Int. J. Appl. Math. Comput. Sci., 2017, Vol. 27, No. 4, 839–852
DOI: 10.1515/amcs-2017-0059

A HIERARCHICAL INFERENTIAL METHOD FOR INDOOR SCENE CLASSIFICATION

JINGZHE JIANG [a], PENG LIU [a], ZHIPENG YE [a], WEI ZHAO [a,*], XIANGLONG TANG [a]

[a] School of Computer Science and Technology
Harbin Institute of Technology, No. 92 West Dazhi Street, Harbin, China
e-mail: zhaowei@hit.edu.cn

Indoor scene classification forms a basis for scene interaction for service robots. The task is challenging because the layout and decoration of a scene vary considerably. Previous studies on knowledge-based methods commonly ignore the importance of visual attributes when constructing the knowledge base. These shortcomings restrict the performance of classification. The structure of a semantic hierarchy was proposed to describe similarities of different parts of scenes in a fine-grained way. Besides the commonly used semantic features, visual attributes were also introduced to construct the knowledge base. Inspired by the processes of human cognition and the characteristics of indoor scenes, we proposed an inferential framework based on the Markov logic network. The framework is evaluated on a popular indoor scene dataset, and the experimental results demonstrate its effectiveness.

Keywords: indoor scene classification, semantic hierarchical structure, rule-based inference, Markov logic network.

1. Introduction

With the rapid development of imaging techniques, the amount of visual information has increased significantly, providing richer data sources for image tasks such as image annotation, recognition, retrieval, and classification (Carneiro *et al.*, 2007; Penatti *et al.*, 2014; Feng *et al.*, 2017; 2016; Feng and Zhou, 2016). Scene classification is a precondition to achieve higher level scene interaction for robots. Consequently, developing efficient tools for automatic scene analysis has drawn considerable attention.

Scene classification is one of the primary goals in computer vision, involving many sub-tasks, such as object detection and recognition. These sub-tasks have been studied intensely over the past few decades, and there is still ample room for improvement (Mottaghi *et al.*, 2013). In general, scene classification refers to the process of learning to answer a "what" question from a given sample, where the answer is naturally determined by what objects a scene contains. Classifying indoor scenes is challenging, and there are no universal models for describing such scenes (Xie *et al.*, 2014b; Khan *et al.*, 2014; Chaojie *et al.*, 2013). This is because the layout and decoration of indoor scenes vary considerably, and the classification performance is easily affected by environmental factors. As a result, indoor scenes are confusing and even difficult for a human to classify.

According to previous research (Ye *et al.*, 2015), algorithms for scene classification can be roughly divided into two types: bio-inspired and feature learning methods. Researchers have investigated and applied existing bio-inspired models, such as the human visual system, to computer-vision applications to further improve performance, and this have been proven effective (Escobar and Kornprobst, 2012; Tang and Qiao, 2014; Huang *et al.*, 2011). This strategy is popular for visual tasks, including field-of-action recognition, image processing, and scene classification (Escobar and Kornprobst, 2012; Delaigle *et al.*, 2002; Alleysson *et al.*, 2005; Siagian and Itti, 2007; Xie *et al.*, 2014a). Meanwhile, scene classifying methods based on visual features can be further divided into three strategies. The first one is based on low-level features for classification, such as color, texture, and shape (Banerji *et al.*, 2013). This strategy is effective, provided that there are only a low number of categories. Also, it is easily affected by external factors such as illumination. The second strategy is devoted to the development of high-level features from a global perspective. This is done

*Corresponding author

by treating the image as a collection of image blobs, and by introducing more descriptive features for precise scene classification (Lazebnik *et al.*, 2006; jia Li *et al.*, 2010; Khan *et al.*, 2014; Yu *et al.*, 2014; 2013; 2012b). This strategy explores the group of features and corresponding metrics to improve the classification performance, and is suitable for a larger number of learning samples. The third one is to introduce semantic features to address the problem of a semantic gap (Tang *et al*., 2012; Zhang *et al*., 2013). Last but not least, rule-and graph-based systems can also be used to solve the problem of classification (Ribeiro *et al*., 2009; Yu *et al*., 2012a; Welter *et al*., 2011; Chaves *et al*., 2012).

