

A context-aware semantic modeling framework for efficient image retrieval

K. S. Arun¹ · V. K. Govindan¹

Received: 8 April 2015 / Accepted: 18 January 2016 / Published online: 6 February 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract In recent years, high-level image representation is gaining popularity in image classification and retrieval tasks. This paper proposes an efficient scheme known as semantic context model to derive high-level image descriptors well suited for the retrieval operation. Semantic context model uses an undirected graphical model based formulation which jointly exploits low-level visual features and contextual information for classifying local image blocks into some predefined concept classes. Contextual information involves concept co-occurrences and their spatial correlation statistics. More expressive potential functions are introduced to capture the structural dependencies among various semantic concepts. The proposed framework proceeds in three steps. Initially, optimal values of model parameters that impose spatial consistency of concept labels among local image blocks are learned from the training data. Then, the semantics associated with the constituent blocks of an unseen image are inferred using an improved message-passing algorithm. Finally, a compact but discriminative image signature is derived by integrating the frequency of occurrence of various regional semantics. Experimental results on various benchmark datasets show that semantic context model can effectively resolve local ambiguities and consequently improve concept recognition performance in complex images. Moreover, the retrieval efficiency of the new semantics based image feature is found to be much better than state-of-the-art approaches.

Keywords Image retrieval · Semantic gap · Contextual information · Graphical models

1 Introduction

In past few years, there has been exponential increase in the size of image databases due to the acceptance and wide usage of various digital imaging techniques. Such image collections will be worthless unless there are some intelligent mechanisms to efficiently manage these large volumes of digital image data. Recently, content based image retrieval (CBIR) has been evolved as an efficient mechanism towards this end. In general, traditional CBIR frameworks involves the design of automated procedures for extracting various low-level visual features such as color, texture, shape etc. by directly analysing image pixels to facilitate the search and the retrieval of desired images from a given repository. Unfortunately, the performance of most CBIR systems are inherently constrained by the low-level visual features because they often fail to capture the high-level semantics perceived by humans. This is known as the semantic gap problem in image retrieval.

Nowadays, semantics-based image modeling appears to be an effective solution for minimizing the discrepancy between low-level image features and their high-level meanings. Attempts made in the literature to model images using high-level semantics can be broadly grouped in to two major categories: object based strategies and bag-of-visual-words (BoVW) based models. Object based strategies relies on objects or concepts (e.g. grass, tree, sky etc.) present in images for high-level modeling and retrieval. Initially, a vocabulary of concept classes is defined and then a statistical model is developed to recognize each of these concepts. In the end, low-level local image features

✉ K. S. Arun
iamarunks@gmail.com

V. K. Govindan
vkg@nitc.ac.in

¹ Department of Computer Science and Engineering, National Institute of Technology, Calicut, India

are mapped to high-level semantics by means of the learned model. Thus, the resulting image search engines support semantic queries and the retrieval operation simplifies to keyword matching. On the other hand, the BoVW [1] model follows a more generalized image representation scheme. In this paradigm, a set of local image descriptors are initially quantized to form a vocabulary of visual words in a completely unsupervised fashion. Then, images are characterized by histograms of assignments of local image features to visual words. As a result, images are modeled at a much higher level of abstraction and it considerably minimizes the semantic gap problem in image retrieval.

Even though some promising results have been obtained with these two formulations, the identified semantics is found to be imprecise under certain circumstances. For example, sky and snow regions highlighted in the natural scene image shown in Fig. 1 are ambiguous and misjudged by most of the object based semantic modeling schemes. This is mainly due to the resemblance of visual features extracted from these conflicting image regions. Moreover, an image contain various high-level concepts in different proportions. The weightage of individual concepts may vary according to their prominence in the image. In fact, majority of the object-based strategies fails to capture the composition of individual concept. Thus, the retrieval performance in such situations is often much lower than what really expected.

Similarly, polysemy and synonymy [2] of visual words creates the same problem in BoVW based image retrieval models. To alleviate this issue, topic models such as probabilistic latent semantic analysis (pLSA) [3, 4] and Latent Dirichlet Allocation (LDA) [5] are introduced. Topic models automatically discover semantically coherent visual words also known as latent topics and subsequently represents images by a vector of probability distribution over these identified latent topics. Image characterization using latent topics gives a quantitative description of high-level image semantics and it appears to be suitable for measuring semantic image similarity. However, the latent topics lack explicit meaning and exact inference in these models is also intractable. One has to depend on inaccurate



Fig. 1 An example which illustrates how contextual information helps to discriminate two local image patches that are visually identical but semantically dissimilar

and slow approximation algorithms for inferring the required latent topic distribution. Therefore, it can be concluded from the above discussion that most of the present day semantic modeling approaches for image retrieval are still not ideal and there is an overwhelming demand for advanced and highly efficient model.

On the other hand, humans can accurately predict complex entities present in images without entirely relying on the appearance based features of visual concepts. Psychophysical experiments in humans demonstrated the fact that recognition of a particular concept of interest (e.g. rock) in a given scene is usually facilitated by the presence of certain other co-occurring concepts (e.g. sea, sand) [6]. Thus, the ability of humans to combine visual data with contextual knowledge plays a crucial role in robust scene understanding. Contextual knowledge can be any statistics that are not directly extractable from the pixel values of an image. Such information can be obtained from the nearby image data, image tags and the presence or locations of objects in the image (e.g. concept co-occurrence, spatial relationship statistics).

Therefore, incorporating contextual information in semantic image modeling rule out unlikely combinations of concept classes and consequently improves the concept recognition accuracy to a large extent. Probabilistic graphical models, especially discriminative random field (DRF) [7] turns out to be a powerful framework to encapsulate visual features of local image regions and contextual information in a single constrained model. More recently, Yu et al. [8] presented a unified graphical modeling framework which assigns one of a finite set of labels to individual image segments by integrating various components that encodes contextual information at multiple levels. However, parameter estimation and inference in these models are known to be tough and time consuming.

This paper proposes a computationally efficient probabilistic graphical model referred to as semantic context model (SCM). In the proposed model, along with visual features of local image blocks two types of contextual information namely concept co-occurrence and their spatial location statistics are incorporated in a meaningful and tractable manner for semantic image modeling and retrieval. To identify local image semantics, an approximate but efficient inference procedure is also introduced. In summary, the main contributions of the proposed work are the following.

- An efficient probabilistic graphical model together with a faster inference scheme that takes long range contextual information into account and thereby improve semantics discovery from local image regions to yield a vector of concept count for efficient image retrieval.

- A new class of potential function is introduced, which considerably enhances the expressive power of the proposed model and outperforms the conventional ones in terms of preserving local image details and context modeling ability.

The rest of the paper is organized as follows: Sect. 2 summarizes state-of-the-art image retrieval frameworks using semantics based image modeling. Section 3 reviews probabilistic graphical models and provides an insight into discriminative random field (DRF) for context based image modeling. A detailed description of the proposed semantic context model (SCM) is presented in Sect. 4. The aggregation of local image semantics towards the creation of a compact yet discriminative image descriptor is presented in Sect. 5. The data sets used for evaluating the proposed image retrieval framework is provided in Sect. 6. Experiments carried out to analyse the context modeling ability and the semantics identification performance of SCM is outlined in Sect. 7. Section 8 evaluates the effectiveness of the inferred semantics based image descriptor for the retrieval operation and summarizes the obtained results. Section 9 concludes the proposed work by listing the outcome of the research and the possible enhancements.

2 Related work

In the context of image retrieval, object-based semantic modeling is a long established but challenging problem. The last decade has witnessed some remarkable advancements in this area but the basic procedure remains the same: (1) define a label vocabulary which involves the set of all visual classes of interest, (2) extract a group of appearance based visual features that are expected to be discriminant for each of the concept classes in the label vocabulary, (3) formulate a sophisticated framework that is capable enough to distinguish all these concept classes, (4) design appropriate strategies to assemble the identified high-level information so that semantic similarity between images can be easily computed. These research efforts contributed several superior image descriptors such as SIFT [10], HoG [11], SURF [12], DAISY [13] etc., certain object identification frameworks namely object bank (OB) [14] and Classemes [15], some advanced classification schemes like support vector machines with kernels specially tuned for semantic concept recognition [16–20] and many other refined statistical models [21–25] for semantics discovery.

Most of these early works annotate target images with multiple class labels enabling the users to specify queries in the form of keywords. Finally, a ranked retrieval list is obtained by locating images having class labels similar to

the given query. Therefore, a simple keyword matching operation will do the trick for all these early models. Moreover, inside the space of trained label vocabulary these models exhibit better generalization performance. In spite of these advantages, they have a severe drawback that images are annotated by the presence or absence of different concepts classes without specifying each of its relevance. This issue does not arise for an image based query, provided it has been allowed to be mapped to a semantic feature space [26] whose dimensions are the concept classes of a predefined label vocabulary. This motivated the introduction of semantics based image representation [26] by which images are characterized by a vector of weights for all the keywords in the label vocabulary. Then, the users can implicitly provide the required semantic composition in the form of query images.

In this direction, Vogel and Schiele [9] proposed concept occurrence vector (COV) to model and retrieve natural scene images. The COV encodes occurrence counts of various semantic concepts present in images that are discovered using SVM based region classifiers. However, there were sufficiently great number of misclassification due to the similarities in visual features of some ambiguous concepts and this in turn adversely affects the discriminative power of COV. A similar procedure is adopted by Rasiwasia and Vasconcelos [26] to derive another sort of semantics based image representation called semantic multinomial (SMN). In this model, the posterior probabilities are computed for all the concepts and images are represented by the concept probability distribution. The major limitation of this model is that density estimation of individual concept classes is computationally intractable for large data sets.

It is worth noting that contextual information is discarded in all these approaches while deriving semantics based descriptors for image retrieval. Studies in cognitive science revealed the fact that concept recognition in humans always use contextual information from global as well as local levels [27, 28]. Global context make use of statistics from the entire image (e.g. a street scene will predict the presence of buildings) while local context take into account information from neighbouring areas of the object of interest (e.g. a boat will predict the presence of water). Motivated by this intuition, many sophisticated frameworks based on probabilistic graphical models [29] have been proposed to integrate contextual information at local or global levels for semantic image modeling.

Among various semantic image modeling techniques based on probabilistic graphical models, formulations relying on conditional random field [30] are known to be the most effective mechanism for incorporating meaningful contextual information. In this direction, Kumar and Hebert [30] proposed discriminative random field (DRF)

model which helps to encode the dependencies between the visual features of local image blocks and their labels together with the dependencies of labels among adjacent image blocks in a unified framework. In their experiments to detect man made structures from natural terrestrial images, the DRF framework achieved good recognition results. However, the absence of efficient inference algorithm has limited the applicability of DRF to only small data sets.

Later on, many researchers have proposed models with higher-order features that enhances the basic DRF formulation by incorporating contextual information beyond adjacent image blocks [7, 31]. To some extent, these higher-order features can enforce label consistency among local image blocks. However, the accuracy of these approaches are necessarily restricted by the expressive power of the higher-order features. After that, hierarchical models have been introduced which are capable to integrate contextual information at local and global levels of images. To this end, He et al. [32] introduced a multi-scale conditional random field (mCRF) that combine classifiers with local and global features for identifying regional semantics. Contextual features at multiple scales helps to discriminate ambiguous concepts to a greater extent. But, the training and inference time increases exponentially with the number of scales.

It is worth noting that Krhenbhl and Koltun [33] recently proposed a dense CRF model to captures long-range interactions among image super-pixels and shows promising results in semantic image segmentation. Dense CRF models incorporate dependencies among all pairs of image super pixels and it has been proven to be instrumental in boosting the accuracy of semantic segmentation. These high-order connectivities substantially improved the labeling accuracy. However, the long-range interactions makes dense CRF computationally much expensive and significant effort has to be devoted to accelerate the inference procedure. To conclude, all these graphical modeling based frameworks focuses more on improving the performance of semantic concept discovery rather than its application to image retrieval. In fact, it is not always clear how the learned semantic information can be properly combined so as to perform the retrieval operation more effectively.

Realizing the above mentioned shortcomings of existing probabilistic graphical models, this paper proposes a new framework called semantic context model (SCM) to derive high-level image descriptors for the retrieval operation. To do so, SCM is defined over non-overlapping, fixed size image blocks to model concept co-occurrences and their spatial relationship statistics in a unified fashion to further enhance block-wise concept recognition. This in turn yield a new image representation comprising the weightage of

individual concepts. A fully connected undirected graphical model based formulation is used by SCM to learn contextual knowledge from training data in a supervised manner to maximize the class label agreement among local image blocks according to spatial and co-occurrence constraints. To the best of our knowledge, SCM is the first attempt that combine contextual information at local as well as global image level in a computationally tractable form to infer comprehensive semantic descriptors applicable for image retrieval.

3 Preliminaries

Before proceeding further, this section provides a brief introduction about probabilistic graphical models. Furthermore, the theoretical foundations of discriminative random field (DRF) based image modeling framework is also included and it serves as the basis of our proposed SCM formulation.

3.1 Probabilistic graphical models

This section provides the formal definition of probabilistic graphical models. It involves a graph $G = (V, E)$ where $V = \{1, 2, \dots, n\}$ is the vertex set such that each of them corresponds to a random variable X_i taking values from some predefined space Ω_i and $E \subseteq V \times V$ is the collection of edges. Then, the multivariate random variable $X = (X_1, X_2, \dots, X_n)$ is supposed to have a joint probability distribution p on $\Omega = \{(x_1, x_2, \dots, x_n) : x_i \in \Omega_i\}$, where Ω represent the space of n-ary Cartesian product $\Omega_1 \times \Omega_2 \times \dots \times \Omega_n$. The fundamental idea of graphical models is that the joint probability distribution p can be factorized according to the structure of G . The presence (or absence) of edges in G indicate the conditional dependence (or independence) of the corresponding pair of random variables given the rest. The edges of the graph may be either directed or undirected leading to two different categories of probabilistic graphical models—directed graphical models and undirected graphical models.

3.1.1 Directed graphical models

Let us first introduce the concept of directed graphical models. Given a directed acyclic graph $G = (V, E)$, a node $s \in V$ is said to be the parent of another node $t \in V$ if there is a directed edge from s to t . Let $\pi(s)$ denote the set of parents of the node s . For $A \subseteq V$, define $X_A = \{X_a : a \in A\}$ be the collection of random variables indexed by the nodes in A . Then, for each vertex s and its parent set $\pi(s)$ G defines a nonnegative potential function $f_s(X_s | X_{\pi(s)})$ such that $\sum_{X_s} f_s(X_s | X_{\pi(s)}) = 1$. Thus, directed graphical

models factorizes a multivariate probability distribution as the product of various potential functions. This can be formally expressed as:

$$p(X) = \prod_{s \in V} f_s(X_s | X_{\pi(s)}) \tag{1}$$

Directed graphical models are also known as Bayesian networks (or Bayes nets) and are successfully used in a variety of application areas where the observed data exhibits causal or temporal relationships.

