A Spatial Layout and Scale Invariant Feature Representation for Indoor Scene Classification

M. Hayat, S. H. Khan, M. Bennamoun, and S. An,

Abstract—Unlike standard object classification, where the image to be classified contains one or multiple instances of the same object, indoor scene classification is quite different since the image consists of multiple distinct objects. Further, these objects can be of varying sizes and are present across numerous spatial locations in different layouts. For automatic indoor scene categorization, large scale spatial layout deformations and scale variations are therefore two major challenges and the design of rich feature descriptors which are robust to these challenges is still an open problem. This paper introduces a new learnable feature descriptor called “spatial layout and scale invariant convolutional activations” to deal with these challenges. For this purpose, a new Convolutional Neural Network architecture is designed which incorporates a novel ‘Spatially Unstructured’ layer to introduce robustness against spatial layout deformations. To achieve scale invariance, we present a pyramidal image representation. For feasible training of the proposed network for images of indoor scenes, the paper proposes a new methodology which efficiently adapts a trained network model (on a large scale data) for our task with only a limited amount of available training data. Compared with existing state of the art, the proposed approach achieves a relative performance improvement of 3.2%, 3.8%, 7.0%, 11.9% and 2.1% on MIT-67, Scene-15, Sports-8, Graz-02 and NYU datasets respectively.

Index Terms—Indoor Scenes Classification, Spatial Layout Variations, Scale Invariance

I. INTRODUCTION

Recognition/classification is an important computer vision problem and has gained significant research attention over last few decades. Most of the efforts, in this regard, has been tailored towards generic object recognition (an image with one or multiple instances of the same object) and face recognition (an image with the face region of the person). Unlike these classification tasks, indoor scene classification is quite different since an image of an indoor scene contains multiple distinct objects, with different scales and sizes and laid across different spatial locations in a number of possible layouts. Due to the challenging nature of the problem, the state of the art performance for indoor scene classification is much lower (69% classification accuracy on MIT-67 dataset with only 67 classes [7]) compared with other classification tasks such as object classification (94% rank-5 identification rate on ImageNet database with 1000 object categories [36]) and face recognition (human level performance on face recognition on real life datasets including Labeled Faces in the Wild and YouTube Faces [39]). This paper proposes a novel method of feature description, specifically tailored for indoor scene images, in order to address the challenges of large scale spatial layout deformations and scale variations.

We can characterize some indoor scenes by only global spatial information [26], [31], whereas for others, local appearance information [5], [16], [23] is more critical. For example, a corridor can be predominantly characterized by a single large object (walls) whereas a bedroom scene is characterized by multiple objects (e.g., sofa, bed, table). Both global and local spatial information must therefore be leveraged in order to accommodate different scene types [30]. This however is very challenging, for two main reasons. First, the spatial scale of the constituent objects varies significantly across different scene types. Second, the constituent objects can be present in different spatial locations and in a number of possible layouts. This is demonstrated in the example images of the kitchen scene in Fig. [1] where a microwave can be present in many different locations in the image with significant variations in scale, pose and appearance.

This paper aims to achieve invariance with respect to the spatial layout and the scale of the constituent objects for indoor scene images. For this purpose, in order to achieve invariance with respect to the spatial scale of objects, we generate a pyramidal image representation where an image is resized to different scales, and features are computed across these scales (Sec. III-C). To achieve spatial layout invariance, we introduce a new method of feature description which is based on a proposed mod-
Fig. 1: The spatial structure of indoor scenes is loose, irregular and unpredictable which can confuse the classification system. As an example, a microwave in a kitchen scene can be close to the sink, fridge, kitchen door or top cupboards (green box in the images). Our objective is to learn feature representations which are robust to these variations by spatially shuffling the convolutional activations (Sec. III).

ified Convolutional Neural Network (CNN) architecture (Sec. III-A).