In this paper, we target the problem of scene classification based on rule inference. Previous research commonly ignores the importance of visual attributes, restricting the performance and flexibility. Inspired by the cognitive principle of the human being, the hierarchical structure of scenes and rule-based inference for determining the category of a scene are investigated and a semantic rule inference (SRI) framework is proposed in this paper. The knowledge base is constructed by both semantic hierarchies extracted by a discriminative part-based model (Felzenszwalb *et al.*, 2010b) and visual attributes (Farhadi *et al*., 2009) extracted from the image. A Markov logic network (MLN) (Richardson and Domingos, 2006) is used to infer the general category of the scene. The MLN deals with conflict rules generated from similar scenes by learning weights. The proposed framework is suitable for other related applications such as image retrieval and understanding.

The remainder of the paper is organized as follows. Related work is introduced in Section 2. Then the SRI indoor scene-classification framework is propose in Section 3. Experimental results are provided in Section 4. Finally, ongoing and future work are summarized in Section 5.

2. Related work

2.1. Object detection. Object detection is a basic and active research topic in computer vision. The research field includes object detection in video sequences and static images. Here we mainly discuss the latter one. The main problem of object detection in static images is that the objects in the image vary with the external factors such as illumination and the viewing angle. Typically speaking, previous research was based on several types of methods, including local image information (Ren and Ramanan, 2013; Lazebnik *et al*., 2006; Shotton *et al*., 2005; Nguyen *et al*., 2013; Teo *et al*., 2015), sparse representation (Rigamonti *et al.*, 2011; 2013), neural networks (Liu *et al*., 2016; Kong *et al*., 2016; Bell *et al*., 2016; He *et al*., 2016) and the SVM (Felzenszwalb *et al.*, 2010b; 2008; 2010a; Sadovnik and Chen, 2011; Felzenszwalb and McAllester, 2007; Girshick *et al.*, 2011; Hosang *et al.*, 2016).

Shotton *et al*. (2005) proposed a partially supervised learning method using local contour-based features. Lazebnik *et al*. (2006) presented a method based on approximate global geometric correspondence by partitioning the image into increasingly fine sub-regions to compute corresponding histograms of local features. Rigamonti *et al*. (2011) evaluated the impact of sparse representation on the performance of image classification, and reduced the computational cost of sparse filters by computing linear combinations of a small number of separable filters (Rigamonti *et al*., 2013).

The discriminatives part-based model (DPM) proposed by Felzenszwalb *et al.* (2010b) is a famous object detection approach that models unknown part positions as latent variables in an SVM framework. The model contains three parts: HoG features, the part model and the latent SVM. Significant object detection performance has been achieved on the PASCAL VOC dataset. Roughly speaking, the model can be considered the improvement over the original HoG by calculating and combining object templates of different scales. Although DPM can solve the problem of pose change to some extent, the computation cost is relatively high. Thus Felzenszwalb *et al.* (2010a) proposed a method of building cascade classifiers for the DPM model to significantly improve its detection speed.

2.2. Markov logic network. A Markov logic network (MLN) is employed for knowledge representation, which is widely used in statistical relational learning that combines the compact expressiveness of first-order logic with the flexibility of probability (Richardson and Domingos, 2006). In MLNs, the Markov random field (MRF) is applied to model logic rules by considering the literals of logic rules binary nodes of the MRF. Entities in a relational domain are represented by predicates and their relationships are represented in first-order logic. For a given rule $f_{e1} \wedge f_{e2} \wedge \ldots \wedge f_{en} \rightarrow f_h$, $f_{e1} \wedge f_{e2} \wedge \ldots \wedge f_{en}$ is called a precondition that can be seen as evidence, and f_h is called a post-condition. The set of logic rules is called a formula. Assigning a value to each logic rule is called a grounding. The learning algorithms are often expressed as stochastic sampling techniques such as Gibbs sampling, Markov chain Monte Carlo, and contrastive divergence. During the learning process, a ground MRF is initialized by the weighted logic formulae. The weights indicate the likelihood of the formulae being true. The knowledge base is defined by the combination of the formulae and the corresponding weights.

