Machine Vision and Applications (2021) 32:52
https://doi.org/10.1007/s00138-021-01168-8

ORIGINAL PAPER

Recognition of varying size scene images using semantic analysis of deep activation maps

Shikha Gupta[1] · **A. D. Dileep**[1] · **Veena Thenkanidiyoor**[2]

Received: 18 April 2020 / Revised: 21 September 2020 / Accepted: 12 January 2021 / Published online: 1 March 2021

Abstract
Understanding the complex semantic structure of scene images requires mapping the image from pixel space to high-level semantic space. In semantic space, a scene image is represented by the posterior probabilities of concepts (e.g., 'car,' 'chair,' 'window,' etc.) present in it and such representation is known as semantic multinomial (SMN) representation. SMN generation requires a concept annotated dataset for concept modeling which is infeasible to generate manually due to the large size of databases. To tackle this issue, we propose a novel approach of building the concept model via pseudo-concepts. Pseudo-concept acts as a proxy for the actual concept and gives the cue for its presence instead of actual identity. We propose to use filter responses from deeper convolutional layers of convolutional neural networks (CNNs) as pseudo-concepts, as filters in deeper convolutional layers are trained for different semantic concepts. Most of the prior work considers fixed-size ($\approx 227 \times 227$) images for semantic analysis which suppresses many concepts present in the images. In this work, we preserve the true-concept structure in images by passing in their original resolution to convolutional layers of CNNs. We further propose to prune the non-prominent pseudo-concepts, group the similar one using kernel clustering and later model them using a dynamic-based support vector machine. We demonstrate that resulting SMN representation indeed captures the semantic concepts better and results in state-of-the-art classification accuracy on varying size scene image datasets such as MIT67 and SUN397.

Keywords Varying size scene images · Scene representation · Semantic multinomial representation · Concept modeling · Pseudo-concept · Scene recognition

1 Introduction

Scene images are composed of many fine and localized semantic concepts (e.g., 'car,' 'chair,' 'book,' 'sky,' etc.) which collectively form abstract semantic entities such as 'coast,' 'bookstore,' and 'dining-room' [20]. Scene image recognition is one of the challenging tasks in computer vision due to multiple factors: (i) high intra-class variability (ii) high inter-class similarity (iii) overlapped complex semantic structure of concepts (iv) large diversity of content of scene images and (v) varying sizes of images [45]. Efficient and effective scene recognition approaches are important as they are the basic task needed in many of the real-world applications such as event recognition, indoor/outdoor robot navigation, automatic camera mode selection, automatic tour guidance, and so on. In the past two decades, a variety of scene recognition approaches have been proposed [2,11,24,39,47]. These approaches focus on either generating the discriminating scene representations or building an effective classifier. Due to the complex semantic nature of scene images, generating the discriminating and descriptive representation is an important task. Scene image representations are broadly based on two type of features, i.e., low-level local image features [29] and learning-based features [24]. Low-level features capture only the local semantics of a scene image whereas high-level learned features capture the global information. However, both the representations fail to capture and quantify the complex concepts information present in a scene. A more suitable representation for scene images is

✉ Shikha Gupta
shikha_g@students.iitmandi.ac.in

A. D. Dileep
addileep@iitmandi.ac.in

Veena Thenkanidiyoor
veenat@nitgoa.ac.in

[1] School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Kamand, H.P. 175001, India

[2] Department of Computer Science and Engineering, National Institute of Technology Goa, Ponda, Goa 403401, India

semantic concept-based representation which quantifies the presence of different concepts in an image [33,39].

Semantic concept-based representation of an image is a vector of concept probabilities corresponding to its semantic content. An approach for obtaining such a representation is known as semantic scene modeling that involves identifying the semantic concepts present in the images and quantifying the extent of the presence of such concepts [43]. Important issues in semantic modeling are to decide (i) what semantic concepts to be quantified and (ii) which method to be used for this quantification. One of the approaches to semantic scene modeling involves representing a scene image using semantic multinomial (SMN) representation [7,22,32,39]. SMN representation, $\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \ldots, \pi_{\mathcal{C}}]^{\mathsf{T}}$ can be considered as a transformation that maps an image $\mathcal{I}$ onto a point in a $\mathcal{C}$-dimensional probability simplex, where each element of π corresponds to the posterior probability of semantic concepts [32]. The posterior probabilities can be computed by building suitable models for the concepts using concept specific features from all the images in the database. To generate such features, it is necessary to have images with true-concept labels. Explicit pixel-wise annotation and feature extraction from concept regions are impractical options [9,33]. So in the absence of true-concept annotated images prominent concept features can be considered as cues to concepts without their true identity. Such features are called as pseudo-concept features. The '*pseudo-concepts*,' give cues for the presence of different concepts. However, they do not give the identity of actual concepts in an image. In [14], clusters of the low-level local image feature corresponding to the database images are considered as cues for pseudo-concepts. Since the local feature vectors are not able to capture the complex semantic structure of scene images effectively, we propose to consider learned features to build the concept models for generating SMN representation of scene images.

In the last few years, convolutional neural networks (CNNs) have achieved great success in vision and other related domains [41,51]. Due to the ability of learning complex visual structures of scene images, CNN-based representations are more effective in comparison with low-level local image features [7,8]. Features extracted from fully connected (FC) layers of a CNN correspond to the global representation of an image but they lack spatial and geometric invariance properties [7,21]. However, convolutional (CONV) layer filters of CNNs preserve spatial structures of concepts. Initial CONV layers capture local structures like the blob, curve, edge, point, etc., while deeper CONV layers capture meaningful semantic information that contributes to concept description [10,49]. This is also illustrated in Fig. 1, where maximally activated image regions from a particular filter are shown in each row. It can be seen from Fig. 1 that images in a particular row correspond to a semantic concept. For, e.g., filter #19 trigger for semantic concept swimming pool, filter #33 for monument structure, and so on. This shows that a filter in a CONV layer responds to a distinct geometric structure corresponding to a semantic concept. Therefore, different activation maps (filter responses) of deeper CONV layers can be thought of as the indicator of various concepts present in the images.

Fig. 1 Visualization of the maximally activated image regions for few filters (19, 33, …, 228) of CONV5 layer of Places365-AlexNet [24] on images of SUN397 scene dataset [45]. Activated regions on same row correspond to a concept

Conventionally deep CNNs take the fixed size of images as input and results in fixed-size feature representation for classification task [24,51]. Resizing varying original resolution images to fixed-size results in loss of concept information before feature extraction [11,18]. This loss is not very crucial for object recognition tasks as object images consist of a single object with uniform artificial background whereas many objects and concepts co-occur in scene images. Also, the scene images may be of varying size with high resolution. Resizing such a high-resolution image to a smaller size results in loss of semantic information corresponding to concepts. In this work, we address this issue by considering the actual resolution images for semantic analysis and data generation for concept model building.

It is also observed that some of the filters in a deeper CONV layer of a pre-trained CNN are not trained for a meaningful semantic structure and hence non-prominent [1]. Other important observation is that some filters are trained

for similar concepts. In this work, we propose an approach to prune the non-prominent filters using a threshold-based approach. We also propose to group the similar filter responses using kernel-based clustering of pseudo-concepts class data. Selected and grouped pseudo-concepts class data is further used for building the pseudo-concept models using kernel-based SVM. To summarize, main contributions of this work are as follows:

- Actual resolution images (without cropping or re-scaling to fixed size) are considered for concept modeling to avoid loss of information.
- A strategy is proposed in which varying size filter responses act as cues for pseudo-concepts in the absence of true-concepts label data.
- Procedure to prune the non-prominent, non-discriminative pseudo-concepts and group the similar one using dynamic kernel-based clustering are proposed.
- Pseudo-concept modeling using modified deep spatial pyramid match kernel (M-DSPMK)-based SVM is proposed to handle varying size activation maps.
- Novel deep CNN-based SMN representation is proposed for varying size scene image recognition.

The remaining paper is organized as follows: Sect. 2 briefly reviews some of the related approaches for scene image representations and recognition. The proposed framework for generation of deep SMN representation with motivation of varying size images is discussed in Sect. 3. In Sect. 4, the experimental studies on scene image recognition are presented. The conclusion and discussion on future possibilities are presented in Sect. 6.

2 Literature review

Scene image recognition has been in the focus of the vision community since the past decades. Many researchers have proposed different types of approaches with the motive of achieving high scene recognition accuracy [2–4,11]. Few focus on holistic spatial envelope properties (degree of naturalness, degree of roughness, and so on) of scene images, others give more importance to local variation in images, while in the last few years focus is completely shifted to learned CNN-based approaches [50]. In this work, we argue that neither handcrafted local image features nor CNN-based global learned features are appropriate for original varying size concept rich images. Alternatively, mid-level intermediate representation encoded from semantic analysis of images captures the discriminative characteristic better.

State-of-the-art intermediate semantic representation of scene images is obtained using either semantic analysis of local image features [9,25,32] or learned features [3,7,23,27,39,44] as a base feature. The work in [32] represents the images by posterior probabilities of appearance-based classifiers which are built using bag-of-visual word representation. The work in [9] also uses low-level local image features along with pattern mining approaches to mine relevant visual primitives of images as a bag of frequent local histograms (FLHs). The work in [25] constructs object bank (OB) representation using object filter responses by considering deformable part-based model as filters, i.e., object detector.

Later, with the advancement of CNNs, local image features as base features are replaced by learned features for generating intermediate scene representation. The work in [7] proposed to represent a scene image using a bag-of-semantics (BoS) representation. Here, scene images are first converted into different scale patch images and then represented by softmax layer output of pre-trained CNN. BoS representation of patch images is further encoded using Fisher vector (FV) embedding and named as semantic FV. The work in [44] used region proposal technique to generate potential patches containing objects, further extract CNN-based features of these patches and harvest discriminative visual objects and part-based representation. The work in [23] proposed to represent images by CNN-based features of mid-level patches using codebooks. The work in [27] proposed to use CNN-based features of image patches and applied pattern mining-based approach to obtain Bag-of-Elements (BoE) and Bag-of-Patterns (BoP)-based representation. The work in [39] proposed a semantic representation using patch features of multi-scale CNN for context modeling with Markov random field framework. The work in [3] proposed to represent image descriptors with the occurrence probabilities of the discriminative object in patches and called their framework as a semantic descriptor with objectness (SDO). In all these approaches, either images patch features are considered to build object classifier or complex model building procedure is used with different encoding techniques. In contrast, we propose an effective and efficient framework to build concept models for varying size scene images. We consider images in their original size for generating the cues for true-concepts. In the absence of true-concept annotated data, we propose concept modeling via pseudo-concepts and generate deep CNN-based SMN representation.