CNNs preserve the global spatial layout in an image. This is desirable for the classification tasks where an image predominantly contains only a single object (e.g., objects in ImageNet database [32]). However, for a high level vision task such as indoor scene classification, an image may contain multiple distinct objects across different spatial locations. We therefore want to devise a method of feature description which is robust with respect to the spatial layout of objects in a scene. Although commonly used local pooling layers (max or mean pooling) in standard CNN architectures have been shown to achieve viewpoint and pose invariance to some extent [9], [14], these layers cannot accommodate large-scale deformations that are caused by spatial layout variations in indoor scenes. In order to achieve spatial layout invariance, this paper introduces a modified CNN architecture with an additional layer, termed ‘spatially unstructured layer’ (Sec. III-A). The proposed CNN is then trained with images of indoor scenes (using our proposed strategy described in Sec. III-B) and the learnt feature representations are invariant to the spatial layout of the constituent objects.

Training a deep CNN requires a large amount of data because the number of parameters to be learnt is quite huge. However, for the case of indoor scenes, we only have a limited number of annotated training data. This becomes then a serious limitation for the feasible training of a deep CNN. Some recently proposed techniques demonstrate that pre-trained CNN models (on large datasets e.g., ImageNet) can be adapted for similar tasks with limited additional training data [3]. However, cross domain adaptation becomes problematic in the case of heterogeneous tasks due to the different natures of source and target datasets. For example, an image in the ImageNet dataset contains mostly centered objects belonging to only one class. In contrast, an image in an indoor scene dataset has many constituent objects, all appearing in a variety of layouts and scales. In this work, we propose an efficient strategy to achieve cross domain adaptation with only a limited number of annotated training images in the target dataset (Sec. III-B).

The major contributions of this paper can be summarized as: 1) A new method of feature description (using the activations of a deep convolutional neural network) is proposed to deal with the large-scale spatial layout deformations in scene images (Sec III-A). 2) A pyramidal image representation is proposed to achieve scale invariance (Sec III-C). 3) A novel transfer learning approach is introduced to efficiently adapt a pre-trained network model (on a large dataset) to any target classification task with only a small amount of available annotated training data (Sec III-B) and 4) Extensive experiments are performed to validate the proposed approach. Our results show a significant performance improvement for the challenging indoor scene classification task on a number of datasets.

II. RELATED WORK

Indoor scene classification has been actively researched and a number of methods have been developed in recent years [16], [28], [30], [31], [37], [38], [45]. While some of these methods focus on the holistic properties of scene images (e.g., CENTRIST [45], Gist descriptor [26]), others give more importance to the local distinctive aspects (e.g., dense SIFT [16], HOG [46]). In this paper, we argue that we cannot rely on either of the local or holistic image characteristics to describe all indoor scene types [30]. For some scene types, holistic or global image characteristics are enough (e.g., corridor), while for others, local image properties must be considered (e.g., bedroom, shop). We therefore neither focus on the global nor the local feature description and instead extract mid-level image patches to encode an intermediate level of information. Further, we propose a pyramidal image representation which is able to capture the discriminative aspects of indoor scenes at multiple levels.

Recently, mid-level representations have emerged as a competitive candidate for indoor scene classification. Strategies have been devised to discover discriminative mid-level image patches which are then encoded by a
feature descriptor. For example, the works \cite{4, 12, 38} learn to discover discriminative patches from the training data. Our proposed method can also be categorized as a mid-level image patches based approach. However, our method is different from previous methods, which require discriminative patch ranking and selection procedures or involve the learning of distinctive primitives. In contrast, our method achieves state of the art performance by simply extracting mid-level patches densely and uniformly from an image (see more details in Sec. III-D).

An open problem in indoor scene classification is the design of feature descriptors which are robust to global layout deformations. The initial efforts to resolve this problem used bag-of-visual-words models or variants (e.g., \cite{1, 16, 47}), which are based on locally invariant descriptors e.g., SIFT \cite{22}. Recently, these local feature representations have been outperformed by learned feature representations from deep neural networks \cite{14, 51, 32}. However, since there is no inherent mechanism in these deep networks to deal with the high variability of indoor scenes, several recent efforts have been made to fill in this gap (e.g., \cite{7, 9}). The bag of features approach of Gong et al. \cite{7} performs VLAD pooling \cite{10} of CNN activations. Another example is the combination of spatial pyramid matching and CNNs (proposed by He et al. \cite{9}) to increase the feature’s robustness. These methods, however, devise feature representations on top of CNN activations and do not inherently equip the deep architectures to effectively deal with the large deformations. In contrast, this work provides an alternative strategy based on an improved network architecture to enhance invariance towards large scale deformations. The detailed description of our proposed feature representation method is presented next.