3.1.2 Undirected graphical models

In the case of undirected graphical models, a multivariate joint probability distribution is decomposed in accordance with an undirected graphical structure $G = (V, E)$. Since the edges in undirected graphs do not have a direction, it is not possible to express the joint distribution as the product of conditional probabilities of individual nodes. Therefore, undirected models rely on cliques (i.e., fully connected subgraphs of G) to define potential functions. Let C be the set of all cliques of the underlying graph G , then the joint probability distribution is factorized as the product of non-negative clique potentials $\psi_c(X_c)$ defined for each clique $c \in C$.

$$p(X) = \frac{1}{Z} \prod_{c \in C} \psi_c(X_c) \tag{2}$$

where Z is a normalizing constant to ensure that $p(X)$ is a true probability distribution and is given by:

$$Z = \sum_X \prod_{c \in C} \psi_c(X_c) \tag{3}$$

Undirected graphical models are also known as Markov random fields (MRFs). Since there is no topological ordering associated with undirected graphical models and are symmetric in nature, they are more suited for problems involving spatial or relational data. One such application is semantic image modeling in which contextual constraints are effectively modeled using undirected graphical models to infer local image semantics. The following subsection briefly summarizes the statistical framework for one such model.

3.2 An overview of DRF-based semantic image modeling

Semantic image modeling frameworks assign one of a several class labels to each local site (i.e., a pixel, a super-pixel or a square block from regular grid) of a given image. It is basically a supervised learning problem and is formally defined as follows: Let $\{X_1, X_2, \dots, X_N\}$ be an observed data from an input image X where $X_i =$

$[X_{i1}, X_{i2}, \dots, X_{id}]$ denotes the visual features associated with a local image site $i \in S$. $S = \{1, 2, \dots, N\}$ is the set of all local sites in an image and d is the dimension of feature vectors extracted from individual image sites. Let $L = \{L_1, L_2, \dots, L_K\}$ be the predefined label vocabulary where K is the total number of concept classes, then the goal of semantic image modeling is to assign a class label $Y_i \in \{L_1, L_2, \dots, L_K\}$ to each image site $i \in S$ in such a way that is Bayes optimal. Finally, the corresponding labels of the whole image are given by $Y = \{Y_1, Y_2, \dots, Y_N\}$.

In conventional MRF, the estimation of the posterior distribution $p(Y | X)$ over the class labels given the observation sequence involves implicit modeling of the likelihood $p(X)$. However, Kumar and Hebert [7] proposed a discriminative random field (DRF) that directly modeled the posterior distribution without specifying a model for the likelihood. Their goal was to classify each non-overlapping local image block in a natural scene as either structured (man-made) or non-structured (natural background). It is modeled as a binary classification problem with $L = \{0, 1\}$ and they introduced an energy function based formulation to directly infer the posterior distribution $p(Y | X)$ of local image labels.

The core of DRF formulation is an undirected graph $G = (V, E)$ where $V = \{Y_1, Y_2, \dots, Y_N\}$ is the vertex set and $E \subseteq V \times V$ the edge set. As an example, Fig. 2 depicts the graphical representation of a DRF model defined over an image decomposed in to local image blocks by a regular grid of size 3×3 . The dependencies between class labels of neighboring image blocks are represented by the edges of the graph while the dashed lines represents interactions among observations at a given image block and the labels of its neighbouring image blocks. Any possible assignment of class labels to these local image blocks as a whole is referred to as a labeling. With this basic formulation, DRF

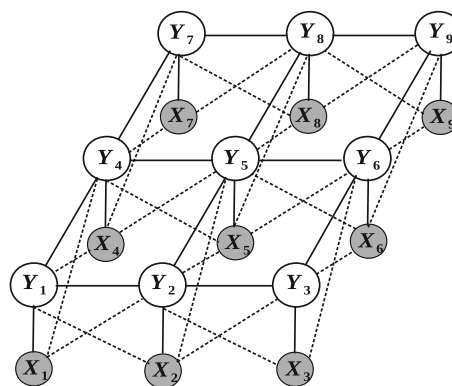


Fig. 2 An illustration of a typical DRF framework defined over 2-D image lattices obtained with a 3×3 grid. The model assumes 4-neighborhood structure. The unshaded circles correspond to label variables and the shaded circles represent the observed data

express the conditional distribution of the labeling Y given the observations X as:

$$p(Y | X) = \frac{1}{Z(X)} \exp(-E(Y | X)) \quad (4)$$

where $E(Y | X) = \sum_{c \in C} \psi_c(X_c, Y_c)$ is the energy of the label configuration Y , $\psi_c(X_c, Y_c)$ denotes non-negative clique potential defined over a collection random variables indexed by the nodes in the clique $c \in C$ and $Z(X)$ is a normalization factor expressed as $Z(X) = \sum_Y \prod_{c \in C} \psi_c(X_c, Y_c)$.

A clique $c \in C$ consist of a set of local image blocks that are interconnected together with a predefined neighborhood system. Thus, cliques can be of type unary (includes only a single local image block), binary (includes a pair of neighboring local image blocks), ternary (includes a triple of neighboring local image blocks) and so on. In the original DRF formulation, potentials up to binary cliques are assumed to be non-zero and all other higher-order potentials are set to zeros. Moreover, the pairs of all neighboring local image blocks are identified on the basis of 4-adjacency relation. Kumar and Hebert [7] experimentally demonstrated that this setting is sufficient enough to obtain satisfactory results. Hence, the energy function in DRF model is defined as:

$$E(Y | X) = \exp \left(\sum_{i \in V} \psi_i(Y_i) + \sum_{i \in V} \sum_{j \in N_4(i)} \psi_{ij}(Y_i, Y_j) \right) \quad (5)$$

where i and j are the indices of local image blocks, $N_4(i)$ is the 4-neighbours of an image block i . With this formulation, it is well understood that an optimal label configuration will always possess a low energy value.

The initial term in the exponent of (5) is the first order clique potential and is generally known as association potential. It links the class label Y_i to an image block i based on the features X_i extracted from it. Kumar and Hebert [7] employed logistic regression model [34], a local discriminative classifier for modelling the association potential by linking it to the conditional probability $p(Y_i | X_i)$ of class Y_i at image block i given the data X_i :

$$\psi_i(Y_i) = \log p(Y_i | X_i) = \log \left(\frac{1}{1 + e^{-Y_i \mathbf{w}^T X_i}} \right) \quad (6)$$

where \mathbf{w} represents the model parameters of the unary potential that have to be determined in the training phase.

Similarly, the next term in the exponent of (5) represent the second order clique potential and is also referred to as interaction potential. The interaction potential is a measure of the influence of the observation X_j and class label Y_j of the neighboring image block i on the class label Y_i of the image block i . To model the interaction potential among a pair of local image blocks (i, j) , DRF formulation extended the classical Ising model [35] and is defined by:

$$\psi_{ij}(Y_i, Y_j) = \begin{cases} \mathbf{v} \Upsilon(X_i, X_j), & \text{if } Y_i == Y_j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where \mathbf{v} denotes the parameters of association potential and $\Upsilon(\cdot)$ characterizes the interaction between local image blocks i and j .

The principle of maximum likelihood provides an efficient means for estimating the model parameters of DRF. Once the parameters are learned from an appropriate set of training samples, an optimal label configuration Y^* over local image blocks of a new test image is computed as:

$$Y^* = \arg \min_{Y \in L^N} E(Y | X_{test}) \quad (8)$$

where X_{test} is the block-wise features of the test image and L^N is the set of all possible label configuration.

Although discriminative random field integrate visual and contextual information in a unified framework for semantic concept discovery, it still need improvements in the following aspects. Primarily, the expressive power of the model has to be enhanced. To capture contextual information, only interactions among image blocks in the 4-neighborhood are incorporated in this model and the long-range dependencies are completely ignored. Secondly, DRF only deals with semantics recognition and does not give any clue towards the formation of semantics based image descriptors for the retrieval task. This paper address the above mentioned limitations and proposes a semantic context model for image retrieval operation.

4 The proposed semantic image modeling framework

The proposed framework for extracting high-level description of images through semantic modeling is illustrated in Fig. 3. The entire framework mainly involves two processing steps. The first step is the training of semantic image modeling framework and it proceeds in four stages: (1) image decomposition (2) feature extraction (3) formulation of semantic context model (SCM) (4) learning the parameters of SCM. Next step is the testing of the learnt SCM formulation. That is, region-wise semantics of an unseen image is inferred using the learnt model and then this local semantics are combined to form a global image representation. A detailed description of these major steps are presented in the remainder of this section.

4.1 Image decomposition

Most region based approaches in the literature rely on image segmentation algorithms for semantics recognition. Feature extraction for visual representation and the generation of training sets are based on such local image

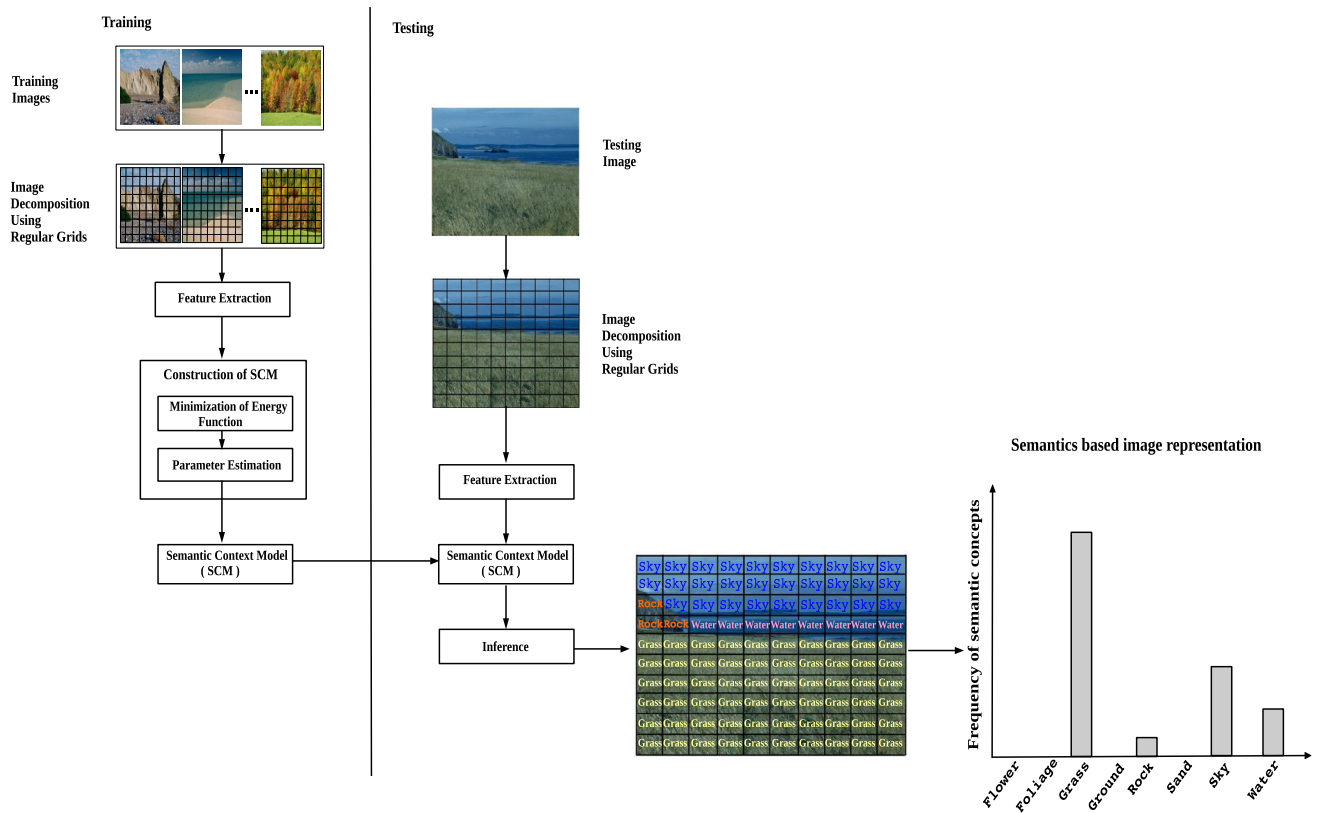


Fig. 3 The outline of the proposed semantic image modeling framework

segments. However, there are intrinsic difficulties in using image segmentation for semantics identification. Primarily, precise extraction of objects from images using automated segmentation is an open problem in computer vision. Secondly, segmentation algorithms often add heavy computation load to the entire image retrieval system. Moreover, some human assisted segmentation algorithms may impose extra burden on users and are infeasible when large database has to be processed.

Considering these facts, an alternative scheme for image decomposition based on regular grids has been suggested by Dagli and Huang [36]. In general, the goal of grid based image decomposition is to identify suitable visual patterns for various semantic concepts by narrowing the influence of noise in the segmentation process. This will also maintain low computational load as compared to the existing image segmentation algorithms. Moreover, Vogel and Schiele [8] demonstrated that semantic modeling of natural scenes using uniform regular grids accomplished good retrieval results. Therefore, the proposed framework use non-overlapping, fixed size local image blocks as the basic units for semantics discovery.

Since block-based decomposition scheme does not always provide precise segmentation of non-scene object categories having clear shape and boundary, this work

restricts the label vocabulary to have only natural scene concepts that are relatively easy to segment out. As an example, Fig. 4 depicts the local image blocks corresponding to various semantic concepts in an image and it includes *sky*, *water*, *rock* and *grass*. The experimental results demonstrate the fact that this simplistic approach for concept extraction does provide a reasonable performance when applied along with the proposed SCM-based formulation for image retrieval.

4.1.1 Creating ground truth data for SCM

Obtaining ground truth data for training and testing the proposed SCM framework is a challenging task. First of all to avoid faulty segmentation, input images are decomposed into uniform grids of $p \times p$ local blocks each comprising 1 % of the total image area. Then, expert annotations are obtained for each of these local image blocks based on a predefined label vocabulary. To do so, relevance judgements given by six domain experts from our research group are taken into consideration. Afterwards, a ground truth for each data collection is generated by accumulating all these human judgements. The same procedure has been adapted in MIRFLICKR [37] dataset to form the ground truth annotation.