The idea of pseudo-concepts for concept model building and obtaining SMN representation using handcrafted local image features are proposed in [14] for smaller datasets, (i.e., MIT8 scene [30] and Vogel–Schiele [43]). In [14], pseudo-concept data is generated using clusters of local feature vectors. Disadvantages of the approach proposed in [14] are (1) concept models are built using features from the complete image instead of concept specific features, as single image comprises of multiple concepts, and (2) handcrafted features are used for building concept models which are local descriptors and do not capture much semantic informa-

tion. The approach proposed was not explored for large-scale datasets. To overcome these limitations, we propose CNN-based pseudo-concepts using the filter responses of original resolution images. The details of the proposed framework are discussed in the next section.

3 Proposed framework for recognition of varying size scene images

In this section, we first present the motivation of using scene images in their original size for varying size scene recognition framework. Further, the proposed concept modeling and deep SMN representation generation procedures are described in subsequent subsections.

3.1 Motivation for using original varying size images for semantic analysis

In real-world scene datasets, the spatial resolution of images approximately varies from 10^4 to 10^8 pixels. This is also illustrated in the scatter plots shown in Fig. 2 which are drawn for large-scale scene image datasets such as MIT67 [31] and SUN397 [45]. Each point in these plots denotes the dimension of an image (in terms of the number of pixels) present in that dataset. It can be observed from the plots that image size varies significantly across the datasets and resizing them all to a fixed small size results in loss of concept information. Hence, there is a need for a scene recognition framework that considers varying size original resolution images.

Also, existing state-of-the-art deep CNNs require fixed-size ($\approx 227\times227$) input images. This requirement may affect the recognition accuracy for the scene images of an arbitrary size/resolution [18]. This is also illustrated in Fig. 3, here, we can visualize the difference in information content of activation maps of the CONV1 layer computed from Places365-AlexNet [51] when the image is resized to '227×227' versus original size '2400×1594.' We observe that fine details and spatial concept layout are preserved and forwarded to later convolution layers when the image is passed to CNN in its original size instead of a fixed reduced size. To avoid such loss, we consider original size images as input to the CNN for concept modeling. Following are the main benefits of doing so:

- Finer semantic concept information is preserved till the last CONV layers of the network hierarchy.
- It avoids the need for cropping or warping of an image in the beginning.
- It results in better features for effective concept modeling.

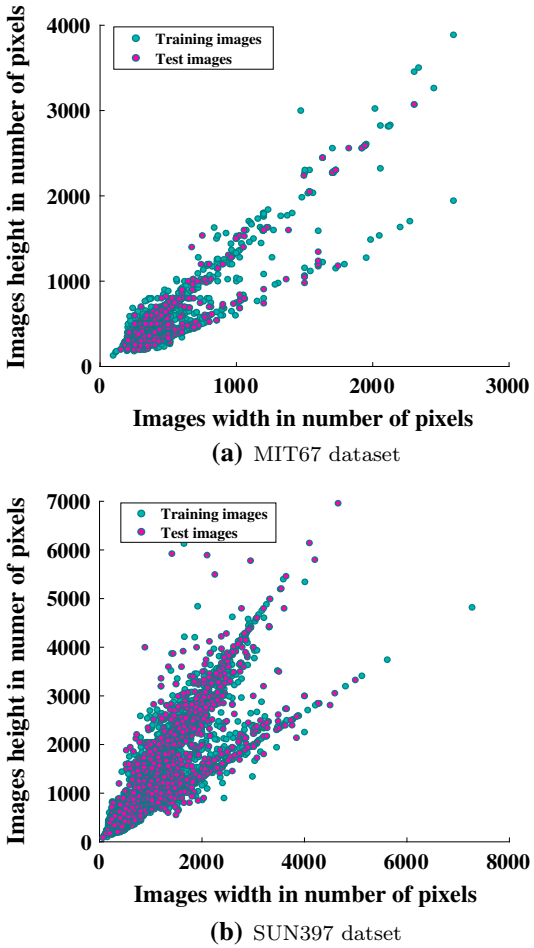

Fig. 2 Scatter plot of original size of images in **a** MIT67 dataset and **b** SUN397 dataset

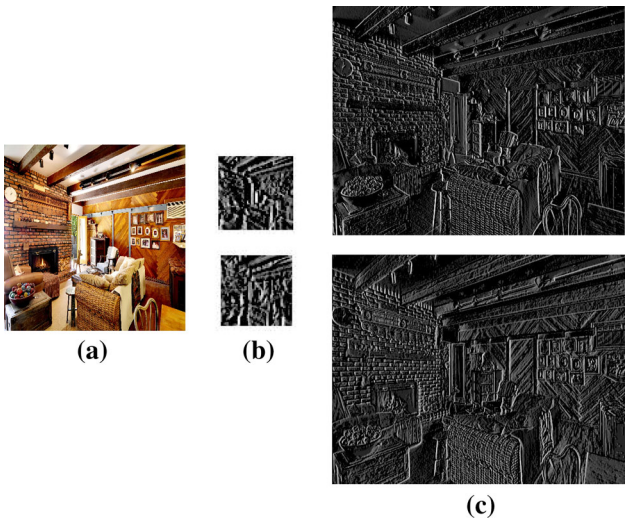

Fig. 3 Visualization of activation maps of the CONV1 layer using Places-AlexNet. **a** Original image, **b** Activation maps of size (55×55) at CONV-1 layer from filter #4 and #44 (when the image is passed with reduced size, i.e., 227×227), **c** Activation maps of size (598×396) at CONV-1 layer from filter #4 and #44 (when the image is passed in its actual size, i.e., 2400×1594)

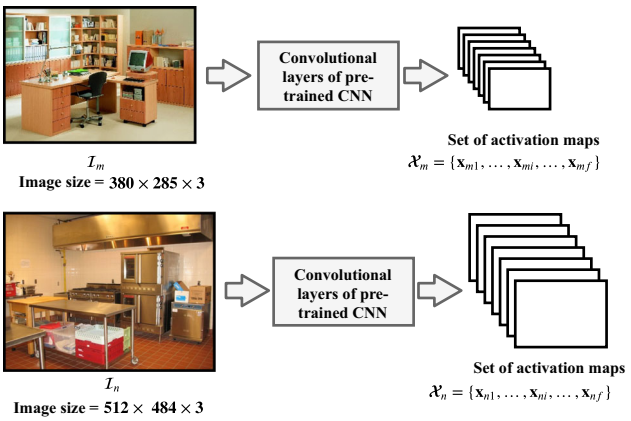

Fig. 4 Illustration of two different resolution images represented as varying size sets of activation maps computed from last CONV layer of pre-trained CNN

3.2 Concept model building using varying size activation maps

The core part of SMN representation generation is building the concept models which requires a concept annotated dataset to generate concept specific features. However, it is impossible to manually annotate concepts present in images at the pixel level. In the absence of true-concept labeled data, we propose to use the response of deeper CONV layer filters as concept cues. Since information regarding the filter identity, i.e., which filter is learning what concept is not known during the training of CNNs, we cannot infer the true-concept identity of the particular filter from its generated activation maps. Moreover, we can only visualize the filter responses using different visualization techniques [1,48]. Figure 1 shows maximally activated image regions of SUN397 images for a few filters (19, 33, …, 228) of the CONV5 layer of Places365-AlexNet [51] using deep-visualization toolbox [48]. It is seen that maximally activated regions from a distinct filter have similar semantic properties and correspond to a specific concept. This enables us to perceive these filter responses as indications for concepts and we have called them as pseudo-concepts. Furthermore, for preserving the spatial concept structure of scenes, we consider original size images which results in varying size activation maps for concept modeling. Figure 4 illustrates the image representations as varying size sets of activation maps, i.e., $\boldsymbol{\mathcal{X}}_m$ and $\boldsymbol{\mathcal{X}}_n$ which are obtained when two images $\mathcal{I}_m$ and $\mathcal{I}_n$ of size '400×285×3' and '512×384×3,' respectively, are passed to CONV layers of pre-trained CNN. It can be observed that the size of all activation maps corresponding to an image is same and that depends upon the size of the input image. Figure 5 shows the block diagram corresponding to the proposed approach to varying size scene recognition and the detailed procedure is given in the following subsections.

3.2.1 Selection of pseudo-concepts using varying size activation maps

In the proposed framework, the filters of deeper CONV layer of a pre-trained CNN are considered as pseudo-concept detector and corresponding responses as features for the pseudo-concept classes. It is determined that few of the filters are not capturing prominent concepts of the given dataset [10] and results in non-discriminative features. So, there is a need to select prominent filters for the effective pseudo-concept model building. The detailed procedure using original resolution images is given in Algorithm 1. In the algorithm, $\mathcal{K}$ pseudo-concepts are selected from f, where f is the total number of filters in a chosen deep CONV layer or an initial number of pseudo-concepts.

Given a dataset of varying size original resolution concept rich images without concept labels, $\mathscr{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m, \ldots, \mathcal{I}_M\}$, where $\mathcal{I}_m \in \mathbb{R}^{i_m \times j_m \times 3}$. Every image of dataset is passed through convolutional layers of pre-trained deep CNN and generates $\{\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_m, \ldots, \boldsymbol{\mathcal{X}}_M\}$ varying size sets of deep activation maps corresponding to M database images from chosen CONV layer. Here, $\boldsymbol{\mathcal{X}}_m = \{\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mi}, \ldots, \mathbf{x}_{mf}\}$ is a set of f activation maps with $\mathbf{x}_{mi} \in \mathbb{R}^{sx_m \times sy_m}$, $sx_m \times sy_m$ is the size of an activation map which depend on chosen pre-trained CNN architecture. As discussed, all the activation maps triggered by their respective filters are not prominent. To prune the non-prominent activation maps corresponding to a filter, varying size activation maps are sum pooled and normalized by its size. A score matrix $\mathbf{P} \in \mathbb{R}^{M \times f}$ is formulated using normalized score values. The threshold specific to a filter or pseudo-concept class is computed as average score distribution of that particular filter responses corresponding to all dataset images. A binary matrix $\mathbf{R}$ is generated from $\mathbf{P}$ using filter specific threshold, where each entry $\mathbf{R}[m, i]$ indicate the presence (1) or absence (0) of ith pseudo-concept in mth image.