III. PROPOSED SPATIAL LAYOUT AND SCALE INVARIANT CONVOLUTIONAL ACTIVATIONS - S^2ICA

The block diagram of our proposed Spatial Layout and Scale Invariant Convolutional Activations (S^2ICA) based feature description method is presented in Fig 2. The detailed description of each of the blocks is given here. We first present our baseline CNN architecture followed by a detailed description of our spatially unstructured layer in Sec. III-A. Note that the spatially unstructured layer is introduced to achieve invariance to large scale spatial deformations, which are commonly encountered in images of indoor scenes. The baseline CNN architecture is pre-trained for a large scale classification task. A novel method is then proposed to adapt this pre-trained network for the specific task of scene categorization (Sec. III-B). Due to the data hungry nature of CNNs, it is not feasible to train a deep architecture with only a limited amount of available training data. For this purpose, we pre-train a ‘TransferNet’, which is then appended with the initialized CNN and the whole network can then be efficiently fine-tuned for the scene classification task. Convolutional activations from this fine-tuned network are then used for a robust feature representation of the input images. To deal with the scale variations, we propose a pyramidal image representation and combine the activations from multiple levels which results in a scale invariant feature representation (Sec. III-C). This representation is then finally used by a linear Support Vector Machine (SVM) for classification (Sec. III-D).

A. CNN Architecture

Our baseline CNN architecture is presented in Fig 3. It consists of five convolutional layers and four fully connected layers. The architecture of our baseline CNN is similar to AlexNet \cite{14}. The main difference is that we introduce extra fully connected layer, and that all of our neighboring layers are densely connected (in contrast to the sparse connections in AlexNet). To achieve spatial layout invariance, the architecture of the baseline CNN is modified and a new unstructured layer is added after the first sub-sampling layer. A brief description of each layer of the network follows next.

Let us suppose that the convolutional neural network consists of $L$ hidden layers and each layer is indexed by $l \in \{1 \ldots L\}$. The feed-forward pass can be described as a sequence of convolution, optional sub-sampling and normalization operations. The response of each convolution node in layer $l$ is given by:

$$a_n^l = f \left( \sum_m (a_m^{l-1} * k_{m,n}^l) + b_n^l \right),$$

(1)

where $k$ and $b$ denote the learned kernel and bias, the indices $(m, n)$ indicate that the mapping is from the $m^{th}$ feature map of the previous layer to the $n^{th}$ feature map of the current layer. The function $f$ is the element-wise Rectified Linear Unit (ReLU) activation function. The response of each normalization layer is given by:

$$a_n^l = \frac{a_n^{l-1}}{\left( \alpha + \sum_{j=\max(0, n-\sigma)}^{\min(N-1, n+\sigma)} (a_j^{l-1})^2 \right)^{\gamma}}$$,

(2)

where $\alpha, \beta, \gamma, \sigma$ are constants and $N$ is the total number of kernels in the layer. The response of each sub-

\footnote{These constants are defined as in \cite{14}: $\alpha = 2, \beta = 10^{-4}, \gamma = 3/4$ and $\sigma = 5/2$.}
Fig. 2: Overview of the proposed Spatial Layout and Scale Invariant Convolutional Activations (S^2ICA) based feature description method. Mid-level patches are extracted from three levels (A, B, C) of the pyramidal image representation. The extracted patches are separately feed-forwarded to the two trained CNNs (with and without the spatially unstructured layer). The convolutional activations based feature representation of the patches is then pooled and a single feature vector for the image is finally generated by concatenating the feature vectors from both CNNs. Figure best seen in color.