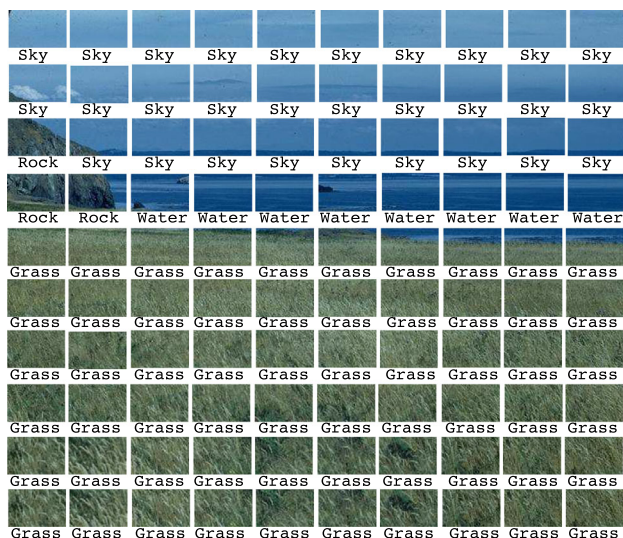


Fig. 4 An example of local image blocks in an image corresponding to various semantic concepts

4.2 Feature extraction

The SCM framework proposed in this paper directly operates on local image blocks to infer semantics based representation for image retrieval. In the proposed framework, the extracted visual features together with the contextual information from its neighboring blocks determines the semantics of a given local image block. Usually the appearance of local image segments are characterized by low-level visual features such as color, texture, shape etc. In practice, these features are not always capable to discriminate complicated and varying semantic concepts. On certain occasions, large intra-class and small inter-class variability of semantic concepts cause ambiguities in classification. Therefore, the visual features should be carefully designed such that they are discriminative and at the same time insensitive to such variabilities.

More recently, such representation are obtained with scattering transform (ST) [38, 39] a deep convolution network based architecture and is found to be effective for applications such as speech recognition [40], image classification [41] etc. Hence, SCM framework relies on scattering transform to have local image features that possess invariance property as well as discriminating power. This will further strengthen the association and the interaction potentials of SCM. The computational aspects of the scattering operator based local image descriptors is presented in the following subsection. Apart from this, SCM use two dimensional positional information $\mathbf{p}_i = (px_i, py_i)$ of each local image block i as a geometric feature to define the interaction potential. px_i and py_i denotes the position offset of the centroid of image block i in the x-y plane.

4.2.1 Scattering representation of image blocks

The scattering transform coefficients for local image blocks are obtained with a deep convolution network, which split up a given image block along multiple paths by the successive application of wavelet convolutions and modulus operators as shown in Fig. 5. The scattering transform perform multiple layers of wavelet transforms, along with complex modulus operations and low-pass filter averaging to yield a representation that guarantees invariance and high discriminative power. A more detailed mathematical description of the scattering operator can be found in Bruna and Mallat [38]. The coefficients in the bottom layer capture minute details present in the input image block. Thus, scattering transform combines a set of coefficients with different degrees of compromise between discrimination and invariance. As coefficients of different compromise degrees are integrated, an appropriate classification boundary is more likely to be found when the concatenated coefficients are employed as a feature descriptor. From the output of the scattering network a local image descriptor $S_{I_j(Z)}$ can be finally composed by:

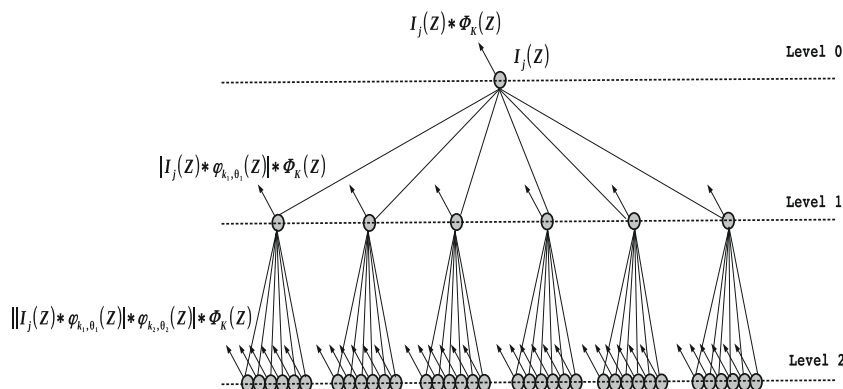
$$S_{I_j(Z)} = \begin{pmatrix} I_j(Z) * \phi_K(Z) \\ |I_j(Z) * \varphi_{k_1, \theta_1}| * \phi_K(Z) \\ ||I_j(Z) * \varphi_{k_1, \theta_1}| * \varphi_{k_2, \theta_2}| * \phi_K(Z) \\ \dots \end{pmatrix} \quad (9)$$

where $I_j(Z)$ is the j th local image block with $Z \in \mathbb{R}^2$ being the pixel positions, φ is the directional wavelet and ϕ is a Gaussian low pass filter. We use complex Gabor function [42] for φ with $\{k_1, k_2, \dots, k_s\}$ and $\{\theta_1, \theta_2, \dots, \theta_r\}$ denote the scale and orientation of the Gabor wavelet φ in each layer of the scattering network. In this work, the convolution network uses three scattering levels (scales) and six orientations for the complex Gabor function φ .

4.3 Formulation of the semantic context model

This section provides a detailed description of the proposed semantic context model (SCM). Given an image decomposed into non-overlapping, fixed sized blocks, SCM assign one of a several class labels from a predefined concept vocabulary to individual image blocks. To ensure accurate labeling of local image blocks, SCM incorporate as much contextual information as possible in a unified framework. SCM basically augments the traditional DRF formulation with more meaningful contextual information. That is, images are modeled using fully-connected undirected graphs that establishes interaction potentials among all pairs of local image blocks. This greatly improves the expressive power and the concept detection accuracy of the proposed model.

Fig. 5 Architecture of the scattering convolution network



SCM poses the task of semantics identification from local image blocks as an energy minimization problem in a fully-connected undirected graph. In the SCM framework, images partitioned into non-overlapping, fixed sized blocks are mapped to fully connected undirected graphs where each image block corresponds to a node in the graph and each node is connected with an edge to every other node. That is, each node is assumed to be a neighbor of every other node. With this assumption, SCM is able to take into account not only neighboring information but also long-range interactions between local image blocks.

More formally, let $B = \{1, 2, \dots, N\}$ be the set of all non-overlapping fixed sized blocks in an image \mathbf{I} . Let y_i be the random variable associated with a local image block $i \in B$, which represents the label assigned to the i th image block. Each variable y_i can take any value from a predefined set of labels $l = \{l_1, l_2, \dots, l_K\}$, where K denotes the total number of semantic classes. Let \mathbf{y} be the vector formed by the random variables $\{y_1, y_2, \dots, y_N\}$ known as the semantic labeling of image \mathbf{I} . Thus, a labelling \mathbf{y} refers to any possible assignment of labels to the local image blocks and it takes values from the set $\mathbb{L} = l^N$. Let the observed data from the set of local blocks B of the given image \mathbf{I} be $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$. Then, SCM is a fully connected undirected graph $G = (V, E)$ where $V = \{y_1, y_2, \dots, y_N\}$ and it characterizes the energy associated with each label configuration \mathbf{y} as:

$$E(\mathbf{y} | \mathbf{x}) = \sum_{c \in C} \Psi_c(\mathbf{x}_c, \mathbf{y}_c) \tag{10}$$

where $\Psi_c(\mathbf{x}_c, \mathbf{y}_c)$ represents non-negative clique potential defined over the random variables indexed by the nodes in the clique $c \in C$. Similar to DRF only the potentials up to binary cliques are considered for defining the energy function and all other higher order clique potentials are ignored. However, in contrast to the 4-adjacency relation to find neighboring pairs of interconnected image blocks in DRF here all pairs of image blocks are interconnected by

an edge. That is, interaction potential is defined among all pairs of local image blocks. Then, the energy function for SCM can be formulated as:

$$E(\mathbf{y} | \mathbf{x}) = \sum_{i=1}^N \Psi_i(y_i, x_i) + \sum_{i=1}^N \sum_{j>i}^N \Psi_{ij}(y_i, y_j) \tag{11}$$

where N is the total number of local image blocks, $\Psi_i(y_i, x_i)$ is the node potential and is computed independently for each local image block i . Node potential gives the likelihood of an image block i to be assigned to one of the labels in the vocabulary l . We use the probabilistic output of an SVM classifier to model the node potential. $\Psi_{ij}(y_i, y_j)$ denotes the interaction potential between node i and j . It is formulated as:

$$\Psi_{ij}(y_i, y_j) = \eta_{ij}(y_i, y_j) \chi_{ij}(y_i, y_j) \tag{12}$$

where $\eta_{ij}(y_i, y_j)$ and $\chi_{ij}(y_i, y_j)$ respectively denotes the feature similarity term and the spatial correlation terms. A detailed description of these newly introduced potential functions are given in the following subsection.

4.3.1 Node potential

Node potential gives the score of assigning a particular class label y_i to a given local image block i . This score is computed from the probabilistic output of an SVM classifier separately trained for each of the concept classes in the predefined label vocabulary. From the definition of the energy function specified in (11), it follows that the most likely labeling \mathbf{y} for a given image must possess a low energy value. In other words, energy should be lower for labels that have a higher class probability. To reflect this basic idea, the negative logarithm of the probabilistic output of the SVM classifier is used as the unary potential. Thus, the unary potential Ψ_i of a local image block i is formally defined as:

$$\Psi_i(y_i, x_i) = -\log(p(y_i = \mathbf{c} | x_i)) \tag{13}$$

where $x_i = S_{I_i(Z)}$ is the scattering wavelet coefficient extracted from the local image block i and $p(y_i = \mathbf{c} \mid x_i)$ is the probability by which i th local image block belongs to class \mathbf{c} .

Since standard SVM formulation does not provide probabilistic outputs, we utilize Platt scaling [43, 44] to convert the SVM outputs into probability values. The Platt scaling operation transforms the discriminant function value $\mathbf{f}_{\mathbf{c}}(x_i)$ of SVM with in the range $[-\infty, +\infty]$ to a posterior probability p using a sigmoid function as:

$$p(y_i = \mathbf{c} \mid x_i) = \frac{1}{1 + \exp(\mathbf{A}_{\mathbf{c}} \cdot \mathbf{f}_{\mathbf{c}}(x_i) + \mathbf{B}_{\mathbf{c}})} \quad (14)$$

where $\mathbf{A}_{\mathbf{c}}$, $\mathbf{B}_{\mathbf{c}}$ denotes the parameters of the sigmoid function for the class \mathbf{c} and are to be estimated from the training samples. Given the feature vector $x_i \in \mathbb{R}^d$ extracted from a local image block i , the SVM discriminant function $f_{\mathbf{c}}(x_i)$ for the concept class \mathbf{c} takes the following form:

$$f_{\mathbf{c}}(x_i) = \langle \rho(x_i), \boldsymbol{\omega}^{\mathbf{c}} \rangle \quad (15)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product, $\rho(x_i) : \mathbb{R}^d \rightarrow \mathbb{R}^D$ is a mapping function that transform the original input data vector x_i to a higher dimensional feature space and $\boldsymbol{\omega}^{\mathbf{c}} = \{\omega_1^{\mathbf{c}}, \omega_2^{\mathbf{c}}, \dots, \omega_D^{\mathbf{c}}\}$ represent model parameters of SVM for class \mathbf{c} . The transformation function is employed to handle nonlinearly separable input vectors in its original feature space. The SVM optimization problem involves inner product in the new feature space \mathbb{R}^D and by using kernel functions, explicit transformation of input features can be avoided. The kernel function usually takes the form: $\kappa(x_i, x_j) = \langle \rho(x_i), \rho(x_j) \rangle$. In this paper, we use RBF kernel as the mapping function and its details are provided in Sect. 4.4.1.

4.3.2 Interaction potential

In this work, we proposes a new kind of interaction potential that can efficiently capture the spatial dependencies among all pairs of local image blocks. This helps to integrate more relevant contextual statistics, which in turn enhance the expressive power of the proposed model. The pairwise interaction between any two local image blocks is modelled as the combination of a feature similarity term and a spatial correlation term. A detailed description of these two components are given below.

Feature similarity term: The term $\eta_{ij}(y_i, y_j)$ compute the difference in visual features between two local image blocks and a higher cost will be assigned if image blocks with similar features attain different class labels. It is formally defined as:

$$\eta_{ij}(y_i, y_j) = \mu(y_i, y_j) \exp(-\mathbb{C}_{ij} / (1 + \|x_i - x_j\|^2)) \quad (16)$$

where $\mu(y_i, y_j)$ denotes a label compatibility function which takes the following form:

$$\mu(y_i, y_j) = \begin{cases} 1, & \text{if } y_i \neq y_j \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

The term $\frac{1}{\|x_i - x_j\|^2}$ specified in (16) computes the difference in visual features of local image blocks i and j . On the other hand, the contextual price matrix \mathbb{C}_{ij} in the numerator of (16) captures the conflict rate for two neighboring image blocks based on their class labels. The normalized confusion matrix obtained after applying the SVM classifiers on the training data set is used as the contextual price matrix. The confusion matrix indicates the number of misclassification for each pair of semantic labels. A large value for an off-diagonal component in \mathbb{C}_{ij} indicate that the labels y_i and y_j have a significant degree of conflict with each other. In order to reduce this ambiguity, a large cost proportional to the conflict rate is assigned for the interaction of these labels. Experimental results proved that incorporating the relationship between all the classes by the interaction potential significantly improves the concept recognition accuracy. Moreover, explicit knowledge about the datasets in the form semantic ontology is not needed to design contextual price matrix. Thus, the feature similarity term imposes high penalty to those local image blocks that are assigned different labels but are similar with respect to their visual features.

Spatial correlation term: The spatial correlation term is designed in such a way that neighboring image blocks have greater impact on the class label of a given image block than distant ones. That is, image blocks closer to the region under consideration will have to assign higher weights and more distant neighbors will be weighted less as shown below:

$$\chi_{ij}(y_i, y_j) = \exp(-\|\mathbf{p}_i - \mathbf{p}_j\|^2 / 2\theta) \quad (18)$$

where the parameter θ controls the degree of nearness. \mathbf{p}_i and \mathbf{p}_j are the coordinate vectors of image sites i and j .

Experimental results on various datasets demonstrated the fact that the proposed feature similarity term and the spatial correlation term in the interaction potential is superior to the existing models in terms of incorporating meaningful contextual information and classification accuracy.

