Vision*. Springer, 2014, pp. 377–391.

[26] M. Marszatek and C. Schmid, "Accurate object localization with shape masks," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[27] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1632–1646, 2008.

[28] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision*. Springer, 2006, pp. 490–503.

[29] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[30] A. Opelt and A. Pinz, "Object localization with boosting and weak supervision for generic object recognition," in *SCIA*. Springer, 2005, pp. 862–871.

[31] M. Oquab, L. Bottou, I. Laptev, J. Sivic *et al.*, "Learning and transferring mid-level image representations using convolutional neural networks," 2014.

[32] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian *et al.*, "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," *arXiv preprint arXiv:1409.3505*, 2014.

[33] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *International Conference on Computer Vision*. IEEE, 2011, pp. 1307–1314.

[34] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2775–2782.

[35] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[36] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2014.

[38] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *European Conference on Computer Vision*. Springer, 2012, pp. 228–241.

[39] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3618–3625.

[40] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *International Conference on Computer Vision Workshops*, 2011.

[41] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *European Conference on Computer Vision*. Springer, 2012, pp. 73–86.

[42] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *International Conference on Computer Vision*. IEEE, 2013, pp. 3400–3407.

[43] D. Tao, L. Jin, Z. Yang, and X. Li, "Rank preserving sparse learning for kinect based scene classification." *IEEE transactions on cybernetics*, vol. 43, no. 5, p. 1406, 2013.

[44] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *European Conference on Computer Vision*. Springer, 2010, pp. 776–789.

[45] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[46] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[47] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid *et al.*, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009.

[48] J. Wu and J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *International Conference on Computer Vision*. IEEE, 2009, pp. 630–637.

[49] ——, "Centrist: A visual descriptor for scene categorization," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492.

[51] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 702–709.

[52] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 2777–2784.

[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[54] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *International Conference on Computer Vision*. IEEE, 2011, pp. 2018–2025.

[55] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 172–185.

[56] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing, "Large margin learning of upstream scene understanding models," in *Advances in Neural Information Processing Systems*, 2010, pp. 2586–2594.

Salman H. Khan received his B.E. degree in Electrical Engineering from National University of Sciences and Technology (NUST), Pakistan in 2012. During 2010-13, he worked with research groups at NUST and FAST-NUCES, Islamabad. He has also been a visiting researcher at National ICT Australia (NICTA), CRL, in 2015. He completed his PhD from the University of Western Australia (UWA) in 2016, where he was awarded prestigious International Postgraduate Research Scholarship (IPRS) for doctoral studies. He is currently working as a Researcher at CSIRO (Data61) since April 2016. His research interests include computer vision, pattern recognition and machine learning.

H e received my Bachelor of Engineering degree from National University of Science and Technology (NUST), Pakistan, in 2009. Later, he was awarded Erasmus Mundus Scholarship for a joint European Masters degree program. He completed my PhD from The University of Western Australia (UWA) in 2015. His PhD thesis received prestigious Robert Street Prize and Deans List Honorable mention award at UWA. He worked as a PostDoc at IBM Research Australia before joining as an Assistant Professor at The University of Canberra. His research interests include computer vision, signal and image processing, pattern recognition and machine learning.

Mohammed Bennamoun received his MSc degree in control theory from Queens University, Kingston, Canada, and his Ph.D. degree in computer vision from Queens/QUT in Brisbane, Australia. He is currently a Winthrop Professor at the University of Western Australia, Australia. He served as a guest editor for a couple of special issues in International journals such as the International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). He was selected to give conference tutorials at the European Conference on Computer Vision (ECCV) and the International Conference on Acoustics Speech and Signal Processing (ICASSP). He organized several special sessions for conferences, e.g., the IEEE International Conference in Image Processing (ICIP). He also contributed in the organization of many local and international conferences. His research interests include control theory, robotics, object recognition, artificial neural networks, signal/image processing, and computer vision. He published more than 300 journal and conference publications.

Ferdous A Sohel received PhD degree from Monash University, Australia in 2009. He is currently a Senior Lecturer in Information Technology at Murdoch University, Australia. Prior to joining Murdoch University, he was a Research Assistant Professor/ Research Fellow at the School of Computer Science and Software Engineering, The University of Western Australia from January 2008 to mid-2015. His research interests include computer vision, image processing, pattern recognition, multimodal biometrics, scene understanding, robotics, and video coding. He has published more than 65 scientific articles. He is a recipient of prestigious Discovery Early Career Research Award (DECRA) funded by the Australian Research Council. He is also a recipient of the Early Career Investigators award (UWA) and the best PhD thesis medal form Monash University. He is a member of Australian Computer Society and the IEEE.

Roberto Togneri (M89-SM04) received the B.E. degree in 1985, and the Ph.D degree in 1989 both from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now currently a Professor. Prof Togneri is a member of the Signals and Systems Engineering Research Group and heads the Signal and Information Processing Lab. His research activities include signal processing and robust feature extraction of speech signals, statistical and neural network models for speech and speaker recognition, audio-visual recognition and biometrics, and related aspects of communications, information retrieval, and pattern recognition. He has published over 150 refereed journal and conference papers in the areas of signals and information systems, was the chief investigator on two Australian Research Council Discovery Project research grants from 2010 to 2013, and is currently an Associate Editor for IEEE Signal Processing Magazine Lecture Notes and Editor for IEEE Transactions on Speech, Audio and Language Processing.