The transformation matrix \( T \) has the following properties:
- \( T = \{t_{ij}\} \) is a permutation matrix (\( T : \{u_{ij}\} \rightarrow \{\hat{u}_{ij}\} \)) since the sum along each row and column is always equal to one i.e., \( \sum_{i} t_{ij} = \sum_{j} t_{ij} = 1 \).
- \( T \) is a bistochastic matrix and therefore according to Birkhoff-von Neumann theorem and the above property, \( T \) lies on the convex hull of the set of bistochastic matrices.
- It is a binary matrix with entries belonging to the Boolean domain \{0, 1\}.
- Its an orthogonal matrix i.e., \( TT^T = I \) and \( T^{-1} = T^T \).

Using the matrix \( T \), we transform \( \hat{U} \) to become:
\[
\hat{U} = (U^T T)^T T = T^T U T.
\]

The updated matrix \( \hat{U} \) contains the new indices of the modified feature maps. If \( Y(\cdot) \) is a function which reads the indices of the blocks stored in the form of tuples in matrix \( \hat{U} \), the layer output are as follows:
\[
a_l^n = r \ast Y(\alpha_l^{n-1}, \hat{U}),
\]

where, \( r \sim \text{Bernoulli}(p) \).

Fig. 4 illustrates the distortion operation performed by the spatially unstructured layer for a different number of blocks.
Fig. 3: The architecture of our proposed Convolutional Neural Network used to learn tailored feature representations for scene categorization. We devise a strategy (see Sec. III-B and Alg. 2) to effectively adapt the learned feature representation from a large scale classification task to scene categorization.

**Algorithm 1** Operations Involved in Spatially Unstructured Layer

**Input:** Feature map : \( F \in \mathbb{M}_{p \times q \times r \times s}(\mathbb{R}) \), Number of Blocks : \( n \)

// \( F \) is a real valued four dimensional matrix

**Output:** Modified feature map (\( F_m \))

\[
\ell = \left\lceil \frac{\sqrt{n}}{2} \right\rceil \\
\text{hpts} \leftarrow (\ell + 1) \text{ linearly spaced points in range}[1 : p] \\
\text{hpts}[\text{end}] + 1 \\
w_{\text{pts}} = \text{hpts} \\
\text{for} \ \forall i \in [1 : \text{length}(\text{hpts}) - 1] \text{ do} \text{ for} \ \forall j \in [1 : \text{length}(w_{\text{pts}}) - 1] \text{ do} \\
F_{\text{tmp}} = F[\text{hpts}(i) : \text{hpts}(i + 1) - 1, \text{wpts}(j) : \text{wpts}(j + 1) - 1, \ddots] \\
F_{\text{tmp}} = F_{\text{tmp}}\left[\frac{\text{rows}(F_{\text{tmp}})}{2} : \text{end}, \ddots, \ddots; F_{\text{tmp}}(1 : \frac{\text{rows}(F_{\text{tmp}})}{2}, \ddots, \ddots)\right] \\
F_{\text{tmp}} = F_{\text{tmp}}(\ddots, \left[\frac{\text{cols}(F_{\text{tmp}})}{2} : \text{end}, \ddots; F_{\text{tmp}}(1 : \frac{\text{cols}(F_{\text{tmp}})}{2}, \ddots, \ddots)\right] \\
F_m[\text{hpts}(i) : \text{hpts}(i + 1) - 1, \text{wpts}(j) : \text{wpts}(j + 1) - 1, \ddots] = F_{\text{tmp}} \\
\text{return } \{R\}
\]

Fig. 4: (left to right) Original image and the spatially unstructured versions with 2\(^{16}\), 2\(^{14}\) and 2\(^2\) blocks respectively.

**B. Training CNNs for Indoor Scenes**

Deep CNNs have demonstrated exceptional feature representation capabilities for the classification and detection tasks (e.g., see ILSVRC’14 Results [32]). Training deep CNNs however requires a large amount of data since the number of parameters to be learnt is huge. The requirement of a large amount of training data makes the training of CNNs infeasible where only a limited amount of annotated training data is available. In this paper, we propose to leverage from the image representations learnt on a large scale classification task (such as on ImageNet [32]) and propose a strategy to learn tailored feature representations for indoor scene categorization. An algorithmic description of our proposed strategy is summarized in Algorithm. 2 The details are presented...
We first train our baseline CNN architecture on ImageNet database following the procedure in [14]. Next, we densely extract mid-level image patches from our scene classification training data and represent them in terms of the convolutional activations of the trained baseline network. The output of the last convolution layer followed by ReLU non-linearity is considered as a feature representation of the extracted patches. These feature representations ($F$) will be used to train our TransferNet.