4.4 Learning the model parameters

When the cardinality of the label variables is large, discriminative training of graphical models becomes highly expensive. In such cases approximate training algorithms

that scale well with label cardinality are preferable. One such technique that has been employed throughout the literature is the piecewise training [45] that divides the model into a set of pieces and then train each piece separately with ground truth instead of a joint learning of parameters of all the potential functions at the same time. Sutton and McCallum [45] very well analysed this piecewise estimation method and concluded that it can have good accuracy and give good performance when the local features are highly informative. The intuition behind piecewise learning is that for a given set of training samples, if each potential function can accurately predict their parameters by their own, then the prediction of the combined model will also be accurate. Therefore, this work employs piecewise training approach for learning the parameters of the SCM model.

4.4.1 Learning the parameters of node potential

A Tikhonov regularization [46] based optimization framework for soft margin SVM is solved to obtain the parameters of the unary potential for a given concept class c . It is basically a minimization problem of the following form:

$$F = \min_{\omega^c} \left(\sum_{i=1}^n \sum_{j=1}^n \omega_i^c \omega_j^c \kappa(x_i, x_j) + \xi \sum_{i=1}^n \mathbf{L}(y_i, f_c(x_i)) \right) \tag{19}$$

where $\{x_i, y_i\}_{1 \leq i \leq n}$ is the given training set such that $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, ξ is a regularization parameter, $\kappa(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ represents the RBF kernel with bandwidth parameter γ and \mathbf{L} is a differentiable quadratic loss function of the following form:

$$\mathbf{L}(y_i, f_c(x_i)) = \max(0, 1 - (\delta(y_i, \mathbf{c})) (y_i f_c(x_i)))^2 \tag{20}$$

where $\delta(y_i, \mathbf{c}) \in \{-1, 1\}$ and it indicate whether the label variable y_i is assigned the class label \mathbf{c} or not. This can be formally expressed as:

$$\delta(y_i, \mathbf{c}) = \begin{cases} 1, & \text{if } y_i == \mathbf{c} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

By defining the kernel matrix \mathbf{K} with $\mathbf{K}_{ij} = \kappa(x_i, x_j)$ and \mathbf{K}_i as the i -th column of \mathbf{K} , (19) can be rewritten as:

$$F = \min_{\omega^c} \omega^c \mathbf{K} \omega^c + \xi \sum_{y_i \mathbf{K}_i^T \omega^c < 1} (1 - y_i \mathbf{K}_i^T \omega^c)^2 \tag{22}$$

the change of limits in the summation follows from the fact that the loss function is non-zero only if $1 - (y_i \mathbf{K}_i^T \omega^c) > 1$ and this happens only when $y_i \mathbf{K}_i^T \omega^c < 1$.

The above formulation can be easily solved using Newton’s method [47]. To find the unknown parameter ω^c ,

each step of the Newton’s method does the following updates

$$\omega^c = \omega^c - \alpha \mathbf{H}^{-1} \nabla \tag{23}$$

where α is a suitable step size and in our simulations we observed that a constant value of $\alpha (= 1)$ did not result in any convergence problem. Therefore, in each iteration of the Newton’s method we fixed the value of α to 1. \mathbf{H} is the Hessian of the objective function F and ∇ denotes the partial derivative of the objective function F with respect to the unknown parameter ω^c .

The gradient of the objective function F at ω is calculated as:

$$\begin{aligned} \frac{\partial F}{\partial \omega^c} &= 2\mathbf{K}\omega^c - 2\xi \sum_{y_i \mathbf{K}_i^T \omega^c < 1} (1 - y_i \mathbf{K}_i^T \omega) y_i \mathbf{K}_i \\ &= 2(\mathbf{K}\omega^c + \xi \mathbf{K} I^0 (\mathbf{K}\omega^c - \mathbf{Y})) \end{aligned} \tag{24}$$

where I^0 is an $n \times n$ diagonal matrix with diagonal entries equal to 1 for those data points x_i for which $y_i \mathbf{K}_i^T \omega^c < 1$ and the others set to zero.

Similarly, the Hessian \mathbf{H} is defined as the matrix of second order partial derivatives of F with respect to ω^c and is easily calculated as:

$$\mathbf{H} = \frac{\partial \nabla}{\partial \omega^c} = 2(\mathbf{K} + \mathbf{K} I^0 \mathbf{K}) \tag{25}$$

In each iteration, ω^c is updated using (23) and the entire steps are repeated until convergence is achieved. That is, the parameter values in successive iterations became same. The whole procedure is summarized in Algorithm 1.

Algorithm 1 Parameter estimation of unary potential

Input: A set of training samples $\{x_i, y_i\}_{1 \leq i \leq n}$ for class \mathbf{c} , Kernel Matrix \mathbf{K} , regularization parameter ξ

- 1: Initialize: $k \leftarrow 1, \mathbf{d}_{(0)} \leftarrow 0, \omega_{(0)}^c \leftarrow \mathbf{0}$
 - 2: $\nabla_{(k)} = 2(\mathbf{K}\omega_{(k)}^c + \xi \mathbf{K} I^0 (\mathbf{K}\omega_{(k)}^c - \mathbf{Y}))$
 - 3: $\mathbf{H}_{(k)} = 2(\mathbf{K} + \mathbf{K} I^0 \mathbf{K})$
 - 4: $d_{(k)} = \mathbf{H}_{(k)}^{-1} \nabla_{(k)}$
 - 5: **if** $d_{(k)} \approx d_{(k-1)}$ **then**
 - 6: **exit**
 - 7: **else**
 - 8: $\omega_{(k+1)}^c \leftarrow \omega_{(k)}^c - d_{(k)}$
 - 9: $k \leftarrow k + 1$
 - 10: **go to** Step 2
 - 11: **end if**
 - 12: **return** $\omega_{(k)}^c$
-

In the case of node potential, the optimum value for the regularization parameter ξ and the kernel parameter γ is learned using cross-validation (CV). A 10-fold cross-validation is conducted on the training data with different choices of ξ and γ . This is done by dividing the available

training data in to ten distinct subsets of almost equal sizes and then for each (ξ, γ) pair nine folds are used for training the model while the remaining one fold is used as the validation set. The training of the nine folds consists of one binary SVM for the given class c to separate the members of that class from members of other classes. The cross validation error is then calculated as the average of incorrectly classified data points. Finally, those values of ξ and γ for which CV error is minimum is picked and used it for training SVM based local image classifier for the semantic class c on the complete training data. The same procedure is repeated for each of the K classes to get the parameters of the respective SVM models.

While learning the parameters, it is ensured that the cross-validation experiments are performed on image level. That is, regions from the same image are either in the test or in the training set but never be in different sets. This is important because image blocks belonging to a particular semantic concept tend to be more similar to other, for example neighbouring, blocks in the same image than the regions in other images.

Finally, the parameters A_c and B_c of (14) are learned using maximum likelihood estimation from a training set $(f_c(x_i), y_i)$. To do so, a new training collection $(f_c(x_i), t_i)$ is created based on the given one, where t_i 's are the target probabilities defined as:

$$t_i = \frac{y_i + 1}{2} \quad (26)$$

Then the parameters A_c and B_c are estimated by minimizing the negative log-likelihood of the training data defined as:

$$\sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \quad (27)$$

where p_i is the class probability value as specified in (14).

Any standard non-linear optimization technique can be employed to solve the minimization problem stated in (27). For robustness, in this paper, we use model-trust minimization algorithm [48] for estimating the parameters of the sigmoid function. It should be noted that the training set remains the same while estimating the model parameters for SVM and the sigmoid function.

4.4.2 Learning the parameters of interaction potential

For interaction potential, we need to learn two parameters. That is, the contextual price matrix \mathbb{C} and the parameter θ that determines the degree of closeness among any two local image blocks. As mentioned earlier, we used the normalized confusion matrix of SVM classifier as the contextual price matrix \mathbb{C} .

To learn the parameter θ , we first sample 3000 different point positions $\mathbf{p}_i = (\mathbf{p}x_i, \mathbf{p}y_i)_{1 \leq i \leq 3000}$ from the local image sites in the training collection. Then, the pairwise distances of the entire point set is computed to form a 3000×3000 distance matrix. Afterwards, the median of the resulting distribution is estimated and that value is used as the estimate of θ .

4.5 Inference

Given a test image decomposed in to local image blocks, the proposed semantic context model has to infer an optimal label configuration over these local image blocks using the learned model parameters. Two commonly used principles for inferring labels from an energy function of the form specified in (11) are Maximum a posteriori (MAP) estimation and maximum posterior margins (MPM). MAP maximizes the energy of a globally compatible label configuration while MPM maximizes the energy of the class labels of individual image blocks separately. These two scenarios are clearly depicted by (28) and (29)

$$\mathbf{y}^{\text{MAP}} = \arg \max_{\mathbf{y}} \mathbf{E}(\mathbf{y} | \mathbf{x}) \quad (28)$$

$$y_i^{\text{MPM}} = \arg \max_{y_i} E(y_i | x_i) \quad (29)$$

Since our objective is to minimize the classification error at each local image site, we adopt MPM based inference scheme to find an optimal label configuration over various local image blocks of a given test image. Due to dense pairwise connectivity, the proposed semantic context model have large number of edges. As a result, the conventional inference mechanisms such as graph cut (GC) [49], belief propagation (BP) [50], Markov Chain Monte Carlo (MCMC) [51] etc. often exhibits high computational complexity. Therefore, in this paper, we introduce a simple but effective technique which can speed up the inference process in the proposed semantic context model. The original belief propagation algorithm is reformulated to enhance the computational efficiency of the inference procedure in SCM.

In general, belief propagation is a *message* passing algorithm proposed by Judea Pearl [52] for performing inference on probabilistic graphical models. The goal of BP is to compute the marginal energy values at every node of the underlying graph by means of local message passing. Formally a message, $msg_{i \rightarrow j}(l_j)$ from node i to j can be interpreted as how likely node i thinks that node j will be in the corresponding state l_j . And a complete message, $msg_{i \rightarrow j}$ at a given node i maintains information about all the possible states of its neighbouring node j . As an indication of uninformative prior, all these message entries are usually

initialized with constant values. Depending up on the optimization criteria used, the initial values of messages can be either set as 0 or 1.

A node i passes a message to an adjacent node j only when it has received all incoming messages, excluding the message from the destination node j . The whole algorithm proceeds in iteration as follows. In each step, messages are sent along both directions of every edge. The outgoing message from a particular node is computed based on the incoming messages to this node from all its neighbours in the previous iteration. When the messages converge to a fixed point or a predefined number of iterations has passed, the final beliefs about the states of individual nodes are computed. The belief $b_i(l_i)$ of a node i in its value l_i can be regarded as how likely the node i thinks it should be in state l_i . At the equilibrium point, the belief of a particular node is proportional to the product of messages from its neighbouring nodes and is regarded as an approximation of the required marginal energy value at that node. It should be noted that message passing is only performed between hidden nodes.

With this basic introduction to belief propagation algorithm, we now describes the proposed inference scheme in detail. It involves an iterative message updating step together with a belief calculation equation as depicted in Algorithm 2. Assuming a fixed label vocabulary $l = \{l_1, l_2, \dots, l_K\}$, the entire procedure starts with a message initialization step. In contrast to uninformative priors in the original belief propagation algorithm, messages in the proposed scheme are informative in nature. That is, rather than initializing the message entries with 0's or 1's, the probabilistic outputs of the SVM are assigned as the starting values of the messages (Steps 3–12 of Algorithm 2). More specifically, a message from a hidden node i to another hidden node j is initialized with the probability of the possible label $l_p \in l$ that the node j can presume. This, in turn, permits the proposed inference algorithm to achieve a reasonably fast rate of convergence. Later on, at each iteration t the messages are updated (Steps 14–21 of Algorithm 2) in the following fashion.

$$m_{i \rightarrow j}^{(t+1)}(l_p) = \sum_{l \in L} \Psi_i(l, x_i) \Psi_{ij}(l, l_p) \prod_{k \in N(i) \setminus j} m_{k \rightarrow i}^{(t)}(l) \quad (30)$$

where l_p can be any label state assumed by the image block i under examination and $N(i) \setminus j$ denotes all neighbors of i in SCM other than node j . This is in contrast to the 4-neighbourhood formulation as in the case of basic belief propagation. After T iterations, a belief vector is computed for each node (Steps 23–26 of Algorithm 2). The dimensionality of the belief vector is equal to the total number of concept classes in the label vocabulary and its individual elements can be calculated as:

$$b_i(l_p) = \Psi_i(l_p, x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(l_p) \quad (31)$$

Finally, the label l_p that maximizes the belief of individual node i is selected as the required semantics.

$$y_i = \arg \max_{l_p} b_i(l_p) \quad (32)$$

Algorithm 2 Proposed inference algorithm for SCM

Input: Semantic label vocabulary $l = \{l_1, l_2, \dots, l_K\}$, Total number of local image blocks N , Features extracted from individual blocks of the test image $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, Estimated SCM parameters

```

1: Initialize:  $A = []$ ;  $t=0$ ;  $T=120$ ,  $n=1$ ;
2: calculation of node potentials
3: repeat
4:   for  $j = 1$  to  $K$  do
5:      $A(n) = [A(n) \mid \Psi_n(l_j, x_n)]$ 
6:   end for
7:    $n = n+1$ ;
8: until  $n=N$ 
9: message initialisation
10: for all  $(i, j) \in V$  of SCM do
11:    $msg_{i \rightarrow j}^{(0)} = A(j)$ ;
12: end for
13: updating the messages iteratively
14: repeat
15:   for all  $(i,j) \in E$  of SCM do
16:     for all  $p=1$  to  $K$  do
17:        $m_{i \rightarrow j}^{(t+1)}(l_p) = \sum_{l \in L} \Psi_i(l, x_i) \Psi_{ij}(l, l_p) \prod_{k \in N(i) \setminus j} m_{k \rightarrow i}^{(t)}(l)$ 
18:     end for
19:   end for
20:    $t = t + 1$ ;
21: until  $t = T$ 
22: belief calculation
23: for  $i=1$  to  $N$  do
24:   for  $p=1$  to  $K$  do
25:      $b_i(l_p) = \Psi_i(l_p, x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(l_p)$ 
26:   end for
27: end for
28: semantics identification
29: for  $i=1$  to  $N$  do
30:    $y_i = \arg \max_{l_p} b_i(l_p)$ 
31: end for

```

5 Image representation and similarity measure

This section describes how to transform the identified local image semantics to a meaningful feature vector suitable for the retrieval operation. Furthermore, the distance metric used to estimate the semantic similarity between images is also discussed.

5.1 Semantics-based image representation

The effectiveness of an image retrieval system ultimately depends on the following factors: the correctness of image

representation, the storage requirements of the extracted representation and finally the computational efficiency of the searching scheme. Hence, selecting an appropriate strategy for representing the images and its compactness have great impact on the final retrieval results.