The effectiveness of the MLN in the realm of data mining have been proven (Singla and Domingos, 2006; Neville and Jensen, 2007). Recently, it has been introduced in computer vision tasks to construct

high-level knowledge to achieve high-level scene analysis (Faria *et al*., 2014; Zhu *et al*., 2014; Kembhavi *et al*., 2010; Xu and Petrou, 2010; Xu *et al*., 2011; Liu and von Wichert, 2013). Zhu *et al*. (2014) proposed a knowledge base representation for reasoning object affordance by harvesting diverse information. The method could achieve outstanding results for predicting the object affordances. Xu and Petrou (2010) proposed a Markov logic network-based logic-rule learning approach for scene interpretation to learn soft-constraint logic rules and label the components of a scene. Liu and von Wichert (2013) combined data driven Markov chain Monte Carlo sampling and inference using rule-based context knowledge, and proposed an abstract model of the perceived environment.

3. Semantic rule inference (SRI) classification framework

The motivation behind our work is to build a knowledge base with both semantic and visual attributes to further improve the performance of indoor scene classification. In this section, we first introduce the hierarchical structure of indoor scenes described by both semantic and visual attributes in Sections 3.1 and 3.2, then present our learning method to construct a knowledge base combining hierarchical semantics and visual attributes implemented by the MLN for classification in Section 3.3. After that the framework is summarized in Section 3.4.

3.1. Hierarchical structure detection. Seeing is not the same as understanding. Obtaining an image is just one small step in the process of acquiring the information associated with it, yet much work remains to be done. Regarding the similarity of an indoor scene, it is often categorized globally by analyzing the entire image. However, this approach is sometimes inappropriate, because only part of the image is similar. This is often ignored in previous research into indoor scene classification. It is common for both similar and dissimilar structures to exist in different indoor scenes, owing to the fact that the same kinds of sub-scenes are common among images. Indeed, similarity is double-edged. On the one hand, it is important for training classifiers. On the other, it risks confusing the classifier and resulting in misclassifications. One reason why the problem of indoor scene classification is challenging is that both similarity and dissimilarity exist among different scenes. That is why the hierarchical structure of indoor scenes should be investigated to deal with this problem. Some sample images are provided in Fig. 1. From the figure we can see that although the ceilings of both a greenhouse and an airport are similar, there exist significant dissimilarities between other hierarchies, providing enough information to tell them apart. This is the basic motivation for the proposed method.

During the past few years, hierarchical methods have been proposed to deal with the characteristics of indoor scenes. The methods can be roughly divided into two types: building hierarchical semantics (Marszałek and Schmid, 2007; Li-Jia *et al.*, 2010; Deng *et al.*, 2011; Bannour and Hudelot, 2012b; 2012a), and developing hierarchical models (Fei-Fei and Perona, 2005; Gupta *et al*., 2009; Porway *et al*., 2010). Building hierarchical semantics is helpful for improving the performance of image classification, making it easier to deal with a large-scale dataset. Developing hierarchical models is another way of describing the classification process in detail. For hierarchical semantics, the hierarchy of a scene is seldom explored, limiting the ability to further improve the accuracy. Methods based on hierarchical models devote to model the whole scene by extracting features from different areas. These two kinds of methods ignore the importance of the visual attributes in the learning process, which is also important in classification (Zhu *et al*., 2014). Moreover, these methods are not able to achieve the inference process like a human does. Thus, in this paper, we model the whole scene with both semantics of different areas and corresponding visual attributes. Three hierarchies were defined for an indoor scene according to the context of the image: the upper, middle, and lower hierarchies, representing the structures such as ceiling, wall and floor that can be obviously distinguished. Each hierarchy chiefly contains a single dominant semantics that constitutes the rules for training the classifier for indoor scenes. The hierarchical structure of an image is constructed upon detected objects since an indoor scene can be easily enumerated, and often composed of several common objects of different positions. Thus we design a relatively fixed hierarchical structure composed of up to three layers, and the number of hierarchies of an image flexibly depends on the results of object detection. Therefore, for each image the number of hierarchies is determined by the detection results. The scene is inferred by combining the hierarchical semantics of different areas and the visual attributes. The motivation behind this work is to focus the classifier on objects located in different parts of the image to improve the quality of the visual words. Unlike traditional methods, the proposed framework is able to reduce the interference between the corresponding hierarchies caused by similarity, because two images can be distinguished insofar as their hierarchy is different. Samples of the hierarchical structure of indoor scenes with different numbers of hierarchies are shown in Fig. 2.