As depicted in Fig 3, our TransferNet consists of three hidden layers (with 4096 neurons each) and an output layer, whose number of neurons are equal to the number of classes in the target dataset (e.g., indoor scenes dataset). TransferNet is trained on convolutional feature representations ($F$) of mid-level patches of the scene classification dataset. Specifically, the input to TransferNet are the feature representations ($F$) of the patches and the outputs are their corresponding class labels. After training TransferNet, we remove all fully connected layers of the baseline CNN and join the trained TransferNet to the last convolutional layer of the baseline CNN. The resulting network then consists of five convolutional layers and four fully connected layers (of the trained TransferNet). This complete network is now fine-tuned on the patches extracted from the training images of the scene classification data. Since the network initialization is quite good (the convolutional layers of the network are initialized from the baseline network trained on imageNet dataset, whereas the fully connected layers are initialized from the trained transferNet), only few epochs are required for the network to converge. Moreover, with a good initialization, it becomes feasible to learn deep CNN’s parameters even with a smaller number of available training images.

Note that the baseline CNN was trained with images from the ImageNet database, where each image predominantly contains one or multiple instances of the same object. In the case of scene categorization, we may deal with multiple distinct objects from a wide range of poses, appearances and scales across different spatial locations. Therefore, in order to incorporate large scale deformations, we train two CNNs: with and without the spatially unstructured layer (learned weights represented by $W$ and $W_{su}$ respectively). These trained CNNs are then used for robust feature representation in Sec. III-D. Below, we first explain our approach in dealing with scale variations.

### Algorithm 2: Training CNNs for indoor scenes

**Input:** Source DB (ImageNet), Target DB (Scene Images)  
**Output:** Learned weights: \{W\}$_{1 \times L}$, \{W$_{su}$\}$_{1 \times L}$

1. Pre-train the CNN on the large-scale Source DB.
2. Feed-forward image patches from target DB to trained CNN.
3. Take feature representations ($F$) from the last convolution layer.
4. Train the ‘TransferNet’ of 4 fully connected layers with $F$ as input and target annotations as output.
5. Append ‘TransferNet’ to the last convolution layer of trained CNN.
6. Fine-tune the complete network with and without the spatially unstructured layer to get \{W\}$_{1 \times L}$ and \{W$_{su}$\}$_{1 \times L}$ respectively.

### C. Pyramid Image Representation

In order to achieve scale invariance, we generate a pyramid of an image at multiple spatial resolutions. However, unlike conventional pyramid generation processes (e.g., Gaussian or Laplacian pyramid) where smoothing and sub-sampling operations are repeatedly applied, we simply resize each image to a set of scales and this may involve up or down sampling. Specifically, we transform each image to three scales, \{0.75 \times D, D, 1.25 \times D\}, where $D$ is the smaller dimension of an image which is set based on the given dataset. At each scale, we densely extract patches which are then encoded in terms of the convolutional activations of the trained CNNs.

### D. Image Representation and Classification

From each of the three images of the pyramidal image representation, we extract multiple overlapping patches of 224 $\times$ 224 using a sliding window. A shift of 32 pixels is used between patches. The extracted image patches are then fed forward to the trained CNNs (both with and without the spatially unstructured layer). The convolutional feature representation of the patches are max-pooled to get a single feature vector representation for the image. This is denoted by A, B and C corresponding to three images of the pyramid in Fig 2. We then max pool the feature representations of these images and generate one single representation of the image for each network (with and without the spatially unstructured layer). The final feature representation is achieved by concatenating these two feature vectors. After encoding the spatial layout and the scale invariant feature representations for the images, the next step is to
perform classification. We use a simple linear Support Vector Machine (SVM) classifier for this purpose.