Similar to bag-of-visual-words [1] and latent topics [3–5] based representations, we introduce an intermediate space defined over the concept classes of a predefined label vocabulary upon which the images are represented and the retrieval operation is performed. This intermediate space is referred to as semantic concept space, where each dimension corresponds to a meaningful semantic concept. The number of concept classes in the label vocabulary determines the dimensionality of the semantic space.

To build a semantics based image representation, the region-wise information obtained with semantic context model (SCM) are integrated in a meaningful fashion. For each concept in the semantic concept space, its occurrence frequency in the image under examination is calculated. This leads to a compact, yet discriminative image descriptor. Such kind of descriptor allows us to make a global statement about the amount of a particular concept being present in the image. Another major strength of the proposed semantics based image representation is that they can also be computed on several overlapping or non-overlapping image regions such that its discriminative power can be further improved. That is, a more detailed image representation can be achieved if multiple semantics-based representations are determined on non-overlapping image areas (e.g. top/middle/bottom) and concatenated. The semantic similarity measure employed in image search experiments is discussed in detail in the following section.

5.2 Image similarity measure

Once the semantics based image representation is obtained, the next step is to determine the correspondence between a query and all images present in the database using an appropriate distance metric. Before going into further details of how to compute the semantic affinity between two images using SCM based image descriptor, a brief review of the existing semantic image similarity measures is presented.

As previously mentioned, there exist three different semantics based image representation paradigms for the retrieval operation. In all these approaches, distinct semantic similarity measures are used. Initially, Vogel and Schiele [8] proposed concept occurrence vector (COV) to semantically characterize images for the grouping and retrieval of natural scene collections. In their experiments they identified five scene categories and COV represent images in these categories by tabulating the weightage of each local semantic concept obtained with pre-trained

SVM classifiers. Then, each scene category is represented by the mean over COV of all images belonging to the respective category. This leads to a prototypical representation p^c for each category c and thereby the retrieval operation corresponds to searching for the nearest scene category. Therefore, the typicality of a query image relative to a specific category c is computed by the sum-squared distance (SSD) between the query COV and the prototype p^c of the respective category.

Later on, Li et al. [13] introduced an object bank (OB) based semantic descriptor for image search and retrieval. In this framework, images are represented by object responses (following the) after applying a set of object filters at various locations and scales by using the sliding window approach. For each scale and each object detector, they obtained an initial response map whose value at each location indicates the possibility of the occurrence of that object. A spatial pyramid pooling for the response map is then constructed to capture the spatial location property of objects. Finally, object bank based representation is built by concatenating the max-pooled object filter responses at each grid and each level of the spatial pyramid. Subsequently, Euclidean distance is used to measure the semantic similarity between images.

More recently, Rasiwasia and Vasconcelos [25] introduced semantic multinomial (SMN) upon which images are represented and retrieval decisions are performed. To derive semantic representations of images, they learned statistical models for all the concept classes in a predefined label vocabulary. Images are then represented as vector of posterior probabilities under each of these independent class models and is known as semantic multinomial (SMN). Thus, SMN is a multinomial probability distribution that characterizes relevance of each semantic concept to a given image. Since SMNs are probability distributions, Kullback-Leibler divergence [53] is the natural choice of similarity measure to rank and retrieve images.

In contrast, this work employs histogram intersection (HI) kernel to measure the semantic similarity of a query image in comparison with images already present in the database. The histogram intersection (HI) kernel has been successfully used as a similarity measure for image retrieval when the feature vectors to be compared are in the form of histograms [53]. Having extracted the concept occurrence histograms from the given query and all the images in the database, the semantic similarity can be estimated as:

$$Sim(\mathbb{H}^q, \mathbb{H}^d) = \sum_{i=1}^K \min(h_i^q, h_i^d) \quad (33)$$

where K is the cardinality of the label vocabulary; \mathbb{H}^q and \mathbb{H}^d represents the semantics based representations of the

query and the database images; h_i^q, h_i^d are the i th component of the feature vectors \mathbb{H}^q and \mathbb{H}^d respectively.

To further speed up the search operation, indexing structures can be incorporated in to the proposed image retrieval framework. This helps images in the database to be organized in an efficient and effective manner. Each dimension of the semantics based image representation signifies a particular concept class from a predefined label vocabulary. Hence, inverted file structure [55] is an appropriate choice for SCM based image retrieval model to organize images belonging to various datasets. In an inverted file, a separate entry is maintained for each of the semantic concept class followed by a list of image identifiers which is having non-zero occurrence frequency for that particular concept class. In order to determine image similarity, a subset of images are initially filtered out from the original collection by intersecting those rows in the inverted index having concept classes with non zero entries in the given query. The relevance score is then calculated only between the images in the resulting subset and the given query. As a result, substantial improvement in search speed is accomplished.

6 Data sets

The proposed SCM model captures various contextual information to infer high-level image descriptors for the retrieval operation. Hence, appropriate data sets are required to evaluate the efficiency of the SCM model and the resulting image retrieval system. It is essential to provide the semantic modeling framework with enough training images containing multiple concept categories under different conditions. The Corel natural scene database [8] and the Groningen natural image database [56] serves as good candidates towards this direction. These two datasets are challenging given the respective number of concept classes and the intrinsic ambiguities that arise from their definition. However, they do not capture the full variations of individual concept categories and their various contexts. To have more realistic datasets in which images are complex and contain multiple semantic concepts in both background and foreground, we additionally created two new natural image collections from the MIT Places database [57] and the ILSVRC database [58]. In all the datasets there exist predefined query images and corresponding retrieval lists for each one of these queries to facilitate the easy evaluation of the conceived retrieval system. The details of these four image collections are summarized below.

Corel natural scene database: It contains 700 natural scenes selected from the Corel Stock photo collection and is originally introduced by Vogel and Schiele [9]. They identified that local image regions in this collection can be mapped into any of the

nine predefined semantic categories. These categories constitute the finite label vocabulary and it includes: $\{flower, foliage, grass, ground, rocks, sand, sky, trunks, water\}$. The images in this collection are selected in such a way that they possess high intra class variance within each semantic concept. The typical size of the images is 720×480 (landscape format) or 480×720 (portrait format).

Groningen natural image database: It is a collection of Dutch natural scene images obtained with a Kodak DCS420 digital camera at a resolution of 760×520 pixels. This mainly involves 3000 images of outdoor scenes, especially natural landscapes. This data set includes highly varying scene categories, for example, the variations in the texture of “foliage” viewed from various distances, the change in color composition of “sky” viewed at regular time intervals during the day time etc. By intuitive inspection of images in the data set, we identified eight discriminant local semantic categories to model the images in a semantic concept space and it includes $\{bushes, foliage, grassland, ground, mountain, sand, sky\}$.

MIT Places-4000 database: It is a scene centric database consists of 4000 images selected from the MIT Places205 collection [57]. The new dataset is constructed in such a way that it contains images for which the same semantic concept may appear in different contexts and their appearance may vary according to these contexts. Here also a set of concept classes are selected in advance that covers the semantic elements of images for experimental purpose. These classes are: $\{cloud, field, flowers, land, mountain, road, rock, shrubs, snow, soil, trunk, water\}$. Therefore, an in depth analysis of the effect of contextual knowledge in improving the semantic modeling ability can be thoroughly investigated.

ILSVRC-5000 database: It is a subset of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [58] and all the images have a fixed resolution of 640×480 . This dataset exhibits notable diversity with images depicting a great variety of semantic concepts having variable appearances, positions, lighting conditions and view points. 12 concept classes are identified to characterize the local image regions. These classes are: $\{flower, grass, ground, leaves, mountain, rock, sand, snow, soil, sky, water, vegetation\}$. Thus, it represents a realistic semantic image modeling scenario. Moreover, this dataset meets all the requirements to evaluate an image retrieval system, due to its moderate size and heterogeneous content.

7 Evaluation of semantic context model

In this section, the ability of the proposed semantic context model to correctly recognize various local image semantics is evaluated. While judging the competence of any

semantic image modeling scheme its expressive power, the rate of convergence of the inference algorithm and overall computational time needs to be thoroughly analysed. In general, the conceived model is expected to capture as much contextual information as possible to resolve local ambiguities and thereby improve the concept recognition rate in complex images. Moreover, the inference procedure needs to be computationally tractable so that it can be applied to real-life scenarios. Therefore, the classification accuracy, convergence rate and the overall execution time of the proposed method is calculated and compared with already existing works such as DRF [7], multi-scale CRF [32] and Dense-CRF [33]. The hardware platform for this work is an Intel Core i7-4570 3.4 GHz CPU, equipped with 16 GB of RAM. All experiments are coded and executed within Matlab 7.14 (R2012a) environment in Windows 7 operating system.

7.1 Experimental protocol

As part of semantic image modeling, ground truth data for SCM training and testing has to be generated for each of the above mentioned data sets. For our experiments, the ground truth is obtained by following the procedure already been explained in Sect. 4.1.1. Afterwards, the class labels and the corresponding scattering transform coefficients of local image blocks in the training set are used to estimate the parameters of the semantic context model. When a test image is presented to the model, the features extracted from its local image blocks together with the parameters already learned from the training data are given as inputs to the modified belief propagation algorithm introduced in Sect. 4.5 to infer the local image semantics.

All experiments have been fully cross validated in order to average out the fact that in diverse databases certain test sets perform better than others. In this paper, 10-fold cross-validation on image level is employed to evaluate the proposed semantic context model in comparison with other approaches. In 10-fold cross validation the available data is initially partitioned into 10 equally sized groups or fold. Subsequently 10 iterations of training and validation are performed such that within each iteration a different fold of data is held out for validation while the remaining 9 fold are used for learning the model. The performance of the proposed model on each validation fold can be tracked using some pre-determined evaluation metric. Upon completion, 10 groups of performance metrics will be obtained and the final measure presented in the subsequent section is the average value computed over these 10 trials.

7.2 Evaluation metric

To evaluate the effectiveness of SCM, we treat the problem of semantic concept discovery from local image blocks as a special case of multi-class classification. Then, the competence of SCM is measured in terms of its ability to generalize unseen images. To do so, confusion matrix [59] is calculated based on the classification result at the level of local image blocks. Confusion matrix maintains information about the actual and the predicted outcomes of a classification system. Table 1 depicts the structure of the confusion matrix for a given class c . In this table, TP_c , TN_c , FP_c , and FN_c respectively stands for the number of true positive, true negative, false positive and false negative instances generated by the model for the specified class c . Based on the confusion matrix, several standard performance metrics have been defined to quantify the efficiency of the proposed model and these measures are summarized in Table 2.

In addition, energy plots are used to evaluate the convergence rate of the modified belief propagation algorithm. It has been already mentioned in Sect. 4.3 that the energy value of the required label configuration should be as small as possible and the inference procedure should ensure a constant reduction in energy values for each iteration until it reach an equilibrium state. Therefore, energy values are computed using (11) and are plotted against the number of iterations to assess the convergence property of the proposed inference scheme.

7.3 Results and discussion

Experiments are conducted on the data sets already mentioned in Sect. 6 to validate the potential of SCM in recognizing local image semantics and to compare its performance with existing methods. The performance evaluation metrics are subsequently determined for each of the methods based on confusion matrix results. Table 3 summarizes the obtained results in terms of the overall classification accuracy, P-micro, P-macro, R-micro, R-macro, F-micro, F-macro values. The experimental

Table 1 Confusion matrix returned by SCM for a given class c

		Expert judgement	
		True	False
Model judgement	True	TP_c	FP_c
	False	FN_c	TN_c

Table 2 Metrics used for evaluating the proposed SCM framework

Evaluation measure	Formula	Intuitive meaning
Accuracy (Acc)	$\frac{\sum_{c=1}^C (TP_c + TN_c)}{\sum_{c=1}^C (TP_c + TN_c + FP_c + FN_c)}$	It is the ratio of total number of correct predictions over the total number of instances evaluated
Precision (P_c)	$\frac{TP_c}{TP_c + FP_c}$	It is the percentage of positive predictions that are correct
Recall (R_c)	$\frac{TP_c}{TP_c + FN_c}$	It is the percentage of positive labeled instances that are predicted as positive
F1-score ($F1_c$)	$2 \times \frac{P_c \times R_c}{P_c + R_c}$	It is defined as the harmonic mean between recall and precision values
Micro precision (P-micro)	$\frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C TP_c + FP_c}$	It is the precision calculated over the test data as a whole
Micro recall (R-micro)	$\frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C TP_c + FN_c}$	It is the recall calculated over the test data as a whole
Micro F1 score (F-micro)	$2 \times \frac{P\text{-micro} \times R\text{-micro}}{P\text{-micro} + R\text{-micro}}$	It is the harmonic mean between micro precision and micro recall
Macro precision (P-macro)	$\frac{\sum_{c=1}^C P_c}{C}$	It is the average of the precision values calculated separately for each of the concept classes in the label vocabulary
Macro recall (R-macro)	$\frac{\sum_{c=1}^C R_c}{C}$	It is the average of the recall values calculated separately for each of the concept classes in the label vocabulary
Macro F1 score (F-macro)	$\frac{\sum_{c=1}^C F1_c}{C}$	It is the average of F1 measures calculated separately for each of the concept classes in the label vocabulary

Table 3 Performance of SCM in comparison with other existing semantic image modeling frameworks

Data set used	Method used	Accuracy	P-micro	R-micro	P-macro	R-macro	F-micro	F-macro
Corel natural scene database	DRF	0.767	0.764	0.748	0.738	0.714	0.750	0.712
	mCRF	0.801	0.798	0.771	0.769	0.753	0.783	0.738
	Dense-CRF	0.821	0.828	0.801	0.799	0.783	0.813	0.776
	SCM	0.925	0.910	0.888	0.904	0.881	0.895	0.872
Groningen natural image database	DRF	0.770	0.761	0.742	0.731	0.710	0.751	0.711
	mCRF	0.798	0.785	0.772	0.751	0.733	0.777	0.741
	Dense-CRF	0.817	0.806	0.791	0.792	0.761	0.798	0.751
	SCM	0.889	0.892	0.865	0.846	0.823	0.878	0.855
MIT Places-4000 database	DRF	0.782	0.778	0.765	0.758	0.731	0.771	0.720
	mCRF	0.791	0.799	0.779	0.770	0.746	0.791	0.752
	Dense-CRF	0.810	0.820	0.803	0.794	0.771	0.810	0.770
	SCM	0.879	0.887	0.870	0.868	0.847	0.880	0.861
ILSVRC-5000 database	DRF	0.713	0.710	0.693	0.674	0.663	0.701	0.665
	mCRF	0.731	0.735	0.720	0.708	0.680	0.727	0.686
	Dense-CRF	0.749	0.758	0.734	0.723	0.708	0.745	0.686
	SCM	0.831	0.840	0.816	0.811	0.786	0.830	0.803

results indicate that the proposed semantic context model has significantly outperformed the baseline methods in terms of all considered evaluation criteria in all the four data sets. The advantage of the newly introduced interaction potential is clear, and on an average it provides 10 % performance improvement over the state-of-the-art semantic modeling approach.