3.2. Evidence of the knowledge base. A knowledge base (KB) refers to a repository of entities and rules that can be used for problem solving. The KB can

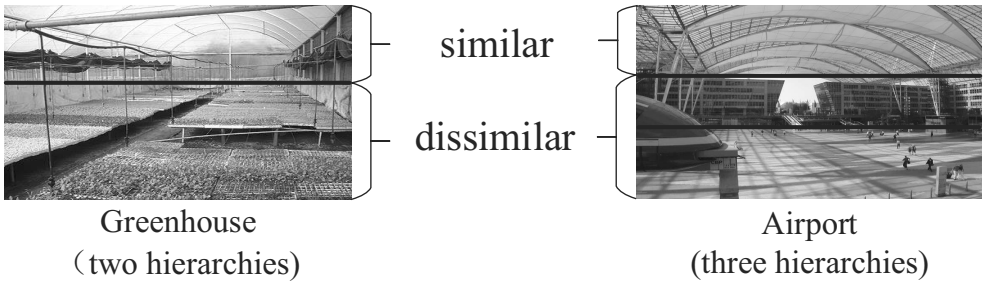


Fig. 1. Sample scenes with similar and dissimilar hierarchies. We can see that due to the context of indoor scenes, despite the number of hierarchies between different images, similar areas are on the same hierarchies. Thus, although there exist similarities between different images, it is still possible to distinguish them by the areas that are not similar. Hence the similarities between different scenes are related to the corresponding hierarchies.

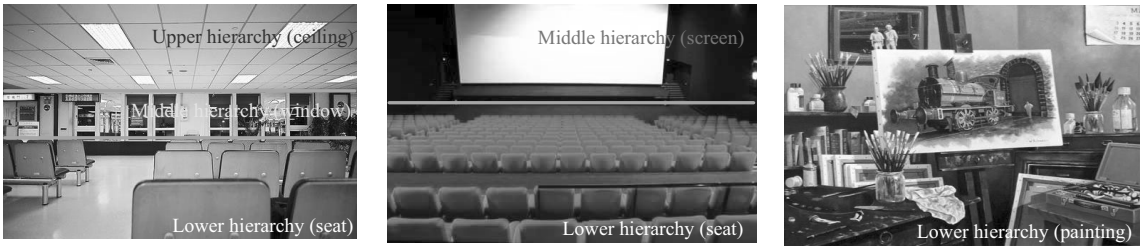


Fig. 2. Sample of indoor scenes with different hierarchical structures.

also be treated as a graph where the nodes denote the entities and the edges denote the general rules between nodes, indicating the relationships. The entities in our KB consist of object attributes. Here we choose two types of attributes to describe an object.

- **Visual attributes** are low-level knowledge acquired from visual features, describing how the object appears. We chose the visual features proposed by Farhadi *et al*. (2009) to describe the visual appearance. Here we chose the visual attributes[1] as listed by Zhu *et al*. (2014).
- **Hierarchical attributes** represent semantic understanding of a human on the object. To construct a more compact KB, hierarchical categories are abstracted to form a higher level of semantics (Lorenza Saitta, 2013) (e.g., closet and shelf belong to the wardrobe). All categorical attributes of indoor scenes are given in Fig. 3.

Here we model the strong correlations between attributes of entities by attribute–attribute relations. Positive weights indicate a positive correlation between two attributes whereas negative weights indicate that these attributes are not likely to co-occur, and zero weights make no statement about whether the relations are probable.

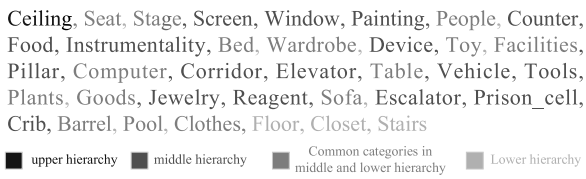


Fig. 3. Semantic category attributes used for constructing the KB.

3.3. Learning of logic rules. For the task of scene classification, the first-order logic is in the form of $L_1 \wedge \ldots \wedge L_n \Rightarrow A$ to represent the relationship between attributes and the result. Here we treat the collected semantic hierarchies and visual attributes as the preconditions and the labels of each scene as the post-condition for the logic rules. Given a set of constants $C = \{c_1, c_2, \ldots, c_{|C|}\}$ containing categories and attributes of different hierarchies, a Markov network is defined as follows (i) the nodes in the Markov

[1]Visual attributes: boxy 2D, boxy 3D, clear, cloth, feather, furniture, arm, furniture back, furniture leg, furniture seat, furry, glass, handlebars, head, horizontal cylinder, label, leather, metal, pedal, plastic pot, rein, round, saddle, screen, shiny, skin, tail, text, vegetation, vertical cylinder, wheel, wood, wool.