IV. EXPERIMENTS AND EVALUATION

The proposed approach is validated through extensive experiments on a number of datasets. To this end, we perform experiments on three indoor scene datasets (MIT-67, NYU and Scene-15). Amongst these datasets, MIT-67 is the largest dataset for indoor scene classification. The dataset is quite challenging since images of many classes are similar in appearance and thus hard to classify (see Fig. 8). Apart from indoor scene classification, we further validate our approach for two other tasks i.e., event and object datasets (Graz-02 and Sports-8). Below (Sec. IV-A), we first present a brief description about each of the datasets and the adopted experimental protocols. We then present our experimental results along with a comparison with existing state of the art in Sec. IV-B. An ablative analysis to study the individual effect of each component on the proposed method is also presented in Sec. IV-B.

A. Datasets

The MIT-67 Dataset contains a total of 15620 images of 67 indoor scene classes. For our experiments, we follow the standard evaluation protocol in [30]. Specifically, 100 images per class are considered, out of which 80 are used for training and the remaining 20 are used for testing. We therefore have a total of 5360 and 1340 images for training and testing respectively.

The 15 Category Scene Dataset contains images of 15 urban and natural scene classes. The number of images for each scene class in the dataset ranges from 200-400. For performance evaluation and comparison with existing state of the art, we follow the standard evaluation protocol in [16], where 100 images per class are selected for training and the rest for testing.

The NYU v1 Indoor Scene Dataset contains a total of 2347 images belonging to 7 indoor scene categories. We follow the evaluation protocol described in [35] and use the first 60% of the images of each class for training and the last 40% images for testing.

The Inria Graz 02 Dataset contains a total of 1096 images of three classes (bikes, cars and people). The images of this dataset exhibit a wide range of appearance variations in the form of heavy clutter, occlusions and pose changes. The evaluation protocol defined in [24] is used in our experiments. Specifically, the training and testing splits are generated by considering the first 150 odd images for training and the first 150 even images for testing.

The UIUC Sports Event Dataset contains 1574 images of 8 sports event categories. Following the protocol defined in [17], we used 70 and 60 randomly sampled images per category for training and testing respectively.

B. Results and Analysis

The quantitative results of the proposed method in terms of classification rates for the task of indoor scene categorization are presented in Tables I, III and V. A comparison with the existing state of the art techniques shows that the proposed method consistently achieves a superior performance on all datasets. We also evaluate the proposed method for the tasks of sports events and highly occluded object classification (Tables II and IV). The results show that the proposed method achieves very high classification rates. The experimental results suggest that the gain in performance of our method is more significant and pronounced for the MIT-67, Scene-15, Graz-02 and Sports-8 datasets. The confusion matrices showing the class wise accuracies of Scene-15, Sports-8 and NYU datasets are presented in Fig. 6. The confusion matrix for the MIT-67 scene dataset is given in Fig. 5. It can be noted that all the confusion matrices have a very strong diagonal (Fig. 5 and Fig. 6). The majority of the confused testing samples belong to very closely related classes e.g., living room is confused with bedroom, office with computer-room, coast with open-country and croquet with bocce.

The superior performance of our method is attributed to its ability to handle a large spatial layout (through the introduction of the spatially unstructured layer in our modified CNN architecture) and scale variations (achieved by the proposed pyramidal image representation). Further, our method is based on deep convolutional representations, which have recently shown to be superior in performance over shallow handcrafted feature representations [9], [31], [32]. A number of compared methods are based upon mid-level feature representations (e.g., [4], [12], [38]). Our results show that our proposed method achieves superior performance over these methods. It should be noted that in contrast to existing mid-level feature representation based methods (whose main focus is on the automatic discovery of discriminative mid-level patches) our method simply densely extracts mid-level patches from uniform locations across an image. This is computationally very efficient since we do not need to devise patch selection and sorting strategies. Further, our dense patch extraction is similar to dense keypoint extraction, which has shown a comparable performance with sophisticated keypoint extraction methods over a number of classification tasks.
Fig. 5: Confusion Matrix for the MIT 67 Indoor Scenes Dataset. Figure best seen in color.