A count vector $\mathbf{pc} \in \mathbb{R}^f$ is computed, where $\mathbf{pc}[i]$ indicates the total number of activation corresponding to ith pseudo-concept class. Few of the filters get triggered for almost all the images of the dataset and few are for very less. For building the effective pseudo-concept models there is a need of some minimum number of activation in a pseudo-concept class. If a particular filter got activated for almost all the images, then it result in non-discriminative pseudo-concept model. To overcome both the scenarios, all such pseudo-concepts are pruned using two threshold, i.e., lb and ub, which indicates at least and at most features needed in a pseudo-concept class for model building. For building the models, all the activation maps are considered as features for the selected pseudo-concept class whose corresponding value in $\mathbf{R}$ matrix is 1. Let $\boldsymbol{\rho} = \{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k, \ldots, \boldsymbol{\rho}_{\mathcal{K}}\}$ indicates pseudo-concept classes data, where $\boldsymbol{\rho}_k = \{\mathbf{x}_{1k}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{m_k k}\}$ is the kth pseudo-concept class and contain varying size activation

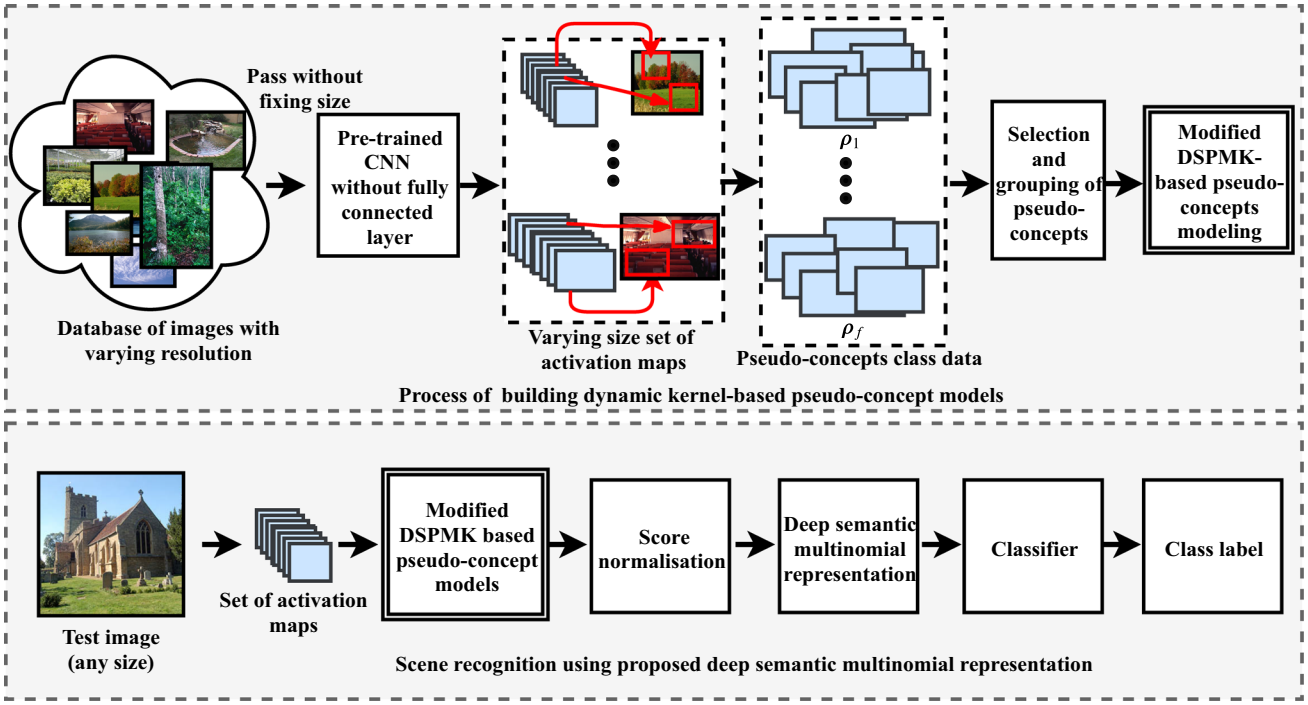

Fig. 5 Block diagram of proposed varying size scene recognition framework. Here, the original resolution images are represented by deep SMN representation which is obtained using pseudo-concept models

maps. Here, $\mathbf{x}_{jk} \in \mathbb{R}^{sx_j \times sy_j}$ and m_k is the total number of varying size activation maps in kth pseudo-concept class. The time complexity of pseudo-concept selection algorithm given the extracted features from pre-trained CNN is $\mathcal{O}(Mf)$. In the next subsection, the procedure for grouping the similar pseudo-concepts using dynamic kernel-based clustering is discussed.

3.2.2 Grouping of similar pseudo-concepts using kernel-based clustering

Using deep-vis toolbox [48], we visualized the maximally activated image region of scene images corresponding to few of the CONV5 layer filters as shown in Fig. 1. It is observed that a particular filter is pre-trained to detect the same semantic concept in the different images of the database. Hence, an intrinsic geometric structure can be assumed. It is also observed that some of the filters respond to similar concepts and will result in redundancy in pseudo-concept modeling (e.g., #19 and #52, #35 and #63 of Fig. 1). To remove such redundancy, we propose to use kernel-based clustering [6] to group the data corresponding to similar varying size pseudo-concept classes. Kernel-based clustering algorithm applies the same approach as k-means clustering but in the transformed space and for computation of distance in transformed space kernel function is used.

From the pseudo-concept selection procedure of Algorithm 1, $\mathcal{K}$ pseudo-concepts class data, i.e., $\{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k \ldots, \boldsymbol{\rho}_{\mathcal{K}}\}$ is obtained, where $\boldsymbol{\rho}_k$ is the data for kth class. A nonlinear mapping $\boldsymbol{\phi}$ is assumed from input feature space (i.e., pseudo-concept data space) to higher-dimensional kernel feature space. In transformed space, pseudo-concept class data is represented using the mapping implicit $\boldsymbol{\phi}$, i.e., $\{\boldsymbol{\phi}(\boldsymbol{\rho}_1), \boldsymbol{\phi}(\boldsymbol{\rho}_2), \ldots, \boldsymbol{\phi}(\boldsymbol{\rho}_{\mathcal{K}})\}$. The dynamic kernel-based clustering procedure [6] is used to group $\mathcal{K}$ pseudo-concepts class data into $\mathcal{C}$ pseudo-concept classes where $\mathcal{C} \leq \mathcal{K}$. In dynamic kernel-based clustering algorithm, inner product in transformed space is computed using kernel function in original space via kernel trick. That is $\boldsymbol{\phi}$ is not computed for $\rho_1^{\mathcal{K}}$ or the cluster means $\boldsymbol{\mu}_1^{\mathcal{C}}$ as algorithm computations are completely depend on kernel evaluations. Few examples of kernel functions are the linear kernel, polynomial kernel, Gaussian kernel, etc. However, all these kernel functions are defined for fixed-length features. Kernel functions designed for varying length patterns are referred to as dynamic kernels [13,17]. Since pseudo-concepts class data is of varying size, we propose modified deep spatial pyramid match kernel (M-DSPMK) as the dynamic kernel to compute the similarity between pseudo-concept classes in transformed feature space. M-DSPMK uses the similar idea as DSPMK proposed in [16,36]. The DSPMK compute the similarity between two different or same size images represented as sets of activation maps. The number of activation maps between two images are same, however, there size are different. The same kernel function is modified further to compute the similarity between varying size pseudo-concept class data and

Algorithm 1 Selection of pseudo-concepts and generation of pseudo-concept class specific data using original resolution images.

Inputs:

(i) Unlabeled varying size image dataset, $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_m, \ldots, \mathcal{I}_M\}$.
(ii) Pre-trained CNN model.
(iii) lb: lower bound on minimum number of feature vectors (activation maps) required in any pseudo-concept class.

Procedure:

1: $\forall\ \mathcal{I}_m \in \mathcal{D}$, extract convolutional layer features using true resolution image, i.e, varying size sets of deep activation maps, $\boldsymbol{\mathcal{X}}_m = \{\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mi}, \ldots, \mathbf{x}_{mf}\}$ from chosen CONV layer of pre-trained CNN where, $\mathbf{x}_{mi} \in \mathbb{R}^{sx_m \times sy_m}$.
2: Generate score matrix, $\mathbf{P}[m,i] = \sum\sum \mathbf{x}_{mi}/(sx_m * sy_m)$, $\forall m = 1, 2, \ldots, M$ and $\forall i = 1, 2, \ldots, f$.
3: $\mathbf{R}[m,i] = \begin{cases} 1, & \text{if } \mathbf{P}[m,i] \geq \frac{1}{M}\sum_{m=1}^{M} \mathbf{P}[m,i] \\ 0, & \text{otherwise} \end{cases}$
4: Initialize k=0, ub =$M/2$ and **pc_index** = $zeros(1,f)$, $\boldsymbol{\rho}_k$={}.
5: $\mathbf{pc}[i] = \sum \mathbf{R}[:,i]$, $\forall i = 1, 2, \ldots, f$
6: **for** i:= 1 **to** f **do**
7: **if** $lb \leq \mathbf{pc}[i] \leq ub$ **then**
8: k=k+1
9: **pc_index**[i]=1
10: **for** m:= 1 **to** M **do**
11: **if** $\mathbf{R}[m,i] \neq 0$ **then**
12: add $\mathbf{x}_{mi}$ in $\boldsymbol{\rho}_k$
13: **end if**
14: **end for**
15: **end if**
16: **end for**
17: $\mathcal{K} = k$ (Total number of selected pseudo-concepts classes)

Outputs:

(i) $\boldsymbol{\rho} = \{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k, \ldots, \boldsymbol{\rho}_{\mathcal{K}}\}$, where $\boldsymbol{\rho}_k = \{\mathbf{x}_{1k}, \mathbf{x}_{2k}, \ldots, \mathbf{x}_{m_k k}\}$ is the set of varying size activation maps corresponding to data of k^{th} pseudo-concept class.
(ii) **pc_index**: index vector corresponding to selected pseudo-concepts.

named as modified DSPMK. In a pseudo-concept class data, each activation map is of different sizes and number of activation maps also varies from one pseudo-concept class to another.