network are groundings of predicates, (ii) all nodes whose corresponding predicates appear in the same formula form a clique in the Markov network, (iii) each clique is associated with a feature. The weight of the feature is the weight of the formula, ω_1. Here we take one formula of Fig. 4, $isA(x, Category) \wedge hasAttribute(x, Attribute) \Rightarrow isA(u, Category)$, for example. Suppose $x \in \{chair, book\}$, $Attribute = \{furniture\_leg\}$ and $Category = \{classroom, closet\}$; then there are four cliques in the Markov network as follows:

$$\begin{aligned}
&\omega\, isA(x, Chair) \wedge hasAttribute(x, furniture_leg)\\
&\Rightarrow isA(u, classroom),\\
&\\
&\omega\, isA(x, book) \wedge hasAttribute(x, furniture_leg)\\
&\Rightarrow isA(u, classroom),\\
&\omega\, isA(x, Chair) \wedge hasAttribute(x, furniture_leg)\\
&\Rightarrow isA(u, closet),\\
&\omega\, isA(x, book) \wedge hasAttribute(x, furniture_leg)\\
&\Rightarrow isA(u, closet).
\end{aligned} \tag{1}$$

The MLN is able to answer arbitrary queries such as "What is the probability that formula $F1$ holds given that $F2$ does?". The goal of inference of the MLN is to assign a value to a variable X which contains the truth values for each predicate. Thus logic rules for scene interpretation have to be learnt first. The joint distribution over a set of variables $X = (X_1, X_2, \ldots, X_n)$, i.e., the possible worlds of the MLN, is given by

$$P(X = x) = \frac{1}{Z} \exp\Big(\sum_{i=1}^{n} \omega_i m_i(x)\Big), \tag{2}$$

where Z is the partition function for normalization, n is the number of the first-order formulae in the MLN, and m_i is the number of true groundings of the formula. For the situation listed above, $X = (isA(chair), isA(book), hasAttribute(furniture\_leg), isA(classroom), isA(closet))$. There are four formulae and one true-grounding, thus $n = 4$, and $m_1 = 1$, m_2 to m_4 equal zero. As shown in Fig. 4, given the data domain D formed by the collected evidence, the formulae F and the predicates P, the purpose is to learn the ground formulae G, and the optimal weights ω^* are learnt by maximizing the pseudo-likelihood using the L-BFGS algorithm (Richardson and Domingos, 2006).

To answer the queries, the MLN infers the probability or the most likely state of each query from the evidence. For a specified logic rule, the probability that the post-condition f_h generated from the testing sample x is true can be queried given the already known MLN M and preconditions as the evidence, expressed by $f_e = \{f_{e1}, \ldots, f_{en}\}$, and can be calculated as

$$\begin{aligned}
P(f_h|f_e, M\}) &= \frac{P(f_h \wedge f_e|M)}{P(f_e|M)}\\
&= \frac{\sum\limits_{x\in(\chi_{f_h}\cap\chi_{f_e})} P(X = x|M)}{\sum\limits_{x\in\chi_{f_e}} P(X = x|M)},
\end{aligned} \tag{3}$$

where χ_{f_h}/χ_{f_e} is the set of worlds in which f_h/f_e holds. For the situation mentioned above, f_{ei} corresponds to the formulae in Eqn. (1). Thus, for a given query formula f_h, the MLN will evaluate the similarities between the input query f_h and the generated rules f_{ei} by Eqn. (2) and determine the corresponding category.

3.4. Proposed framework. Given the collected evidence, we construct the knowledge base by learning the relations. In this paper, we choose the MLN as the inference engine. This is because according to the characteristics of the MLN the framework can be easily extended for transfer learning problems. This means that the proposed framework is able to manipulate the previously acquired knowledge to answer a new question (Bottou, 2013; Zhu *et al.*, 2014). Also, compared with the naive strategy, the MLN is more effective in dealing with conflict rules generated from similar scenes. The structure of the knowledge base of the proposed SRI framework implemented by the MLN is summarized in Fig. 4. The knowledge base is constructed by domain D, predicates P, random variables X, formulae F and ground formulae G. Both X and F are pre-defined for the task of indoor scene classification. Here, X and F are defined according to the number of hierarchies as shown in Fig. 2. Semantic attributes are represented by the predicate isA, and visual attributes are represented by the predicate $hasVisualAttribute$. Random variables are generated by an assignment of P with evidence collected from indoor scenes. G stands for the true formula after the assignment of F with X. The semantic and visual attributes are merged by predicates of the formulae. From a top-down view, the MLN serves as a formalism that provides particular probabilistic semantics. From a bottom-up view, the MLN is a particular way of compactly representing generalized features.