Fig. 6: Confusion matrices for Scene-15, Sports-8 and NYU scene classification datasets. Figure best seen in color.

| Method                          | Accuracy (%) | Method                          | Accuracy (%) |
|---------------------------------|--------------|---------------------------------|--------------|
| ROI + GIST [CVPR’09] [30]       | 26.1         | OTC [ECCV’14] [23]              | 47.3         |
| MM-Scene [NIPS’10] [50]         | 28.3         | Discriminative Patches [ECCV’12] [37] | 49.4         |
| SPM [CVPR’06] [16]              | 34.4         | ISPR [CVPR’14] [21]             | 50.1         |
| Object Bank [NIPS’10] [18]      | 37.6         | D-Parts [ICCV’13] [58]          | 51.4         |
| RBoW [CVPR’12] [29]             | 37.9         | VC + VQ [CVPR’13] [20]          | 52.3         |
| Weakly Supervised DPM [ICCV’11] [28] | 43.1         | IFV [CVPR’13] [12]              | 60.8         |
| SPM [ECCV’12] [15]              | 44.0         | MLRep [NIPS’13] [4]             | 64.0         |
| LPR-LIN [ECCV’12] [13]          | 44.8         | CNN-MOP [ECCV’14] [7]           | 68.9         |
| BoP [CVPR’13] [12]              | 46.1         | CNNaug-SVM [CVPRw’14] [31]      | 69.0         |
| Hybrid Parts + GIST + SP [ECCV’12] [49] | 47.2         | Proposed SICA                   | 71.2         |

TABLE I: Mean accuracy on the MIT-67 indoor scenes dataset.
UIUC Sports-8 Dataset

| Method                  | Accuracy (%) |
|-------------------------|--------------|
| GIST-color [IJCV’01]    | 70.7         |
| MM-Scene [NIPS’10]      | 71.7         |
| Graphical Model [IJCV’07] | 73.4       |
| Object Bank [NIPS’10]   | 76.3         |
| Object Attributes [ECCV’12] | 77.9      |
| CENTRIST [PAMI’11]      | 78.2         |
| RSP [ECCV’12]           | 79.6         |
| SPM [CVPR’06]           | 81.8         |
| SPM-MA [ECCV’12]        | 83.0         |
| Classenses [ECCV’10]    | 84.2         |
| HIK [ICCV’09]           | 84.2         |
| LScSPM [CVPR’10]        | 85.3         |
| LPR-RBF [ECCV’12]       | 86.2         |
| Hybrid Parts + GIST + SP [ECCV’12] | 87.2       |
| LCSR [CVPR’12]          | 87.2         |
| VC + VQ [CVPR’13]       | 88.4         |
| IFV                      | 90.8         |
| ISPR [CVPR’14]          | 89.5         |
| Proposed S^ICA          | 98.7         |

TABLE II: Mean accuracy on the UIUC Sports-8 dataset.

NYU Indoor Scenes Dataset

| Method                  | Accuracy (%) |
|-------------------------|--------------|
| BoW-SIFT [ICCV’11]      | 55.2         |
| RGB-LLC [TC’13]         | 78.1         |
| RGB-LLC-RPSL [TC’13]    | 79.5         |
| Proposed S^ICA          | 81.2         |

TABLE III: Mean Accuracy for the NYU v1 dataset.

Graz-02 Dataset

| Method                  | Cars  | People | Bikes | Overall |
|-------------------------|-------|--------|-------|---------|
| OLB [SCIA’05]           | 70.7  | 81.0   | 76.5  | 76.1    |
| VQ [ICCV’07]            | 80.2  | 85.2   | 89.5  | 85.0    |
| ERC-F [PAMI’08]         | 79.9  | -      | 84.4  | 82.1    |
| TSD-IB [BMVC’11]        | 87.5  | 85.3   | 91.2  | 88.0    |
| TSD-k [BMVC’11]         | 84.8  | 87.3   | 90.7  | 87.6    |
| Proposed S^ICA          | 98.7  | 97.7   | 97.7  | 98.0    |

TABLE IV: Equal Error Rates (EER) on Graz-02 dataset.