Modified DSPMK function for computation of the similarity between two pseudo-concept classes is given in Algorithm 3. Let $\boldsymbol{\rho}_i = \{\mathbf{x}_{1i}, \ldots, \mathbf{x}_{pi}, \ldots, \mathbf{x}_{m_i i}\}$ and $\boldsymbol{\rho}_j = \{\mathbf{x}_{1j}, \ldots, \mathbf{x}_{qj}, \ldots, \mathbf{x}_{m_j j}\}$ is two pseudo-concept class data with, $\mathbf{x}_{pi} \in \mathbb{R}^{sx_p \times sy_p}$ and $\mathbf{x}_{qj} \in \mathbb{R}^{sx_q \times sy_q}$. Here, m_i is the number of activation maps in the ith pseudo-concept class and m_j is the number of activation maps in the jth pseudo-concept class. Modified DSPMK works by building L spatial pyramid levels ranging from 0, 1 to $L-1$. At each level l, activation maps corresponding to $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ are spatially partitioned into 2^{2l} equal size blocks. Each block is max pooled and results in $\mathbf{X}_i^l \in \mathbb{R}^{m_i 2^{2l} \times 1}$ and $\mathbf{X}_j^l \in \mathbb{R}^{m_j 2^{2l} \times 1}$ vectors. $\mathbf{X}_i^l$ and $\mathbf{X}_j^l$ are ℓ_2 normalized and used to compute the intermediate matching score S_l using Eq. 2 of Algorithm 3. Here, the matching score S_l computed at level l also includes all the matches found at the finer level $l+1$. Therefore, the number of new matches found at level l is given by $S_l - S_{l+1}$ for $l = 0$, …, L-1. The final matching score is computed as a weighted sum of the new matching score at different levels of the spatial pyramid as in Eq. 4 of Algorithm 3. Pictorial illustration of modified DSPMK for L=2 is shown in Fig. 6. Once the prominent pseudo-concepts are identified and similar ones are grouped then, the next step is to build the pseudo-concept models, which is presented in the next subsection.

Algorithm 2 Grouping of similar pseudo-concept classes using dynamic kernel-based clustering.

Inputs:

(i) $\mathcal{K}$: pseudo-concept classes data, i.e., $\{\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_k \ldots, \boldsymbol{\rho}_{\mathcal{K}}\}$, here $\boldsymbol{\rho}_k$ is the k^{th} pseudo-concept class data.
(ii) $\mathcal{C}$: effective number of pseudo-concepts.

Procedure:

1: Consider pseudo-concepts class data as points in higher dimensional space as $\{\phi(\boldsymbol{\rho}_k)\}_{k=1}^{\mathcal{K}}$.
2: Randomly pick $\mathcal{C}$ points as cluster representative in transformed higher dimensional space i.e., $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_j, \ldots, \boldsymbol{\mu}_{\mathcal{C}}\}$.
3: Compute the distance of each pseudo-concept class and cluster center $\boldsymbol{\mu}_j$ in transformed space using below distance measure:

$$\mathcal{D}(\phi(\boldsymbol{\rho}_i), \boldsymbol{\mu}_j) = \left\|\phi(\boldsymbol{\rho}_i) - \boldsymbol{\mu}_j\right\|^2$$
$$= \phi(\boldsymbol{\rho}_i)^{\intercal}\phi(\boldsymbol{\rho}_i) - 2\phi(\boldsymbol{\rho}_i)^{\intercal}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\intercal}\boldsymbol{\mu}_j, \text{ here, } \boldsymbol{\mu}_j = \frac{\sum_{\forall \boldsymbol{\rho}_j \in \boldsymbol{\Omega}_j} \phi(\boldsymbol{\rho}_j)}{|\boldsymbol{\Omega}_j|}$$
$$= K(\boldsymbol{\rho}_i, \boldsymbol{\rho}_i) - 2\frac{\sum_{\forall \boldsymbol{\rho}_j \in \boldsymbol{\Omega}_j} K(\boldsymbol{\rho}_i, \boldsymbol{\rho}_j)}{|\boldsymbol{\Omega}_j|} + \frac{\sum_{\forall \boldsymbol{\rho}_j \in \boldsymbol{\Omega}_j} K(\boldsymbol{\rho}_j, \boldsymbol{\rho}_j)}{|\boldsymbol{\Omega}_j|},$$

here, $\boldsymbol{\Omega}_j$ denote j^{th} cluster, $|\boldsymbol{\Omega}_j|$ indicate the number of pseudo-concepts class in j^{th} cluster and $K()$ indicate the modified DSPMK.

4: Assign pseudo-concept class to that cluster center whose distance is minimum.
5: Recompute the cluster centers as,

$$\boldsymbol{\mu}_j = \frac{\sum_{\forall \boldsymbol{\rho}_j \in \boldsymbol{\Omega}_j} \phi(\boldsymbol{\rho}_j)}{|\boldsymbol{\Omega}_j|} \tag{1}$$

6: Repeat from step 3, until there is no change in cluster center.

Outputs:

(i) $\{\hat{\boldsymbol{\rho}}_1, \ldots, \hat{\boldsymbol{\rho}}_c, \ldots, \hat{\boldsymbol{\rho}}_{\mathcal{C}}\}$, $\mathcal{C}$ pseudo concepts class data after grouping. Here, $\hat{\boldsymbol{\rho}}_c = \{\mathbf{x}_{1c}, \mathbf{x}_{2c}, \ldots, \mathbf{x}_{m_c c}\}$

3.2.3 Pseudo-concept modeling

From the grouping procedure, $\mathcal{C}$ pseudo-concepts class data is obtained, i.e., $\{\hat{\boldsymbol{\rho}}_1, \ldots, \hat{\boldsymbol{\rho}}_c, \ldots, \hat{\boldsymbol{\rho}}_{\mathcal{C}}\}$ for building the concept models. Here, $\hat{\boldsymbol{\rho}}_c = \{\mathbf{x}_{1c}, \mathbf{x}_{2c}, \ldots, \mathbf{x}_{m_c c}\}$ is the cth class data which contain varying size activation maps. For every pseudo-concept class, the modified DSPMK-based SVM model is built that discriminates corresponding pseudo-concept data from the rest of the pseudo-concepts class data. The modified DSPMK as dynamic kernel is used for building the SVM-based model instead of linear ker-

Algorithm 3 Modified deep spatial pyramid match kernel $K_{\text{M-DSPMK}}(\boldsymbol{\rho}_i, \boldsymbol{\rho}_j)$

Inputs:

(i) Pseudo-concepts class data,
$\boldsymbol{\rho}_i = \{\mathbf{x}_{1i}, \ldots, \mathbf{x}_{pi}, \ldots, \mathbf{x}_{m_i i}\}$ where, $\mathbf{x}_{pi} \in \mathbb{R}^{sx_p \times sy_p}$
$\boldsymbol{\rho}_j = \{\mathbf{x}_{1j}, \ldots, \mathbf{x}_{qj}, \ldots, \mathbf{x}_{m_j j}\}$ where, $\mathbf{x}_{qj} \in \mathbb{R}^{sx_q \times sy_q}$
(ii) L: total number of pyramid levels.

Procedure:

1: **for** $l = 0$ **to** $L - 1$ **do**
2: At level l, spatially partitioned each activation map of $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ into 2^{2l} blocks.
3: Apply max pooling over each spatially partitioned block of activation maps and compute $\mathbf{X}_i^l \in \mathbb{R}^{m_i 2^{2l} \times 1}$ and $\mathbf{X}_j^l \in \mathbb{R}^{m_j 2^{2l} \times 1}$ vectors.
4: ℓ_1- normalize the generated feature vectors $\mathbf{X}_i^l$ and $\mathbf{X}_j^l$

$$\widehat{\mathbf{X}}_i^l = \frac{\mathbf{X}_i^l}{\sum_{r=1}^{m_i 2^{2l}} \mathbf{X}_i^l[r]}, \quad \widehat{\mathbf{X}}_j^l = \frac{\mathbf{X}_j^l}{\sum_{s=1}^{m_j 2^{2l}} \mathbf{X}_j^l[s]} \tag{2}$$

5: Compute level-wise matching score using histogram intersection function as

$$S_l = \sum_{r=1}^{m_i 2^{2l}} \sum_{s=1}^{m_j 2^{2l}} min(\widehat{\mathbf{X}}_i^l[r], \widehat{\mathbf{X}}_j^l[s]) \tag{3}$$

6: **end for**
7: Compute final similarity score between $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ using level-wise matching score,

$$K_{\text{M-DSPMK}}(\boldsymbol{\rho}_i, \boldsymbol{\rho}_j) = \sum_{l=0}^{L-2} \frac{1}{2^{(L-l-1)}} (S_l - S_{l+1}) + S_{L-1} \tag{4}$$

Outputs:

(i) $K_{\text{M-DSPMK}}(\boldsymbol{\rho}_i, \boldsymbol{\rho}_j)$.

nel as activation maps are of varying size. Inputs to the modified DSPMK are pair of activation maps of same or different sizes, i.e., $K_{\text{M-DSPMK}}(\mathbf{x}_i, \mathbf{x}_j)$ instead of pair of sets of activation maps. The proposed framework of building the concept model is free from any image label, concept or region annotation, and segmentation. A set of concept rich images with corresponding activation maps are sufficient for pseudo-concept selection and modeling. As an image may contain multiple pseudo-concepts, the process of modeling the pseudo-concepts is considered as weakly supervised.