4. Experimental results

In this section, the overall performance of the proposed SRI framework was evaluated on the MIT's indoor scene-recognition database. Tests were divided into two parts containing vertical and horizontal comparisons to demonstrate the effectiveness of our work. First, we focused on the performance of SRI with different module settings. Then, SRI was compared with other relevant

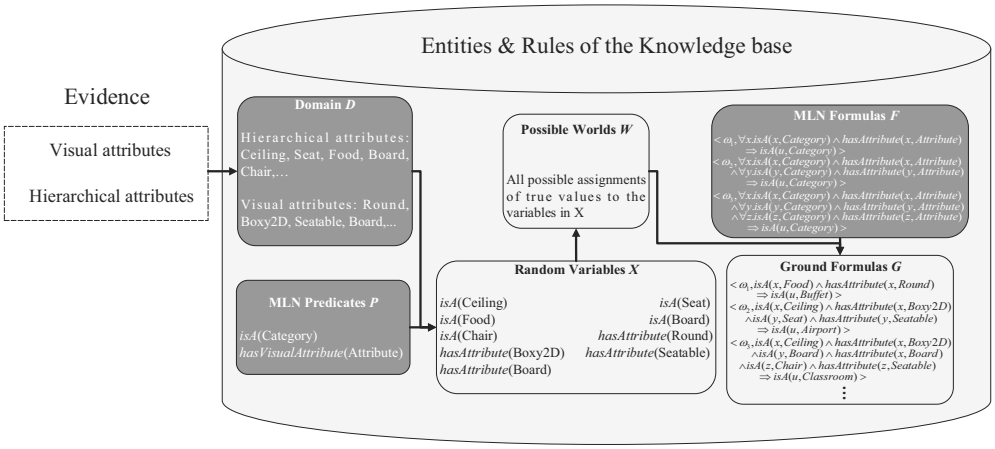


Fig. 4. Overview of the knowledge base learning process.

methods for a horizontal comparison. The corresponding results are provided in the following subsections.

4.1. Experimental settings.

4.1.1. Dataset. The MIT's indoor scene-recognition dataset (MIT, n.d.) contains 67 indoor categories in a total of 15,620 images loosely divided into five abstract categories: home, store, public places, leisure, and working places. The resolution of the smallest axis for all images in this dataset is larger than 200 pixels. The uniqueness of this dataset lies in the fact that, unlike outdoor scenes that can be roughly described with global scene statistics, indoor scenes tend to be much more variable in terms of the objects they contain. As such, unlike in other datasets such as Caltech-101, the distance between different categories is not significant. With the dataset from the MIT, it is sometimes confusing even for humans to distinguish between pairs of samples.

4.1.2. Detailed settings. In experiments, the MIT dataset was equally and randomly divided into training and testing sets. The knowledge base is constructed from the training dataset with ProbCog (Hall *et al*., 2009). The method implemented by Vondrick *et al*. (2013) was used to visualize the object detection results. Two-fold cross validation was used to compare our proposal with other methods, and the mean precision was reported. Mean-average precision (MAP) was used as a metric to evaluate the performance of the approaches.

We conduct several experiments, including vertical comparison, horizontal comparison, partial observation and diverse query, to prove the effectiveness and robustness of the proposed framework. In vertical comparison, different settings of the proposed SRI framework are evaluated for the overall comparison. In horizontal comparison, SRI is compared with other relative classification methods. In partial observation and diverse query, SRI is trained and tested by partial evidence and data. For the experiments of classification, we compare the performance of seven methods, including both single-feature and multi-feature ones:

(a) The traditional bag of visual words classification method (Csurka *et al*., 2004). BoVW is one of the most widely used methods for object classification based on vector quantization. SIFT was chosen as the feature descriptor. The size of the code book is 1000. The SVM is utilized as the classifier.

(b) The method proposed by Quattoni and Torralba (2009). SIFT and GIST were chosen as the feature descriptors.