To get a clear understanding of the expressive power of SCM and to what extent it can resolve the ambiguities in concept discovery, the semantic modeling experiments in the Corel natural scene collection is further analysed in detail. The confusion matrix which summarizes the block-wise classification results over all the nine concept classes of the Corel natural scene data set is shown in Table 4. This

Table 4 The confusion matrix obtained for various semantic image modeling schemes when the validation set corresponding to Corel natural scene database is utilized (the results are in percent)

DRF	Flower	Foliage	Grass	Ground	Rock	Sand	Sky	Trunk	Water
Flower	79.24	9.46	4.42	1.85	0	0	0	4.93	0
Foliage	6.25	80.20	3.84	1.70	1.70	1.13	0	5.11	0
Grass	4.30	8.45	78.64	2.15	1.61	1.61	0	3.22	0
Ground	1.39	2.32	4.18	77.20	7.90	6.97	0	0	0
Rock	0	1.53	1.53	9.96	78.92	6.84	0	1.14	0
Sand	0	0	3.00	9.50	7.00	79.00	0	1.50	0
Sky	0	1.63	0	0	4.34	1.35	76.90	3.53	1.22
Trunk	2.94	8.82	5.14	4.41	0	0	0	78.67	0
Water	0	0	0	1.09	2.50	4.37	12.60	0	78.75
mCRF	Flower	Foliage	Grass	Ground	Rock	Sand	Sky	Trunk	Water
Flower	82.99	7.98	3.84	1.24	0	0	0	3.67	0
Foliage	5.55	82.96	2.98	1.16	1.21	1.08	0	4.75	0
Grass	3.87	7.32	81.84	1.92	1.14	1.12	0	2.86	0
Ground	1.04	1.65	3.68	81.47	6.43	5.67	0	0	0
Rock	0	0	0	8.40	82.64	6.06	0	1.04	0
Sand	0	0	2.34	8.33	6.34	81.14	0	1.09	0
Sky	0	1.12	0	0	3.76	1.02	79.63	2.85	10.97
Trunk	1.86	7.74	4.59	3.64	0	0	0	81.74	0
Water	0	0	0	1.32	2.17	4.10	10.34	0	81.09
Dense-CRF	Flower	Foliage	Grass	Ground	Rock	Sand	Sky	Trunk	Water
Flower	85.80	6.17	3.70	1.85	0	0	0	2.47	0
Foliage	2.84	85.20	5.6	0	1.14	0	0	3.98	0
Grass	2.69	5.84	85.56	1.61	0	1.08	0	2.69	0
Ground	0	0	2.33	86.51	6.05	3.26	0	1.86	0
Rock	0	1.55	1.15	7.05	83.16	5.75	0	1.15	0
Sand	0	0	2.00	7.00	6	83.50	0	0	1.00
Sky	0	1.90	0	0	2.42	0	81.43	3.40	10.05
Trunk	1.47	6.62	3.68	2.21	2.21	0	0	83.82	0
Water	0	0	0	0	2.0	4.06	10.70	0	82.15
SCM	Flower	Foliage	Grass	Ground	Rock	Sand	Sky	Trunk	Water
Flower	92.11	1.36	2.29	1.24	0	0	0	3.00	0
Foliage	1.47	92.78	2.17	0	1.06	0	0	2.52	0
Grass	2.36	3.17	91.33	1.03	0	0	0	2.11	0
Ground	0	0	1.53	90.97	3.69	2.48	0	1.33	0
Rock	0	0	0	5.51	89.87	4.62	0	0	0
Sand	0	0	1.05	5.31	4.02	87.43	0	1.05	1.14
Sky	0	4.88	0	0	3.26	0	87.93	1.63	2.30
Trunk	0	5.86	1.41	0	0	0	2.86	89.87	0
Water	0	0	0	0	2.11	3.06	5.19	0	89.64

highlights the superior performance of SCM over other baseline methods. It can be seen that the concept classes “sky”, “water” and “foliage” obtain the greatest gain from the newly introduced interaction potential. These concept categories possess unique spatial relationships in a scene.

The proposed semantic context model effectively encodes the long-range interactions among these concepts and achieved better results in the classification of local image regions as compared to other approaches.

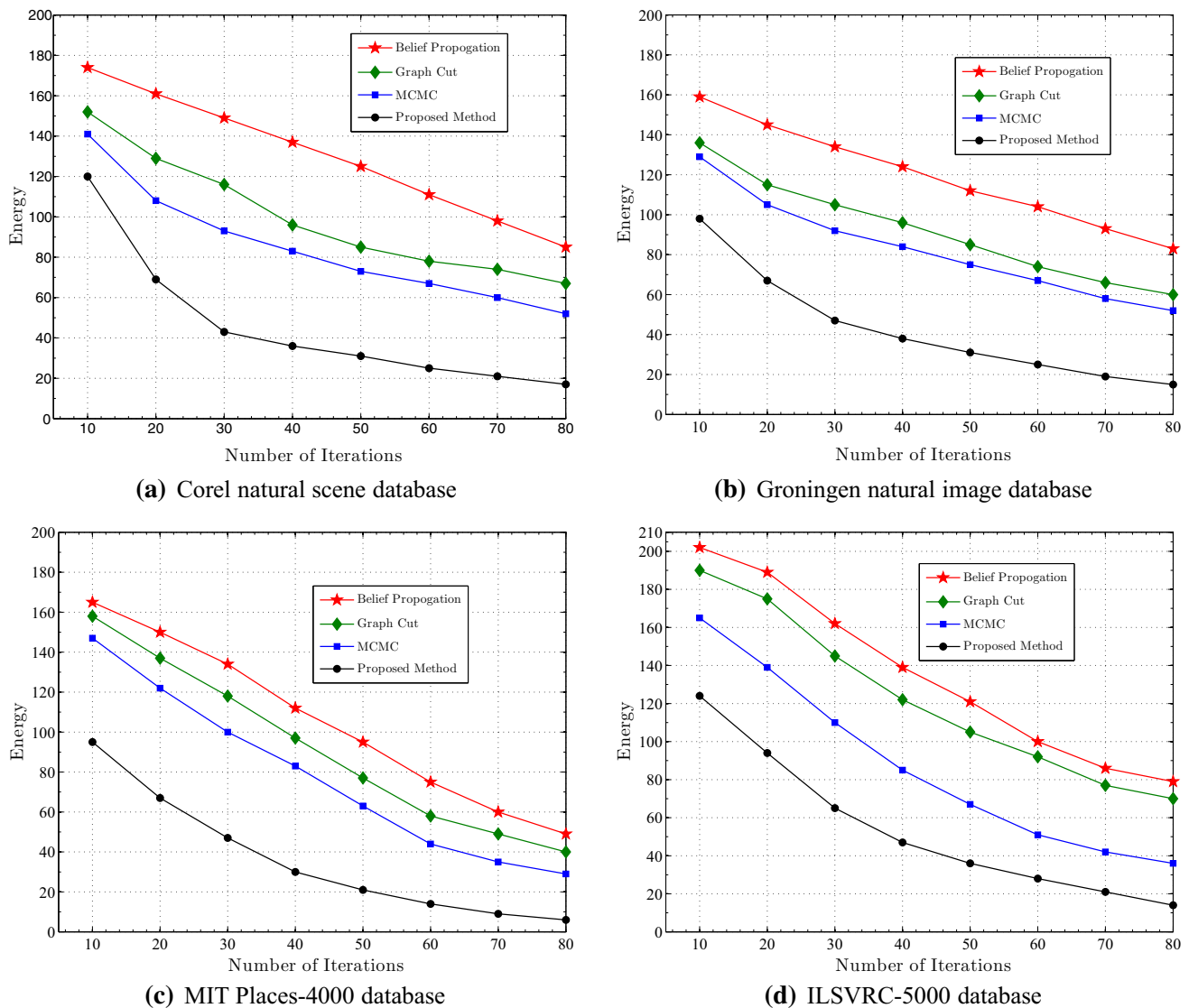


Fig. 6 Energy plots depicting the rate of convergence of the proposed inference algorithm in comparison with the existing algorithms

The dominance of the modified message-passing algorithm over the state-of-the-art inference techniques is demonstrated by means of energy plots. Figure 6 shows the plot of energy values versus number of iterations for the proposed inference scheme and other compared algorithms on various data sets. The energy values are calculated as an average over the test image collections from each data sets. From these results, it can be easily concluded that the proposed inference algorithm exhibits faster convergence rate as compared to the commonly used inference techniques. Among the conventional approaches, Belief propagation (BP) suffers from loopiness of the graph, resulting a poor convergence rate. As they do not provide any uncertainty measure associated with the generated label configuration, Graph cuts based inference are also prohibitively slow. In each iteration, only a few variables of

the current label configuration are allowed to vary in MCMC. This in turn slow down the inference procedure in MCMC. In contrast to this, a single-threaded implementation of the proposed inference algorithm produces an optimal labeling in a limited number of iterations. Therefore, a comparatively lesser number of iterations is required by the proposed inference algorithm for finding an optimal solution.

Finally, Table 5 summarizes the computation time of various semantic modeling schemes averaged over ten folds of individual data sets. In Algorithm 2, the calculation of node potentials requires $\mathcal{O}(N)$ time, where N is the total number of nodes in SCM. It is then followed by a message updating step which iterate over all the labels K and all the edges $|E|$ of SCM, leads to a computational complexity of $\mathcal{O}(TK(|E|+N))$, where T is the number of iteration. Finally,

Table 5 Computation time of the proposed SCM framework in comparison with existing image modeling schemes

Data set used	Semantic modeling scheme	Computation time for training + inference (s)
Corel natural scene database	DRF	1026
	mCRF	1776
	Dense-CRF	871
	SCM	826
Groningen natural image database	DRF	2232
	mCRF	2927
	Dense-CRF	1041
	SCM	997
MIT Places-4000 database	DRF	3675
	mCRF	3008
	Dense-CRF	2017
	SCM	1672
ILSVRC-5000 database	DRF	4763
	mCRF	4176
	Dense-CRF	3171
	SCM	2978

The computation time is calculated as an averaged of 10-fold cross validation

belief calculation for individual node requires summation over all the K labels, which takes $\mathcal{O}(NK)$ time. Thus, the inference procedure summarized in Algorithm 2 will have a total complexity $\mathcal{O}(N+TK(|E|+N)+NK)$. Hence, all these results indicate the fact that the proposed semantic context model outperform the existing frameworks in terms of both classification accuracy and speed.

8 Evaluation of semantic descriptors in image search

This section evaluates the retrieval efficiency of the proposed semantics based image descriptors and provides empirical evidences to demonstrate its superior performance over the traditional approaches. The quantitative indices used to measure image retrieval accuracy, the experimental set up and the results of image search experiments conducted in various benchmark data sets are described in detail in the following subsections.

8.1 Evaluation metric

For evaluating its effectiveness, the relevance of the retrieval results yielded by the proposed model in response to a given query image collection needs to be thoroughly analysed. Among various evaluation metrics available precision and recall are possibly the most commonly used

measures for assessing the performance of any image retrieval system. Precision denotes the potential of a system to retrieve only relevant items while recall specifies the ability of a system to return all the relevant items from the database for a given query image.

$$Precision = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (34)$$

$$Recall = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the set}} \quad (35)$$

Since precision and recall are inversely proportional to each other, it is quite hard to differentiate two pairs of precision-recall values. Therefore, several attempts have been made to combine these two measures for deriving a single value which quantifies the overall retrieval performance. One such metric is the 11-point interpolated average precision. In the context of an ordered retrieval list, the top k results often constitute the required outcome for a given query. Interpolated precision values are calculated at 11 recall levels $\{0.0, 0.1, 0.2, \dots, 1.0\}$ for such an outcome. The interpolated precision p_{int} at a recall level r_i is defined as the largest observed precision for any recall value r between r_i and r_{i+1} :

$$P_{int}(r_i) = \max_{r_i \leq r \leq r_{i+1}} Precision(r) \quad (36)$$

Then, the average of interpolated precisions across all values of recall between 0 and 1 gives 11-point interpolated average precision. Another important single value based evaluation measure is the mean average precision (mAP). For a set of m query images the Mean Average Precision is defined as:

$$Mean\ Average\ Precision(mAP) = \frac{1}{m} \sum_{i=1}^m AP(q) \quad (37)$$

where $AP(q)$ is the average precision for a given query q and is defined as the ratio of the sum of precision values from rank positions where a relevant image is found in the retrieval result to the total number of relevant images in the database.