3.3 Deep SMN representation generation for varying size scene images

To generate SMN representation of original resolution scene image $\mathcal{I}_m$, convolutional layer activation maps set $\boldsymbol{\mathcal{X}}_m = \{\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mi}, \ldots, \mathbf{x}_{mf}\}$ is obtained from a chosen CONV layer of a pre-trained CNN model, where $\mathbf{x}_{mi} \in \mathbb{R}^{sx_m \times sy_m}$. Details of used pre-trained CNN models are discussed in Sect. 4.2. Pruning of activation maps is done by removing non-significant filter responses using **pc_index** (computed in Algorithm 1). Final set of activation maps are given as input to every pseudo-concept model and corresponding output score is computed. The output score s_c, for the cth pseudo-concept model is mapped onto a pseudo-probability using a logistic function as

$$\tilde{P}(c \mid \tilde{\mathbf{X}}_m) = \frac{1}{1 + \exp^{(-\alpha s_c)}} \tag{5}$$

where α is a free parameter that controls the slope of the logistic function. The pseudo-probability value is normalized to obtain the posterior probability value corresponding to pseudo-concept c as follows:

$$\pi_c = P(c \mid \tilde{\mathbf{X}}_m) = \frac{\tilde{P}(c \mid \tilde{\mathbf{X}}_m)}{\sum_{c=1}^{C} \tilde{P}(c \mid \tilde{\mathbf{X}}_m)}. \tag{6}$$

SMN representation for an image $\mathcal{I}_m$ is given as $\boldsymbol{\pi}_m = [\pi_{m1}, \pi_{m2}, \ldots, \pi_{mc}, \ldots, \pi_{mC}]^\intercal$ corresponding to each of the pseudo-concepts.

4 Experimental studies

In this section, we evaluate the effectiveness of proposed deep SMN representation for varying size scene recognition task using different datasets. We also discuss the various parameters and other fine details of the proposed framework.

4.1 Datasets

We evaluate the proposed approach on widely used scene recognition datasets namely Vogel–Schiele (VS) [43], MIT8 scene [30], MIT67 [31] and SUN397 [45].

- *MIT8* scene dataset consists of 8 outdoor scene categories with a total of 2,688 images. Scene categories include 'coast,' 'forest,' 'mountain,' 'open-country,' 'highway,' 'inside-city,' 'tall building' and 'street.' 100 images from every class are randomly selected for training and rest for testing in 5 trials similar to [30]. The average classification accuracy of 5 trials is presented in the result section.
- *Vogel–Schiele (VS)* dataset consists of 6 natural scene categories with a total of 700 images. Scene categories include, 'coasts,' 'forests,' 'mountains,' 'open-country,' 'river' and 'sky-clouds.' The average classification accuracy of fivefold is presented by considering 80% of images from each class for training and the rest for testing.
- *MIT67* dataset consists of 5 main categories, i.e., store, home, public spaces, leisure, and working place where

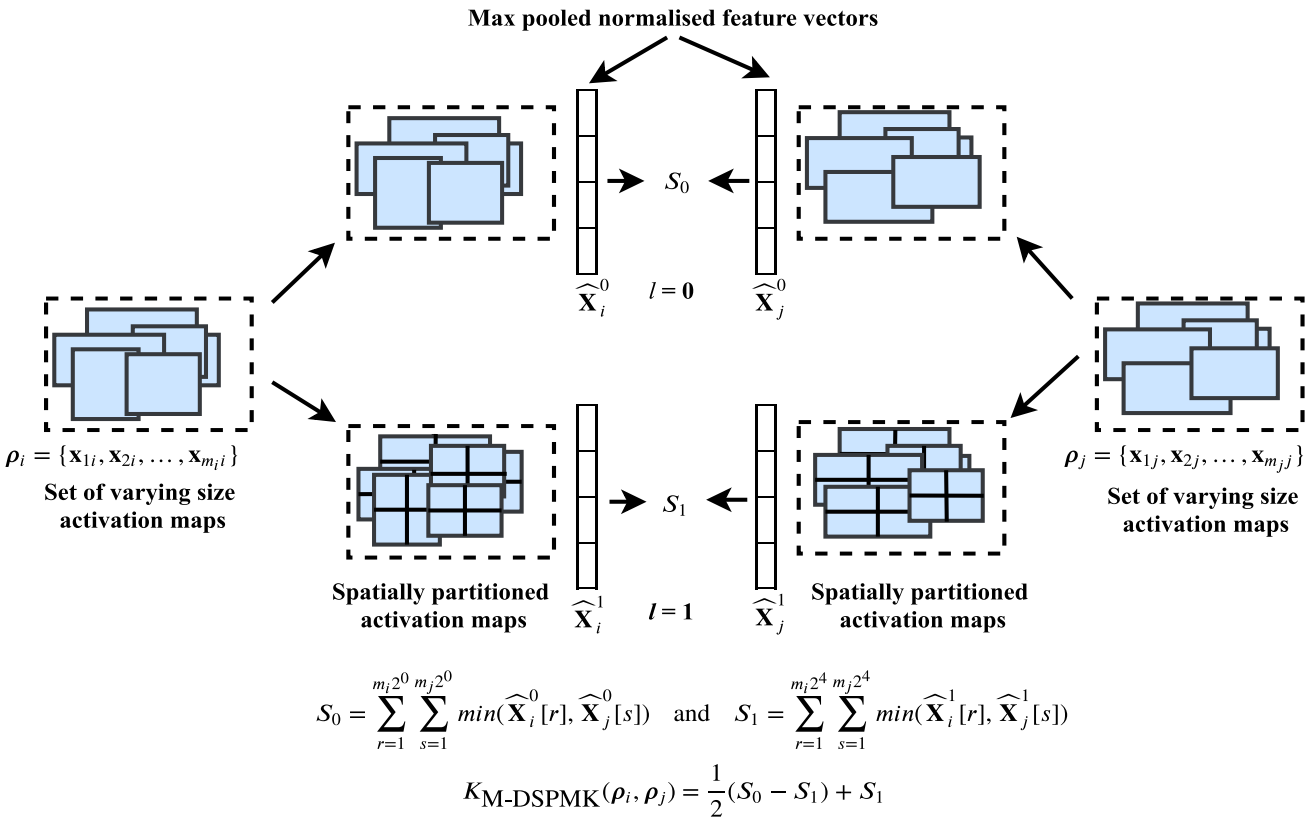

Fig. 6 Illustration of computation of similarity score between two varying size pseudo-concept class data using modified DSPMK with the number of levels as 2

each category is further divided into several sub-categories. Overall there are 67 indoor scene classes with a total of 15,620 images. This dataset is very challenging when compared to outdoor scene datasets as it consists of very confusing indoor classes (like bookstore–library, jewelry shop–shoes shop, and so on). Another main challenge includes the presence of varying size images, i.e., the original resolution of images varies approximately from 10^4 to $12*10^6$ also shown in Fig. 2. Classification results are presented on single split defined in [31] with approximately 80 images per training class and 20 per testing class.

– *SUN397* dataset is large-scale dataset consists of 397 classes of outdoor, indoor, and urban scene images with varying in size. Resolution of images in this dataset varies from 10^3 to 10^7 as shown in Fig. 2. It consists of at least 100 images per class. Training and testing splits are fixed and publicly available [45]. Each split contains 50 training and 50 testing images per category. For the scene image recognition task, we used the first three splits and the results are presented as average classification accuracy for 3 splits.

4.2 Experimental setup

In our studies, we have utilized four different CNN architectures namely AlexNet [24], GoogLeNet [41], VGGNet16 [37], and ResNet152 [19]. These architectures are pre-trained on ImageNet [5] (i.e., object-centric) and Places-(205 & 365) [51] (i.e., scene-centric) databases. Publicly available (Caffe Model zoo[1]) pre-trained weights are used without any fine-tuning. Pre-trained CNNs are used without its FC layers as CONV layers do not restrict the input image size and also preserve the semantic structure of images. We consider deep activation maps of last CONV (LC) and second-to-last CONV (SLC) layers for the selection of pseudo-concepts and model building.

The number of filters for different CNN architectures in LC and SLC layers are given in Table 1. Since we have considered the different CNN architectures, the initial number of pseudo-concepts also varies and that depends on the total number of filters in the chosen CONV layer and CNN architecture. We select $\mathcal{C}$ filters as pseudo-concepts detector from f number of filters in chosen CONV layer using Algorithms 1 and 2. The dimension of SMN representation is restricted with number of built pseudo-concept models. The

[1] https://github.com/BVLC/caffe/wiki/Model-Zoo.

Table 1 Number of filters f in LC layer and SLC layer of different CNN architectures considered in this study

	LC layer	SLC layer
AlexNet	256	384
GoogLeNet	1024	832
VGGNet16	512	512
ResNet152	2048	2048

effect of the final number of pseudo-concepts on classification performance is discussed in Sect. 4.6. For building the pseudo-concepts model of a particular dataset, all the training images associated with different classes of that dataset are collectively considered as the database of M images without any class or concept level information. Although any collection of concept rich images are well suitable for the pseudo-concept modeling process. The value of lb in Algorithm 1 (i.e., minimum number of features required in a particular pseudo-concept class) for concept modeling is chosen as 30. The number of spatial pyramid level L is chosen as 2 in Algorithm 3.