The contributions of the extracted mid-level patches towards a correct classification are shown in the form of heat maps for some example images in Fig 7. It can be seen that our proposed spatial layout and scale invariant convolutional activations based feature descriptor gives automatically more importance to the meaningful and information rich parts of an image.

The actual and predicted labels of some miss-classified images from MIT-67 dataset are shown in Fig 8. Note the extremely challenging nature of the images in the presence of high inter-class similarities. Some of the classes are very challenging and there is no visual indication to determine the actual label. It can be seen that the miss-classified images belong to highly confusing and very similar looking scene types. For example, the image of inside subway is miss-classified as inside bus, library as bookstore, movie theater as auditorium and office as classroom.

An ablative analysis to assess the effect of each individual component of the proposed technique towards the overall performance is presented in Table VI. Specifically, the contributions of the proposed spatially unstructured layer, pyramid image representation, training of the CNN on the target dataset and pooling (mean pooling and max pooling) are investigated. In order to investigate a specific component of the proposed method, we only modify (add or remove) that part, while the rest of the pipeline is kept fixed. The experimental results in Table VI show that the feature representations from trained CNNs with and without the spatially unstructured layer complement each other and achieve the best performance. Furthermore, the proposed pyramidal image representation also contributes significantly towards the performance improvement of the proposed method. Our proposed strategy to adapt a deep CNN (trained on a large scale classification task) for scene categorization also proves to be very effective and it results in a significant performance improvement. Amongst the pooling strategies, max pooling provides a superior performance compared with mean pooling.

V. CONCLUSION

This paper proposed a novel approach to handle the large scale deformations caused by spatial layout and scale variations in indoor scenes. A pyramidal image representation has been contrived to deal with scale variations. A modified Convolutional Neural Network Architecture with an added layer has been introduced to deal with the variations caused by spatial layout changes. In order to feasibly train a CNN on tasks with only a limited annotated training dataset, the paper proposed an efficient strategy which conveniently transfers learning
TABLE V: Mean accuracy on the 15 Category scene dataset. Comparisons with the previous best techniques are also shown.

![Image of misclassified images from MIT-67 indoor scenes dataset. Actual and predicted labels of each image are given. Images from highly similar looking classes are confused amongst each other. For example, the proposed method misclassifies library as bookstore, office as classroom and inside subway as inside bus.](image)

Fig. 8: Some examples of misclassified images from MIT-67 indoor scenes dataset. Actual and predicted labels of each image are given. Images from highly similar looking classes are confused amongst each other. For example, the proposed method misclassifies library as bookstore, office as classroom and inside subway as inside bus.

Baseline CNN (w/o Spatially Unstructured layer) 65.4%  
Modified CNN (with Spatially Unstructured layer) 65.9%  
Baseline CNN + Modified CNN 71.2%  
w/o pyramidal representation 68.5%  
with pyramidal representation 71.2%  
CNN trained on imageNet 67.3%  
CNN trained on imageNet+MIT-67 71.2%  
Mean-pooling 65.7%  
Max-pooling 71.2%  

TABLE VI: Ablative analysis on MIT-67 dataset. The joint feature representations from baseline and modified CNNs gives the best performance. The proposed pyramidal image representation results in a significant performance boost.

from a large scale dataset. A robust feature representation of an image is then achieved by extracting mid-level patches and encoding them in terms of the convolutional activations of the trained networks. Leveraging on the proposed spatial layout and scale invariant image representation, state of the art classification performance has been achieved by using a simple linear SVM classifier.

ACKNOWLEDGEMENTS

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 DP120102960. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

REFERENCES

[1] A. Bosch, A. Zisserman, and X. Muoz, “Scene classification using a hybrid generative/discriminative approach,” *PAMI*, 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 *CVPR*. 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 *BMVC*, 2014.
[50] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing. “Large margin learning of upstream scene understanding models,” in *NIPS*, 2010, pp. 2586–2594.