In addition, the performance of image search engines can be further characterized by the following measures such as F1 score, Precision at k ($P@k$) and R-precision. Among them, F1 score is defined as the harmonic mean of precision and recall

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (38)$$

Precision at k ($P@k$) is the value of precision calculated at a cut off point k , where k is the number of top items considered for evaluating the performance of the system. Similarly, R-Precision for a given query is defined to be the precision after retrieving R images from the image data base and is expressed as:

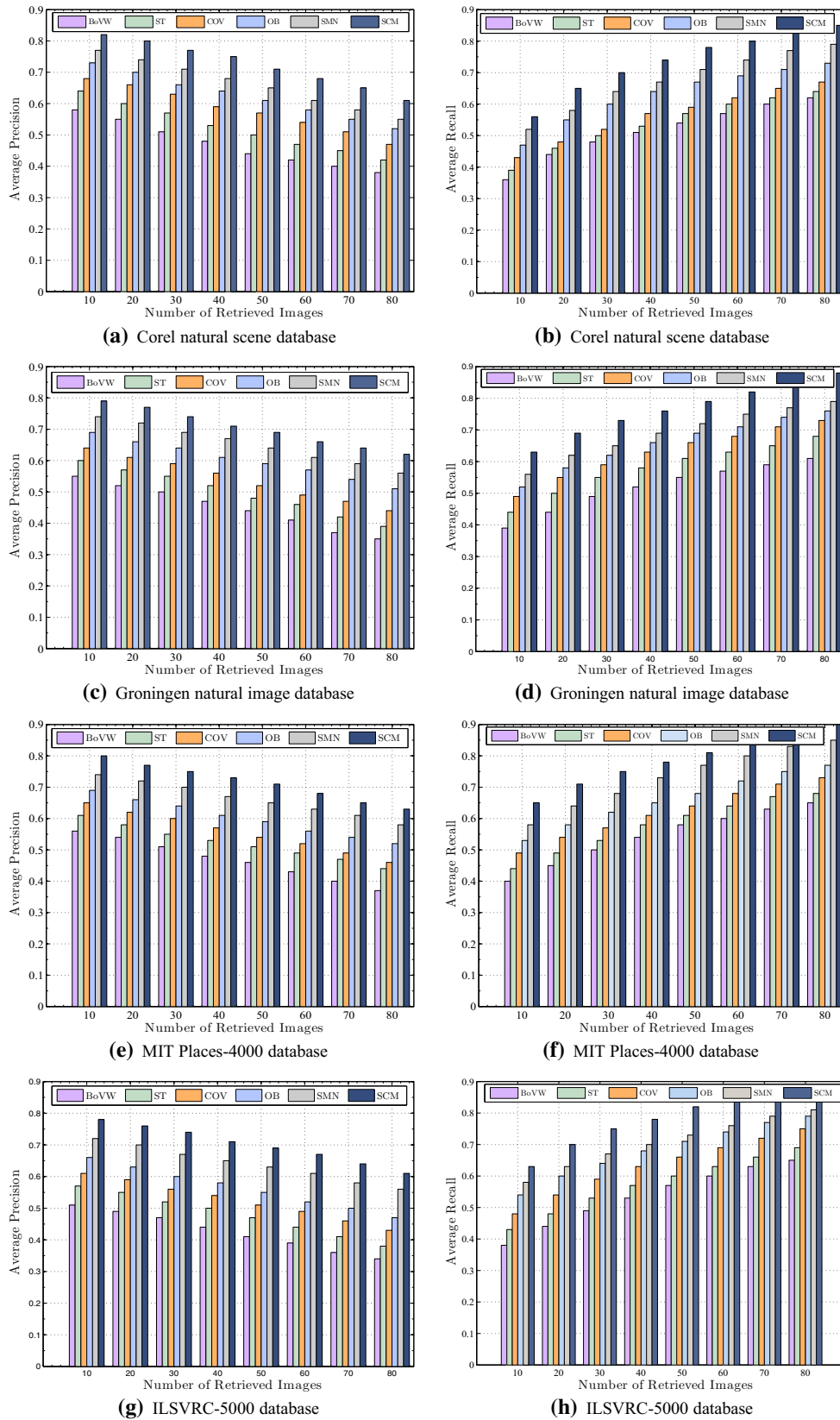


Fig. 7 Comparative evaluation of average precision and average recall across various image retrieval frameworks

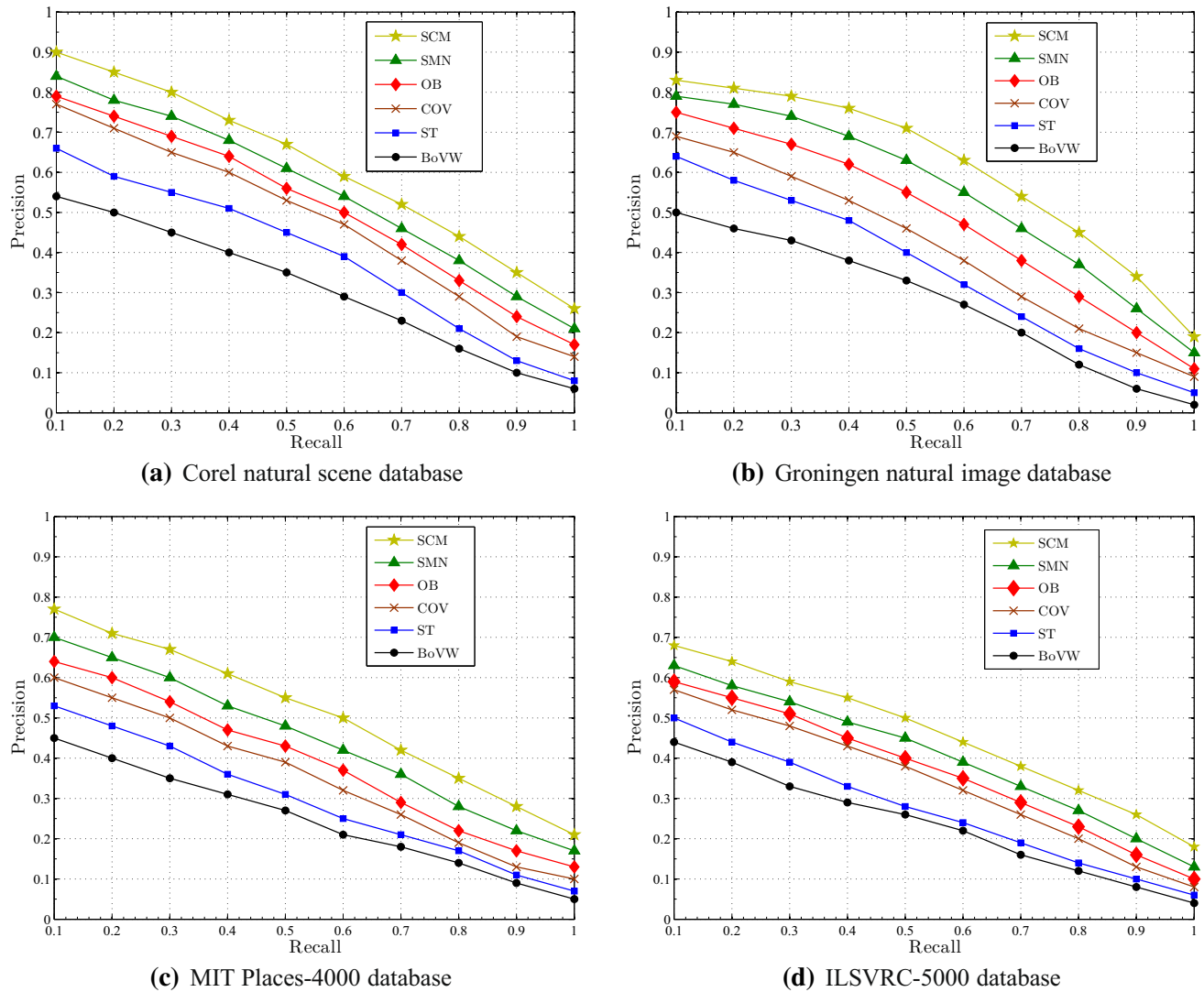


Fig. 8 Comparison of various image retrieval systems based on 11-point interpolated average precision

$$R\text{-Precision} = \frac{1}{R} \sum_{j=1}^R \text{Rel}(j) \quad (39)$$

where R is the total number of relevant images in the database for the given query and $\text{Rel}(j)$ is an indicator function returns the value 1 when the image present at the j th location of the retrieval list is relevant with respect to the given query. For the proposed model, a comprehensive analysis of the retrieval performance is done using almost all of the above mentioned evaluation metrics.

8.2 Experimental set up

The retrieval experiments are conducted using 10-fold cross validation technique. In order for this to work, images in individual database are arbitrarily divided into ten folds of approximately equal sizes. In each successive

experiments, nine image subsets are used for training purpose and the remaining one will act as the query. Thus, each subset is used once as the query. The evaluation metrics presented in subsequent sections refers to the average values computed over these ten trials. The same cross-validation folds are used for all other compared methods.

8.3 Retrieval results and discussion

In this section, the retrieval efficiency of the SCM based framework is compared with state-of-the-art systems. The proposed SCM based semantic descriptors for image retrieval is respectively compared with four existing techniques in performance namely, concept occurrence vector (COV) [9], object bank (OB) [14], automatic annotation and retrieval [21] with scattering transform based visual

Table 6 Precision, F1-score and mean average precision (mAP) for each data sets calculated at selected rank thresholds of 10 and 20 along with average R-precision values

Method Used	P@10	P@20	F1@10	F1@20	mAP@10	mAP@20	Average R-Precision
Corel natural scene database							
BoVW	0.518	0.539	0.532	0.554	0.522	0.547	0.588
ST	0.543	0.571	0.561	0.583	0.556	0.592	0.644
COV	0.606	0.638	0.625	0.649	0.614	0.639	0.681
OB	0.638	0.664	0.651	0.676	0.644	0.667	0.704
SMN	0.662	0.681	0.679	0.699	0.663	0.686	0.733
SCM	0.725	0.747	0.738	0.756	0.724	0.742	0.797
Groningen natural image database							
BoVW	0.468	0.504	0.513	0.538	0.504	0.528	0.579
ST	0.512	0.535	0.527	0.545	0.526	0.541	0.626
COV	0.571	0.597	0.586	0.604	0.583	0.606	0.694
OB	0.608	0.632	0.622	0.647	0.613	0.633	0.712
SMN	0.638	0.667	0.653	0.681	0.644	0.667	0.743
SCM	0.692	0.732	0.711	0.744	0.704	0.738	0.805
MIT Places-4000 database							
BoVW	0.451	0.487	0.495	0.514	0.471	0.496	0.544
ST	0.505	0.532	0.516	0.537	0.519	0.534	0.592
COV	0.563	0.586	0.574	0.592	0.573	0.592	0.656
OB	0.592	0.628	0.605	0.623	0.612	0.636	0.691
SMN	0.625	0.654	0.633	0.654	0.635	0.652	0.732
SCM	0.683	0.717	0.693	0.713	0.692	0.717	0.784
ILSVRC-5000 database							
BoVW	0.404	0.447	0.435	0.456	0.425	0.447	0.513
ST	0.449	0.487	0.464	0.499	0.468	0.485	0.552
COV	0.518	0.552	0.546	0.564	0.533	0.558	0.598
OB	0.543	0.578	0.567	0.584	0.563	0.584	0.648
SMN	0.579	0.607	0.591	0.618	0.591	0.620	0.680
SCM	0.638	0.676	0.668	0.686	0.649	0.685	0.744

features (ST), semantic multinomial (SMN) [28] and the tf-idf weighted bag-of-visual-words model (BoVW) [60]. The outcomes of image search experiments conducted on various data sets are presented in the remainder of this section.

Initially, the proposed model is evaluated as a function of average precision/recall values against the top k -retrieved images. For each image in the query collection, a set of precision and recall values are calculated by varying the number k of the returned images. For different values of k , the average precision and recall values are calculated over all images in the query set. The final results are then shown in Fig. 7 for all the four data sets with the number n of retrieved images on the X-axis and the precision/recall values averaged over a set of queries for each data sets in Y-axis. In this experiment, the average recall values obtained are relatively better for the proposed approach and it significantly increases with the number n of returned images. These results validate the efficacy of the proposed semantics based image representation scheme for the desired retrieval operation.

Figure 8 shows the 11-point interpolated average precision curve for the proposed image retrieval scheme in comparison with all the other approaches. For all the image collections, it seems that the precision achieved by the proposed SCM based semantic image representation is significantly better than existing ones across all values of recall.

In order to further confirm the retrieval efficiency, the performance of the SCM framework and other existing approaches are compared in terms of precision, F1 score and mAP values at selected rank thresholds of 10 and 20. The obtained values are summarized in Table 6. Apart from this, the table includes R-precision values achieved by various image retrieval models averaged over the set of all query images. If we assume a user will only view the top 10 or 20 results returned by the retrieval model, then an overall relative improvement of up to 6 % would be possible with the SCM based framework. These findings once again underline the effectiveness of the proposed SCM model in semantics based image search.

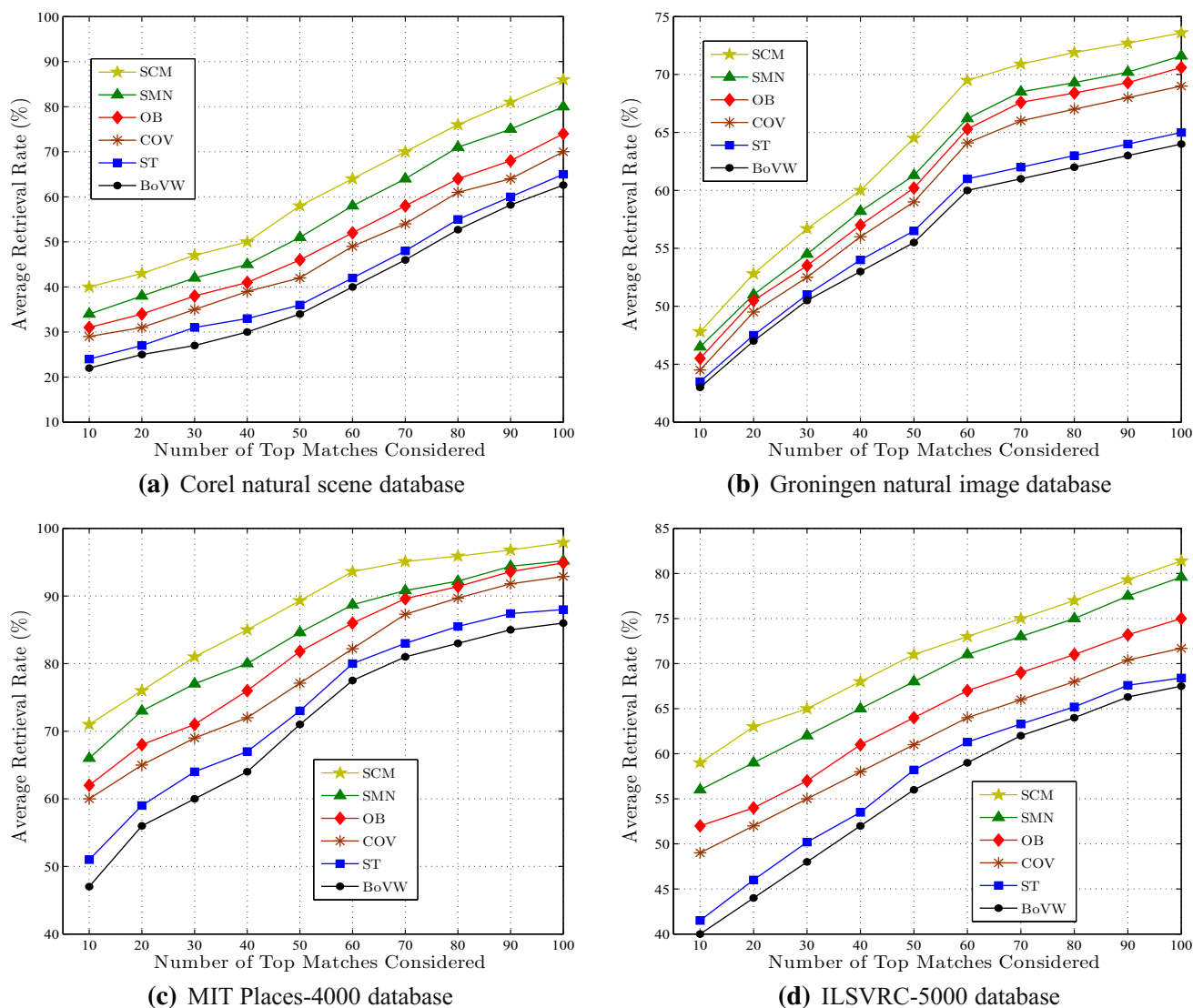


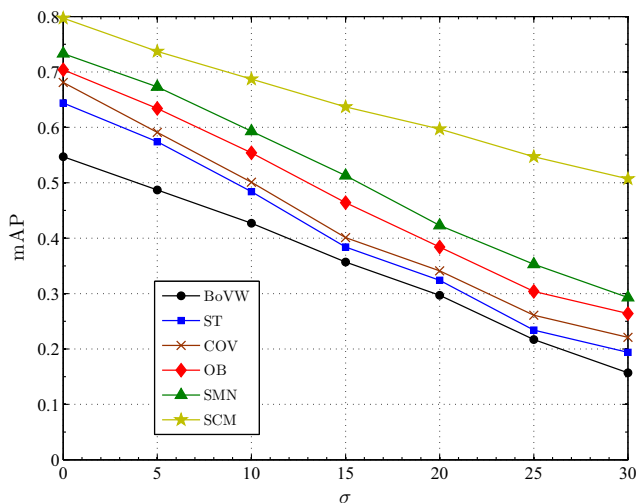
Fig. 9 Comparison of proposed SCM framework with other existing models in terms of average retrieval rate

Finally, Fig. 9 shows the comparative average retrieval rates for all the six retrieval models. It is evident from Fig. 9 that the average retrieval rate of the proposed model is always higher than the existing approaches. Hence, SCM model gives the best performance in terms of average retrieval rates for all the evaluated data sets.