4.3 Experimental results and analysis

Tables 2 and 3 compare the classification accuracy of proposed representation for different varying size scene image datasets. Final classifiers are trained using proposed representation and χ^2-kernel [26]-based SVM classifier. An effective number of pseudo-concepts, i.e., $\mathcal{C}$, is chosen empirically and bounded by number of filters in CONV layer. It is observed that the large number of pseudo-concepts are selected for MIT67 and SUN397 datasets in comparison with MIT8 and VS datasets, as these datasets are semantically complex and having a large number of classes. In Tables 2 and 3, best values of $\mathcal{C}$ are given for different CNNs and datasets. It is also observed that from the SLC layer less number of pseudo-concepts are selected compared to the LC layer. This is mainly because the LC layer considers SLC layer output responses as input and learn more distinct concepts. It is seen that the performance of SVM-based classifiers with SMN representation from VGGNet16 and ResNet152 is significantly better than that obtained using other architectures like GoogLeNet and AlexNet. The reason being VGGNet16 and ResNet152 are deeper networks compared to other architectures and it learns the hierarchical representation of visual data more efficiently. From Table 2, it is observed that MIT8 and VS datasets show comparable results with deep SMN representation obtained using various pre-trained CNN architectures. The reason for this behavior is that MIT8 and VS have less number of classes. For MIT67 and SUN397 datasets, an immense difference in results is observed between the SMN representation obtained using AlexNet and ResNet152. Since ResNet152 captures complex semantics in indoor scene images of MIT67 ('children room,' 'kitchen,' 'clothing store,' etc.) and diverse scene images in SUN397 ('airplane cabin,' 'village,' 'street,' etc.) more effectively. Scene classification for MIT67 using deep SMN representation obtained with Places205-VGGNet16 is found to be better than Places365-VGGNet16 as Places205 dataset has a larger number of images per classes in comparison of Places365 and contain mostly indoor scene categories [50,51]. For SUN397, deep SMN representation obtained using Places365-VGGNet16 performs better in comparison with Places205-VGGNet16 as Places365 dataset covers a large number of scene categories with diverse images. ImageNet is an object-centric dataset whereas MIT67 and SUN397 are scene-centric datasets. So for pseudo-concept modeling of scene-centric datasets, activation maps are needed from CNN architecture trained with scene-centric datasets. Due to this reasoning, SMN representation obtained using Places-CNNs performs better for MIT67 and SUN397 datasets.

4.4 Comparison of proposed approach to state-of-the -art approaches

Table 4 compares the classification results of the proposed approach with state-of-the-art techniques. The best performing representation using the proposed approach is based on concept modeling using Places365-ResNet152 features as cues to pseudo-concepts. The proposed approach considers the images in their original resolutions for building the concept models. Compared to the traditional methods, the proposed representation is compact (bounded by the number of pseudo-concept models) and achieves the best performance. The Fisher vector (FV) framework considered in [34] uses the handcrafted local image descriptors like SIFT and converts them into fixed length for classification using the linear classifier. The work in [3,7,8,11,12,16,42,44,50] proposed scene classification framework using CNN-based architectures. The work in [8] generates the global image representation of scene images using CNN trained on object database. Later, the work in [50] proposed the CNN architecture trained with Places dataset. This architecture gives a significant improvement in classification results compared to CNN trained with an object database. Results with both the architectures are reported on fully connected layer features with an SVM-based linear classifier. The work in [12] extracts CNN-based features from local patches of the image at multiple scales and generates an orderless vector of locally aggregated descriptors at every scale separately, and then concatenate the representation from different scales. The resulting representation is known as multi-scale orderless pooling (MOP-CNN). The work in [7] obtained the semantic FV using standard Gaussian Mixture encoding for

Table 2 Classification accuracy (CA) (in %) using Deep CNN-based SMN representation and the χ^2 kernel-based SVM classifier

Pre-trained CNNs considered for obtaining deep CNN-based SMN representation	MIT8				VS			
	SLC		LC		SLC		LC	
	$\mathcal{C}$	CA	$\mathcal{C}$	CA	$\mathcal{C}$	CA	$\mathcal{C}$	CA
ImageNet-AlexNet	55	89.23	65	92.45	85	80.45	90	81.28
Places205-AlexNet		90.12		93.41		79.81		81.48
Places365-AlexNet		91.23		93.81		80.85		83.01
ImageNet-GoogLeNet	95	90.23	105	93.12	100	81.98	110	82.32
Places205-GoogLeNet		91.21		93.81		83.23		84.98
Places365-GoogLeNet		91.52		93.48		84.12		85.11
ImageNet-VGGNet16	110	91.56	145	93.87	155	82.81	165	84.52
Places205-VGGNet16		92.75		94.07		84.27		85.33
Places365-VGGNet16		92.88		94.09		83.20		86.62
ImageNet-ResNet152	175	94.76	195	94.57	170	86.01	185	86.96
Places365-ResNet152		**95.42**		**95.11**		**87.76**		**88.23**

Base features for building modified DSPMK-based pseudo-concept models and generating SMN representation are extracted from second-to-last CONV (SLC) layer and last CONV (LC) layer of AlexNet, GoogLeNet, VGGNet16, and ResNet152 which are pre-trained deep networks on ImageNet and Places dataset. $\mathcal{C}$ is the number of final pseudo-concept also the length of SMN rep. The highest accuracy of each column is marked in bold

Table 3 Classification accuracy (CA) (in %) using Deep CNN-based SMN representation and the χ^2 kernel-based SVM classifier

Pre-trained CNNs considered for obtaining deep CNN-based SMN representation	MIT67				SUN397			
	SLC		LC		SLC		LC	
	$\mathcal{C}$	CA	$\mathcal{C}$	CA	$\mathcal{C}$	CA	$\mathcal{C}$	CA
ImageNet-AlexNet	250	56.21	350	64.28	320	44.52	230	46.78
Places205-AlexNet		56.89		64.82		49.82		53.42
Places365-AlexNet		62.31		68.91		52.33		53.89
ImageNet-GoogLeNet	310	74.39	380	75.21	705	45.67	710	47.61
Places205-GoogLeNet		73.21		77.89		55.21		59.21
Places365-GoogLeNet		74.88		76.92		56.88		60.23
ImageNet-VGGNet16	355	73.11	410	77.08	490	48.55	500	50.12
Places205-VGGNet16		75.23		80.45		59.25		62.11
Places365-VGGNet16		72.33		77.52		60.54		64.11
ImageNet-ResNet152	850	75.79	890	77.94	1280	56.81	1380	57.91
Places365-ResNet152		**81.76**		**82.83**		**64.78**		**65.89**

Base features for building modified DSPMK-based pseudo-concept models and generating SMN representation are extracted from second-to-last CONV (SLC) layer and last CONV (LC) layer of AlexNet, GoogLeNet, VGGNet16, and ResNet152 which are pre-trained deep networks on ImageNet and Places dataset. $\mathcal{C}$ is the number of final pseudo-concept. The highest accuracy of each column is marked in bold

Table 4 Comparison of classification accuracy (CA) (in %) of the proposed approaches with state-of-the-art approaches on different varying size scene recognition datasets

Methods	MIT8	VS	MIT67	SUN397
SIFT+BOVW [29]	79.13	67.49	35.86	24.82
SIFT+FV [34]	79.56	70.45	58.23	43.30
DeCaF [8]	82.45	72.56	59.50	43.76
Places-CNN-FC7 [50]	88.30	76.02	68.24	54.32
MOP-CNN [12]	89.45	76.81	68.88	51.98
Hybrid-CNN-FC7 [50]	91.23	78.56	70.80	53.86
FC8-FV [7]	88.43	79.56	72.86	54.40
VGGNet + DSP [11]	92.34	81.34	76.34	57.27
MetaObject-CNN [44]	90.45	81.54	78.90	58.11
Bag-of-Elements [27]	91.23	81.34	77.63	–
Patch-based SMN [22]	–	–	79.63	57.47
MLR+CFV [46]	–	–	81.47	64.14
G-MS2F [42]	92.19	84.01	79.25	63.29
SDO [3]	94.23	85.89	81.01	64.23
SPMK-based SVM using true size images [16]	95.09	84.68	77.76	62.31
LK-based SVM using LC layer features of ImageNet-ResNet-152 [original size images]	93.56	82.41	71.33	52.26
LK-based SVM using LC layer features of Places-ResNet-152 [original size images]	93.90	84.54	75.42	58.68
ImageNet-ResNet-152 fine-tuning [fixed-size images]	93.81	83.09	72.67	53.23
Places365-ResNet-152 fine-tuning [fixed-size images]	94.01	84.89	76.67	60.23
SMN from LC+SLC layers of Places-ResNet152 with pseudo-concept modeling using modified DSPMK-based SVM [original size images]	96.84	89.17	83.25	66.11
SMN from LC layer of Places-ResNet152+ImageNet-ResNet152 with pseudo-concept modeling using modified DSPMK-based SVM [original size images]	**97.01**	**89.89**	**84.43**	**66.87**

CNN-based features. The work in [11] used the generative model-based approach and built the dictionary of the activation maps to obtain the FV representation for the different spatial regions. The work in [16] proposed a dynamic kernel-based framework where images are considered in true resolution. In this work, images are represented by varying size sets of activation maps and the classification is performed using dynamic kernel-based SVMs. The work in [44] uses region proposals and discriminative patch mining instead of dense sampling of patches for generating the final image representation by pooling the feature response maps of all the learned metaobjects at multiple spatial scales. The results are further improved by the work of Xie *et al.* [46], they propose two dictionary-based representations, namely mid-level local representation (MLR) and convolutional Fisher vector representation (CFV) using AlexNet and VGGNet fine-tuned networks. The work in [42] proposed a CNN model, which learns the features in the convolutional neural network in multi-stage, they named the network as GoogLeNet-based multi-stage feature fusion (G-MS2F). The work in [3] tries to capture the co-occurrence pattern of all the object across scenes by considering the image patch features and named the approach as semantic descriptor with objectness (SDO).

The proposed classification results are also compared with fine-tuned networks, we fine-tune the ResNet-152 pre-trained architecture using all the four scene recognition datasets. It is observed that the proposed approach performs better than the fine-tuned network.

In our earlier work [14], we proposed an SMN representation with handcrafted features. We observe that results are better in comparison with FV representation for MIT8 and VS dataset. MIT67 is a complex indoor scene dataset, low-level features fail to capture the semantic concepts information hence results in lower classification accuracy. From Table 4, we observe that proposed SMN representation gives better classification results when pseudo-concept models are built using original resolution images in comparison with the fixed reduced size. Combining the classifier output score of proposed SMN representation using LC and SLC layer activation maps with χ^2 kernel-based SVM classifier performs better than other related approaches. It is also observed that scene recognition results are varying with image size.

Though recent few works on scene recognition [35,38,40] show better classification performance which comes at the cost of complex and hybrid CNN architecture with specific training procedures (patch based or multiple scales based), but our method uses only a pre-trained CNN architecture with few parameter tuning for generation of SMN representation which gives us an edge to distinguish it from other methods.

4.5 Ablation study

In this section, we perform an ablation study to analyze the effectiveness of different components in the proposed procedure. The number of pseudo-concepts in pseudo-concept modeling determines the length of SMN representation, i.e., $\mathcal{C}$. In order to show the effectiveness of a step, we keep all the others step fixed. Table 5 shows a summary of the ablation study results on MIT67 and SUN397 datasets for various scenarios. For building the pseudo-concept models, base features are extracted from Places-ResNet152 architecture. We have considered both the pre-trained and fine-tuned architectures. To show the importance of true-resolution images in semantic analysis, we have compared the results with fixed-size images also.

First, we built the pseudo-concept models without selection procedure, i.e., all the filters in the last CONV layer are considered as concept detectors and further grouped using the proposed kernel-based grouping procedure. Next, the experiments are performed without considering the grouping procedure in the pipeline. It is observed that the number of final pseudo-concepts will be more in all cases and results in lower classification accuracy. This experiment shows the importance of selection and grouping procedure in the proposed pipeline.

Second, we compare the results of building the pseudo-concept models and generating the SMN representation using pre-trained network with fixed and true-resolution images. The fixed-size images result in the same size sets of activation maps across the image representations so we build the concept models via linear kernel-based SVM [15]. For true-resolution images, varying size set of activation maps are obtained, so the pseudo-concept models are built via modified DSPMK. It is observed that modified DSPMK results in better concept modeling and accuracy. This also indicates the advantages of considering true-resolution images for building concept models and performing classification.

In addition to the building pseudo-concept models using pre-trained CNN architectures, we perform the similar experiments using fine-tuned networks. To compare, we consider the ResNet152 architecture and fine-tune this network on MIT67 and SUN397 datasets. We consider both fixed and original varying resolution images. It is observed that considering the fine-tuned networks results in a slight improvement of accuracy but it comes at the cost of fine-tuning.

Table 5 Classification results (in %) on MIT67 and SUN397 datasets for varying pipeline configurations

Configuration	MIT67		SUN397	
	$\mathcal{C}$	CA	$\mathcal{C}$	CA
Without PC selection procedure [pre-trained network+original size images]	1450	77.89	1620	60.85
Without PC grouping procedure [pre-trained network + original size images]	1220	78.56	1550	61.25
Without PC selection and grouping procedure [pre-trained network + original size images]	2048	76.25	2048	60.52
Linear kernel-based PC modeling [pre-trained network + fixed-size images]	755	80.34	1240	64.25
Modified DSPMK-based PC modeling [pre-trained network + original size images]	890	82.83	1380	65.89
Linear kernel-based PC modeling [fine-tuned network + fixed-size images]	720	81.45	1180	64.31
Modified DSPMK-based PC modeling [fine-tuned network + original size images]	780	83.52	1250	66.12

Base features are extracted from ResNet152 architecture which is fine-tuned using respective dataset or pre-trained using Places dataset. Images for PC modeling and SMN representation generation are considered in fixed or true resolutions. Here, $\mathcal{C}$ indicates the number of final pseudo-concepts (PC) and length of final SMN representation

4.6 Analysis of effective number of pseudo-concepts

Figure 7 shows classification accuracy versus number of pseudo-concepts for different datasets. The dimensionality of final SMN representation is same as number of pseudo-concept models. It is seen that for small datasets like MIT8 and VS less number of pseudo-concepts are enough. However, for MIT67 and SUN397, a large number of pseudo-concepts are needed. It is also observed that without pseudo-concepts selection and grouping if all the filters response is considered for pseudo-concepts model building then the classification accuracy is 2-3% lower for all the cases. The detailed experiments with different numbers of pseudo-concepts demonstrate that prominent selected filters are enough to generate descriptive and discriminative SMN representation. Considering fewer pseudo-concepts results in low recognition rate whereas too many results in poor generality.

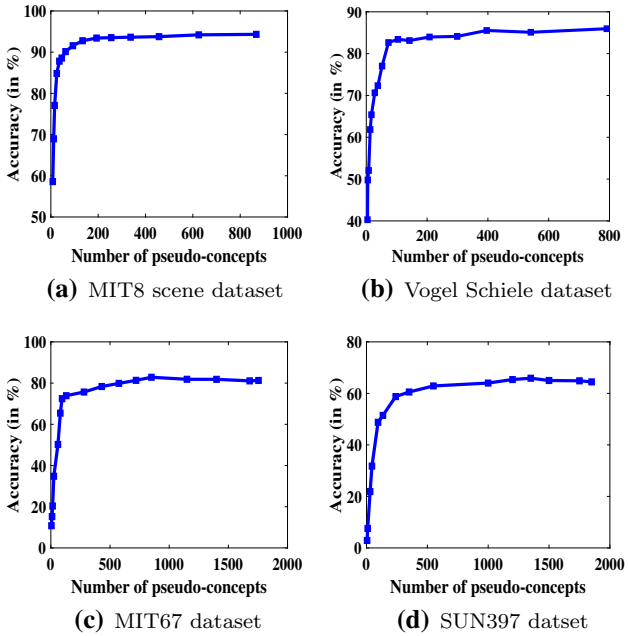

Fig. 7 Classification accuracy (in %) versus the number of pseudo-concepts for different datasets. LC layer activation maps of true-resolution images are computed from Places365-ResNet152 for pseudo-concepts modeling

4.7 Visualization of pseudo-concept versus true-concept

Figure 8 shows the correspondence between maximally activated image regions generated from semantic analysis of fixed reduced size versus original size for three of the MIT67 dataset images using deep-visualization toolbox [48]. True-concept annotations for every image are specified on the right side. Images in the second column are of varying size although here they are shown in a fixed size. Generated concepts annotation for fixed-size images captures a few of the concepts whereas almost all the concepts are captured in original size images. For example, in the images of the first row, concepts such as 'window,' 'cabinets' and 'food' are captured since they are large whereas failed to capture the small concepts such as 'stove,' 'sink' and 'oven.' The main reason is that converting the true-resolution image (i.e., 1200 × 900) to fixed reduced size (i.e., 227 × 227) results in loss of concepts which are of small size in the original image.

Figure 9 shows the number of true-concepts versus generated pseudo-concepts distribution in MIT8 scene dataset. For MIT8 dataset, pixel-wise true-concept annotations are available [28]. We can observe that only 'building,' 'car,' 'mountain,' 'road,' 'sky' and 'tree' concepts are prominently present in the dataset in contrast concepts like 'bird,' 'balcony,' 'cow,' 'crosswalk,' 'moon,' 'sun,' etc. are rarely present in the dataset, hence building concept model for them is very challenging. On the other hand, CNN-based pseudo-concepts are based on their visual appearance and semantic

Fig. 8 Illustration of the presence of pseudo-concepts versus true-concepts in few of the images of MIT67 indoor scene dataset. Pseudo-concepts are marked with red bounding box

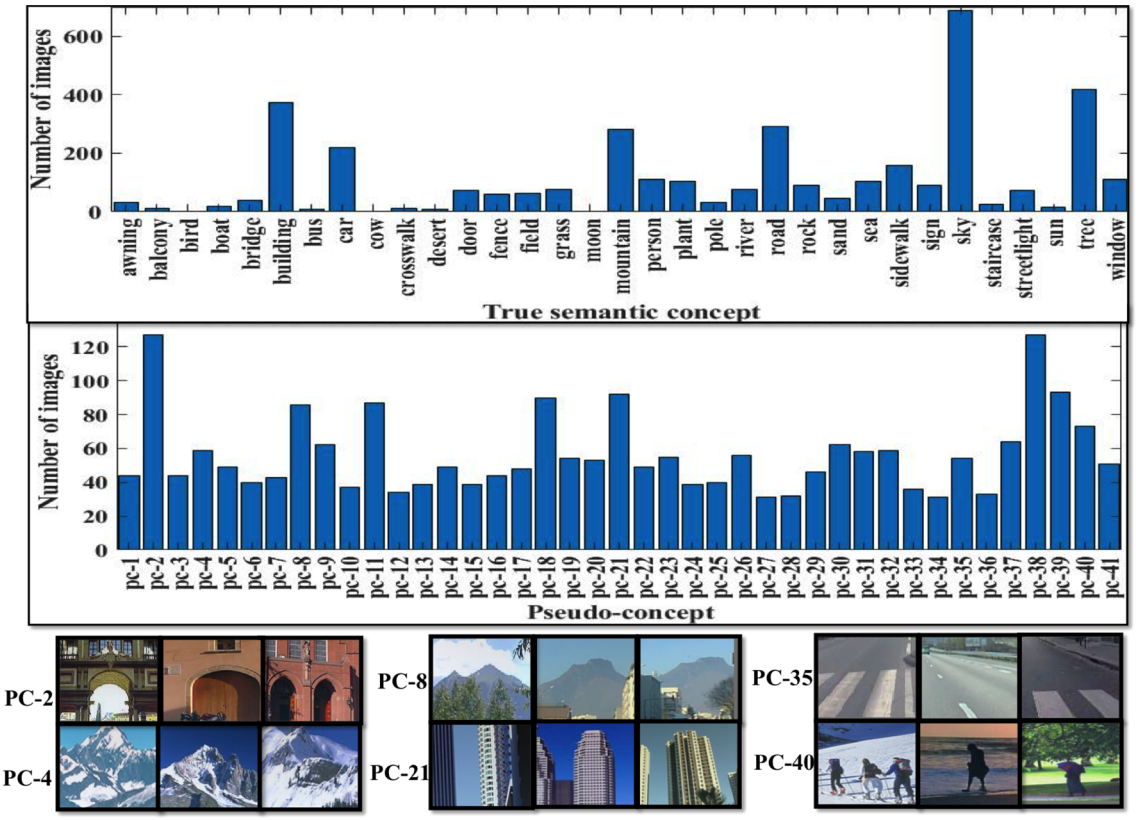

Fig. 9 Illustration of the presence of true-concepts versus pseudo-concepts (PC) in the MIT8 scene dataset (Images can be better visualized in color)

structure, for example, PC-4 and PC-8 both correspond to the 'mountain' concept but as they are very different in appearance so they are captured by separate pseudo-concepts. The concept like, 'crosswalk' is present in very few images but still captured as pseudo-concept. So, we can conclude that pseudo-concepts captured by CNNs are according to the geometric structure of the concepts present in images.