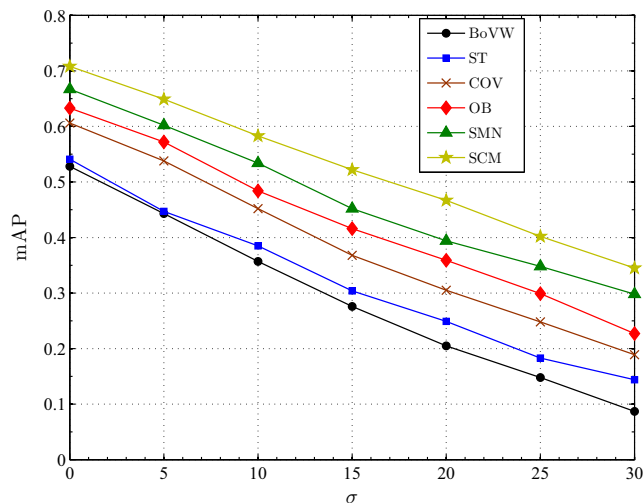
In general, the quality of the training images significantly affects the retrieval performance. Image quality is usually determined by many factors and among them noise contents present in images needs primary attention. Hence in this paper, retrieval experiments are carried out by varying the quality of the training images with noise addition. To this end, Gaussian noise is added to the training images and mean average precision (mAP) obtained for various retrieval systems are plotted against different values of the variance σ as shown in Fig. 10. By

analysing the result it can be seen that the retrieval performance decreases with the increase of σ for all the existing approaches. However, the influence of noise is comparatively less in SCM based formulation due to the contextual information incorporated in the semantic modeling step.

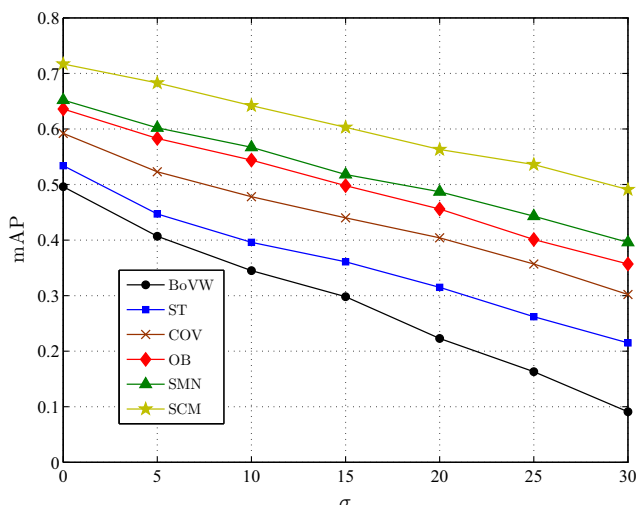
All these encouraging results reveals the fact that SCM based image modeling consistently surpassed all their available counterparts in terms of retrieval efficiency. The proposed semantic context model is capable of extracting highly rich semantic information from complex images in the form of a compact yet discriminative feature vector. The inferred semantic descriptor encode the percentage of each concept present in individual images and thus the similarity between images can be estimated at a much higher level of abstraction. Hence, the SCM based



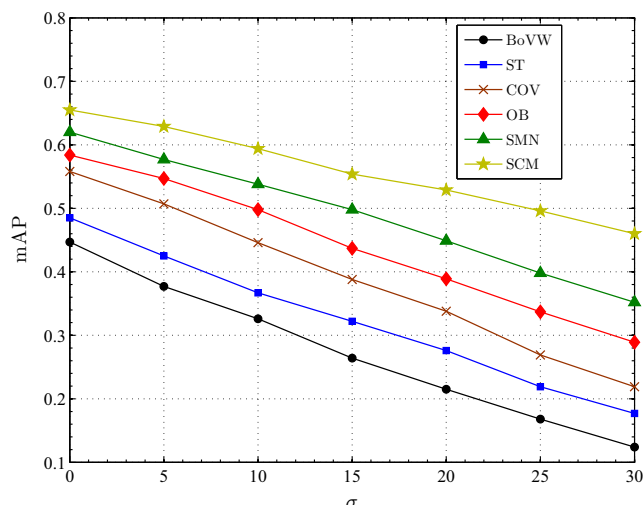
(a) Corel natural scene database



(b) Groningen natural image database



(c) MIT Places-4000 database



(d) ILSVRC-5000 database

Fig. 10 The influence of added Gaussian noise on image retrieval performance

approach to model the semantic contents of images for the retrieval task is found to be promising.

9 Conclusion

In this paper, we proposed a new framework called semantic context model (SCM) to represent images in a semantic concept space for the task of image retrieval. SCM is formulated as a fully connected undirected graphical model with newly introduced potential functions and an efficient inference procedure for discovering local image semantics. It has been shown that SCM is suitable to encode long-range dependencies among semantically meaningful local image blocks without requiring much human efforts. Experimental

results demonstrated the fact that the proposed model is capable to resolve local ambiguities while recognizing image semantics and can further boost the classification performance as compared to other existing image modeling approaches. Moreover, it considerably reduces the semantic gap problem in image retrieval by deriving a compact but discriminative image representation. Retrieval experiments are carried out on four different datasets. The obtained results show that the proposed semantic image representation has strong discrimination power and is more reliable than state-of-the-art semantic image descriptors in terms of overall retrieval performance. In future, we would like to incorporate more sophisticated potential functions in to the proposed SCM formulation to further improve the concept recognition accuracy.

References

- Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Proceedings of Ninth IEEE international conference on computer vision, vol 2, pp 1470–1477
- Duan M, Wu X (2010) Visual polysemy and synonymy: toward near-duplicate image retrieval. *Front Electr Electron Eng China* 5(4):419–429
- Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42(1–2):177–196
- Zhang R, Zhang Z (2007) Effective image retrieval based on hidden concept discovery in image database. *IEEE Trans Image Process* 16(2):562–572
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Biederman I, Mezzanotte R, Rabinowitz J (1982) Scene perception: detecting and judging objects undergoing relational violations. *Cogn Psychol* 14(2):143–177
- Kumar S, Hebert M (2006) Discriminative random fields. *Int J Comput Vis* 68(2):179–201
- Yu L, Xie J, Chen S (2012) Conditional random field-based image labelling combining features of pixels, segments and regions. *IET Comput Vis* 6(5):459–467
- Vogel J, Schiele B (2007) Semantic modeling of natural scenes for content-based image retrieval. *Int J Comput Vis* 72(2):133–157
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, vol 1, pp 886–893
- Bay, H., Tuytelaars, T., Van Gool, L (2006) Surf: speeded up robust features. In: Proceedings of the 9th European conference on computer vision, pp 404–417
- Tola E, Lepetit V, Fua P (2010) Daisy: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans Pattern Anal Mach Intell* 32(5):815–830
- Li LJ, Su H, Lim Y, Fei-Fei L (2014) Object bank: an object-level image representation for high-level visual recognition. *Int J Comput Vis* 107(1):20–39
- Torresani L, Szummer M, Fitzgibbon A (2010) Efficient object category recognition using classemes. In: Proceedings of 11th European conference on computer vision. Springer, Berlin, Heidelberg, pp 776–789
- Chan A, A., Vasconcelos., N, (2005) Probabilistic kernels for the classification of auto-regressive visual processes. In: Proceedings of IEEE conference on computer vision and pattern recognition, vol 1, pp 846–851
- Zhang H, Berg A, Maire M, Malik J (2006) Svm-knn: discriminative nearest neighbor classification for visual category recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2126–2136
- Cai D, He X, Han J (2007) Efficient kernel discriminant analysis via spectral regression. In: Proceedings of Seventh IEEE international conference on data mining, pp 427–432
- Grauman K, Darrell T (2007) The pyramid match kernel: efficient learning with sets of features. *J Mach Learn Res* 8:725–760
- Bosch A, Zisserman A, Munoz X (2008) Scene classification using a hybrid generative/discriminative approach. *IEEE Trans Pattern Anal Mach Intell* 30(4):712–727
- Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, pp 119–126
- Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 524–531
- Sivic J, Russell B, Efros A, Zisserman A, Freeman W (2005) Discovering object and their localization in images. In: Proceedings of the tenth IEEE international conference on computer vision, vol 1, pp 370–377
- Sudderth E, Torralba A, Freeman W, Willsky A (2005) Learning hierarchical models of scenes, objects and parts. In: Proceedings of the tenth IEEE international conference on computer vision, vol 2, pp 1331–1338
- Carneiro G, Chan A, Moreno P, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans Pattern Anal Mach Intell* 29(3):394–410
- Rasiwasia N, Vasconcelos N (2012) Holistic context models for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 34(5):902–917
- Bar M (2004) Visual objects in context. *Nat Rev Neurosci* 5(8):617–629
- Bar M, Ullman S (1993) Spatial context in recognition. *Perception* 25:343–352
- Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge, p 1280
- Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning, pp 282–289
- Kohli P, Torr PH (2009) Robust higher order potentials for enforcing label consistency. *Int J Comput Vis* 82(3):302–324
- He X, Zemel RS, Carreira-Perpindn MA (2004) Multiscale conditional random fields for image labeling. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 695–702
- Krhenbhl P, Koltun V (2012) Efficient inference in fully connected crfs with Gaussian edge potentials. [arXiv:1210.5644](https://arxiv.org/abs/1210.5644)
- Efron B (1975) The efficiency of logistic regression compared to normal discriminant analysis. *J Am Stat Assoc* 70(352):892–898
- Kindermann R, Snell JL (1980) Markov random fields and their applications, vol 1. American Mathematical Society, Providence
- Dagli C, Huang TS (2004) A framework for grid-based image retrieval. In: Proceedings of the 17th IEEE international conference on pattern recognition, vol 2, pp 1021–1024
- Huiskes MJ, Lew MS (2008) The MIR Flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on multimedia information retrieval, pp 39–43
- Bruna J, Mallat S (2013) Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell* 35(8):1872–1886
- Mallat S (2012) Group invariant scattering. *Commun Pure Appl Math* 65(10):1331–1398
- Andn J, Mallat S (2011) Multiscale scattering for audio classification. In: ISMIR, pp 657–662
- Oyallon E, Mallat S, Sifre L (2013) Generic deep networks with wavelet scattering. [arXiv:1312.5940v3](https://arxiv.org/abs/1312.5940v3)
- Lee TS (1996) Image representation using 2D Gabor wavelets. *IEEE Trans Pattern Anal Mach Intell* 18(10):959–971
- Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 10(3):61–74
- Wu TF, Lin CJ, Weng RC (2004) Probability estimates for multi-class classification by pairwise coupling. *J Mach Learn Res* 5:975–1005
- Sutton C, McCallum A (2007) Piecewise pseudo likelihood for efficient training of conditional random fields. In: Proceedings of

- the 24th ACM international conference on machine learning, pp 863–870
46. Beck A, Ben-Tal A (2006) On the solution of the Tikhonov regularization of the total least squares problem. *SIAM J Optim* 17(1):98–118
 47. Kelley CT (1999) *Iterative methods for optimization*. Frontiers in applied mathematics. Siam, Philadelphia, PA
 48. Gill PE, Murray W, Wright MH (1981) *Practical optimization*, vol 5. Academic press, London
 49. Lempitsky V, Rother C, Roth S, Blake A (2010) Fusion moves for markov random field optimization. *IEEE Trans Pattern Anal Mach Intell* 32(8):1392–1405
 50. Murphy KP, Weiss Y, Jordan MI (1999) Loopy belief propagation for approximate inference: an empirical study. In: *Proceedings of the Fifteenth International conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc, pp 467–475
 51. Murray I, Ghahramani Z (2004) Bayesian learning in undirected graphical models: approximate MCMC algorithms. In: *Proceedings of the 20th International conference on uncertainty in artificial intelligence*. AUAI Press, pp 392–399
 52. Pearl J (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Francisco, CA
 53. Johnson D, Sinanovic S (2001) Symmetrizing the kullback-leibler distance. <http://www-dsp.rice.edu/~dhj/resistor.pdf>
 54. Barla A, Odone F, Verri A (2003) Histogram intersection kernel for image classification. In: *Proceedings of international conference on image processing*, vol 3, pp 513–516
 55. Zobel J, Moffat A, Ramamohanarao K (1998) Inverted files versus signature files for text indexing. *ACM Trans Database Syst* 23(4):453–490
 56. van Hateren JH, van der Schaaf A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R Soc Lond B Biol Sci* 265(1394):359–366
 57. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: *Advances in Neural Information Processing Systems (NIPS)*, pp 487–495
 58. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2014) Imagenet large scale visual recognition challenge. *Int J Comput Vis*, pp 1–42
 59. Kohavi R, Provost F (1998) Glossary of terms. *Mach Learn* 30(2–3):271–274
 60. Chum O, Philbin J, Zisserman A (2008) Near duplicate image detection: min-Hash and tf-idf weighting. In: *Proceedings of British machine vision conference*, vol 810, pp 812–815