5 Discussion

This paper discusses mainly two issues of varying size scene image recognition, (i) loss of information in the pre-processing stage of CNNs when varying size original resolution images are converted to a fixed reduced size, (ii) challenges in the building of concept model when true-concept annotated dataset is unavailable. For the first issue, varying size original images are considered as input to the CNN and corresponding activation maps from chosen CONV layers are computed without any concept information loss. For the second issue, we propose the idea of pseudo-concepts and consider deeper CONV layer filter responses as cues to pseudo-concepts. Finally, deep SMN representation is generated using built pseudo-concept models. This performs better in contrast to the low-level or high-level feature representation. We also noted that results are further improved by 2-3% on combining the SMN representation obtained using activation maps of the last two CONV layers or from different pre-trained CNN architectures. Furthermore, the proposed representation performed significantly better in terms of accuracy with relatively low dimension and bounded by the number of pseudo-concepts. It also outperforms the other CNN-based approaches without combining any complementary features. The requirement of the proposed approach includes the need for a pre-trained CNN architecture and to choose an effective number of pseudo-concepts in the grouping procedure.

6 Conclusion

We proposed a novel semantic concept-based representation for the recognition of varying size scene images. The loss of concept information in the pre-processing stage is reduced by considering original resolution images as input to the CNNs. In the absence of a true-concept annotated image database, varying size deeper CONV layer activation maps are considered as cues for the concepts and corresponding feature maps for model building. Non-prominent and non-discriminative pseudo-concepts are pruned using the proposed algorithm. Grouping of similar pseudo-concepts and model building is performed in kernel-space using modified DSPMK. The proposed SMN representation captures the information of diverse and varying size concepts present in the images without significant loss of semantic content and hence results in state-of-the-art classification accuracy.

In future, proposed approach can be extended to scene video understanding by building the generalized pseudo-concept models using the different CNN architectures.

References

1. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations, pp. 3319–3327 (2017)
2. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of the British Machine Vision Conference (BMVC 2011), Dundee, Scotland, vol. 2, p. 8 (2011)
3. Cheng, X., Lu, J., Feng, J., Yuan, B., Zhou, J.: Scene recognition with objectness. Pattern Recogn. **74**, 474–487 (2018)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proceedings of Workshop on Statistical Learning in Computer Vision (ECCV 2004), Prague, Czech Republic, vol. 1, pp. 1–2 (2004)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Florida, USA, pp. 248–255 (2009)
6. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 551–556 (2004)
7. Dixit, M., Chen, S., Gao, D., Rasiwasia, N., Vasconcelos, N.: Scene classification with semantic Fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, Massachusetts, pp. 2974–2983, https://doi.org/10.1109/CVPR.2015.7298916 (2015)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the International Conference on Machine Learning (ICML 2014), Beijing, China, pp. 647–655 (2014)
9. Fernando, B., Fromont, E., Tuytelaars, T.: Mining mid-level features for image classification. Int. J. Comput. Vis. **108**(3), 186–203 (2014)
10. Fong, R., Vedaldi, A.: Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks (2018) Preprint arXiv:1801.03454
11. Gao, B.B., Wei, X.S., Wu, J., Lin, W.: Deep spatial pyramid: the devil is once again in the details (2015). Preprint arXiv:1504.05277
12. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Proceedings of European Conference on Computer Vision (ECCV 2014), Zurich, pp. 392–407 (2014)
13. Gupta, S., Dileep, A.D., Thenkanidiyoor, V.: Segment-level pyramid match kernels for the classification of varying length patterns of speech using svms. In: Proceedings of the European Signal Processing Conference (EUSIPCO 2016), Budapest, Hungary, pp. 2030–2034 (2016)
14. Gupta, S., Dileep, A.D., Thenkanidiyoor, V.: The semantic multinomial representation of images obtained using dynamic kernel based pseudo-concept SVMs. In: Proceedings of National Conference on Communication (NCC 2017), Chennai, India, pp. 1–6 (2017)

15. Gupta, S., Dinesh, D.A., Thenkanidiyoor, V.: Deep cnn based pseudo-concept selection and modeling for generation of semantic multinomial representation of scene images. In: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pp. 336–339 (2018a)
16. Gupta, S., Pradhan, D.K., Aroor, Dinesh D., Thenkanidiyoor, V.: Deep spatial pyramid match kernel for scene classification. In: International Conference on Pattern Recognition Applications and Methods ICPRAM, pp. 141–148 (2018b)
17. Gupta, S., Karanath, A., Mahrifa, K., Dileep, A.D., Thenkanidiyoor, V.: Segment-level probabilistic sequence kernel and segment-level pyramid match kernel based extreme learning machine for classification of varying length patterns of speech. Int. J. Speech Technol. **22**(1), 231–249 (2019)
18. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016, Las Vegas, USA, pp. 770–778 (2016)
20. Henderson, J.: Introduction to real-world scene perception. Vis. Cogn. **12**(6), 849–851 (2005)
21. Herranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, USA, pp. 571–579 (2016)
22. Jiang, S., Chen, G., Song, X., Liu, L.: Deep patch representations with shared codebook for scene classification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **15**(1s), 5 (2019)
23. Khan, S.H., Hayat, M., Bennamoun, M., Togneri, R., Sohel, F.A.: A discriminative representation of convolutional features for indoor scene recognition. IEEE Trans. Image Process. **25**(7), 3372–3383 (2016)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS 2012), Nevada, USA, pp. 1097–1105 (2012)
25. Li, L.J., Su, H., Lim, Y., Fei-Fei, L.: Object bank: an object-level image representation for high-level visual recognition. Int. J. Comput. Vis. **107**(1), 20–39 (2014). https://doi.org/10.1007/s11263-013-0660-x
26. Li, P., Samorodnitsk, G., Hopcroft, J.: Sign cauchy projections and chi-square kernel. In: Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS 2013), Harrah's Lake Tahoe, USA, pp. 2571–2579 (2013)
27. Li, Y., Liu, L., Shen, C., Van Den Hengel, A.: Mining mid-level visual patterns with deep cnn activations. Int. J. Comput. Vis. **121**(3), 344–364 (2017)
28. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: Label transfer via dense scene alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Florida, USA, pp. 1972–1979 (2009)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
30. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
31. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Florida, USA, pp. 413–420 (2009)
32. Rasiwasia, N., Vasconcelos, N.: Holistic context models for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 902–917 (2012)
33. Rasiwasia, N., Moreno, P.J., Vasconcelos, N.: Bridging the gap: query by semantic example. IEEE Trans. Multimed. **9**(5), 923–938 (2007)
34. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
35. Seong, H., Hyun, J., Kim, E.: Fosnet: an end-to-end trainable deep neural network for scene recognition. IEEE Access **8**, 82066–82077 (2020)
36. Sharma, K., Gupta, S., Dileep, A.D., Rameshan, R.: Scene image classification using reduced virtual feature representation in sparse framework. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2701–2705 (2018)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition.Preprint arXiv:1409.1556 (2014)
38. Sitaula, C., Xiang, Y., Zhang, Y., Lu, X., Aryal, S.: Indoor image representation by high-level semantic features. IEEE Access **7**, 84967–84979 (2019)
39. Song, X., Jiang, S., Herranz, L.: Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Trans. Image Process. **26**(6), 2721–2735 (2017)
40. Sun, N., Li, W., Liu, J., Han, G., Wu, C.: Fusing object semantics and deep appearance features for scene recognition. IEEE Trans. Circuits Syst. Video Technol. **29**(6), 1715–1728 (2019)
41. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, Massachusetts, pp. 1–9 (2015)
42. Tang, P., Wang, H., Kwong, S.: G-ms2f: Googlenet based multi-stage feature fusion of deep cnn for scene recognition. Neurocomputing **225**, 188–197 (2017)
43. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. In: Proceedings of the International Conference on Image and Video Retrieval (CIVR 2004), Dublin, Ireland, pp. 207–215 (2004)
44. Wu, R., Wang, B., Wang, W., Yu, Y.: Harvesting discriminative meta objects with deep cnn features for scene classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, pp. 1287–1295 (2015)
45. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, CA, pp. 3485–3492 (2010)
46. Xie, G.S., Zhang, X.Y., Yan, S., Liu, C.L.: Hybrid cnn and dictionary-based models for scene recognition and domain adaptation. IEEE Trans. Circuits Syst. Video Technol. **27**(6), 1263–1274 (2015)
47. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Florida, USA, pp. 1794–1801 (2009)
48. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. In: Proceedings of the Deep Learning Workshop in International Conference on Machine Learning (ICML 2015) (2015)
49. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV 2014), Zurich, pp. 818–833 (2014)
50. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS 2014), Montreal, Canada, pp. 487–495 (2014)

51. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)

Shikha Gupta received her B.tech. degree in Computer Science and Engineering from Krishna Engineering College, U.P.T.U, India, in 2013. She received her Ph.D. degree in Computer Science from the School of Computing and Electrical Engineering of the Indian Institute of Technology Mandi, India. She is currently working in Vehant Technology Pvt. Ltd. as Senior Research Engineer. Her research interests include machine learning, pattern recognition, scene understanding, kernel methods, and deep learning.

Dr. A. D. Dileep received his B.E. degree in Computer Science and Engineering Bhalki from Gulbarga University, Karnataka, India, in 2000. He received his M.Tech. degree in Computer Science and Engineering and Ph.D. degree in Computer Science and Engineering from Indian Institute of Technology (IIT) Madras in 2006 and 2013, respectively. He joined as Assistant Professor in the School of Computing and Electrical Engineering at IIT Mandi, Himachal Pradesh, in 2013. Since 2019, he has been working as an Associate Professor in the School of Computing and Electrical Engineering at IIT Mandi. His current research interests are in machine learning, kernel methods, and deep learning with applications in speech technology, computer vision, and cloud network resource utilization.

Veena Thenkanidiyoor received her B.E. degree from Manipal Institute of Technology Manipal, Mangalore University, India, in 1998. She received her M.S. degree from Manipal Academy of Higher Education, Manipal in 2000. She received her Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Madras, Chennai, India, in 2014. She was an Assistant Professor from 2013 to 2018 in the department of Computer Science and Engineering, NIT Goa, India. Since 2018, she has been working as an Associate Professor in the same department. Her research interests include deep learning, kernel methods, computer vision, speech processing, content-based information retrieval, pattern recognition.