(c) DPM (Felzenszwalb *et al.*, 2010b). The DPM is a successful object detector that directly improves the traditional HoG. The object detector implemented by the authors is utilized for the object detection task.

(d) Object bank (Li *et al*., 2014). The object bank method offers high-level encoding of an object's appearance and spatial location information for image recognition. HoG, texture, location and geometry were chosen to train its SVM object detector (Felzenszwalb *et al.*, 2010b) and the Hoiem classifier (Hoiem *et al*., 2005).

(e) Multi-scale BoVW (Zhou *et al*., 2013). This framework introduces multi-scale information to the original BoVW to improve the performance of classification. The size of the codebook is 200. SIFT is chosen as the feature descriptor.

(f) ISPRs (Lin *et al*., 2014). The ISPR classification method jointly learns spatial-pooling regions with discriminative part appearance in a unified framework for

scene classification. Settings described by the authors are used for comparison.

(g) CNN-SVM (Sharif Razavian *et al*., 2014). We use the implementation of a CNN described by the authors, and data augmentation is done by providing 16 representations for each sample (original image, 5 crops, 2 rotations and their mirrors).

4.2. Experimental results.

4.2.1. Vertical comparison results. In the vertical experiments, we first show the statistical distribution of rules of the dataset, and then evaluate the performance of the proposed framework with different modular settings. Figure 5 shows the results of evidence collection. Both the hierarchical and visual attributes are visualized in the figure. The process of automatical hierarchy detection is achieved by the DPM (Felzenszwalb *et al.*, 2010b). The hierarchical structures are automatically constructed according to their spatial context in the image. Visual attributes are detected with the method proposed by Farhadi *et al*. (2009). Evaluation on different settings of the proposed method is given in Table 1.

For a single hierarchical structure, the model degraded to a common classifying method. The results show a substantial improvement in the performance with the proposed hierarchical structure, and using both hierarchical and visual attributes to construct the knowledge base is beneficial for improving the performance classifying indoor scenes. This is due to the fact that indoor scenes are typically complex. It is common to find multiple objects in the same scene, and classifying such samples merely with global features introduces noise that affects the performance. The introduction of a hierarchical structure divides and annotates the indoor scenes into several hierarchies, and combines the categories from each hierarchy to obtain a universal category, considerably improving the quality of the classification. Also, visual attributes are an important factor for us to interact with the real world. Top positive and negative weighted relations are listed in Fig. 6. Detailed classification results on concluded and detailed classes (Quattoni and Torralba, 2009) are given in Fig. 7 and Table 2.

Table 1. Performance of the proposed framework with different settings.

Strategy	MAP
Single hierarchy – Visual attribute	40.6
Single hierarchy + Visual attribute	43.7
Multiple hierarchy – Visual attribute	49.1
Multiple hierarchy + Visual attribute	53.4

4.2.2. Horizontal comparison results. In this subsection, we compare the proposed framework with other methods based on visual and semantic attributes, including the classic BoVW (Csurka *et al*., 2004), the prototype-based model (Quattoni and Torralba, 2009), the DPM (Felzenszwalb *et al.*, 2010b), the object bank (Li *et al*., 2014), the multi-scale BoVW (Zhou *et al*., 2013), and the important spatial pooling regions (ISPRs) (Lin *et al*., 2014).

Table 3 shows the results of the proposed SRI compared to other scene-classification methods. Performance is evaluated by mean average precision (MAP). BoVW was used as baseline. The proposed SRI framework detected the objects in each hierarchy, exhibiting their spatial relationship according to the semantic hierarchical structure. Furthermore, SRI introduced visual attributes to construct the knowledge base. Thus, the proposed SRI framework is effective for indoor scene classification, and it consistently outperformed other methods.

4.2.3. Robustness and diverse query. The ability of inferring from partial observation is an important feature of humans. The trend of knowledge-based querying methods is to achieve performance comparable with that of a human. Besides, robustness and diversity are also important factors to reflect the quality of the KB. In this section we will show the robustness and diversity of the proposed model. Inspired by Zhu *et al*. (2014), we designed the tests using partial evidence and different semantic granularity. First we demonstrate the robustness of SRI by testing the performance of our model in classification, given a randomly selected portion of evidence for the learning process. Then we evaluate the performance of SRI with rules generated from different indoor scenes of the same abstract category, i.e., we test the performance of the generated KB from a coarse-grained level, which differs from Section 4.2.2 in that the performance of the generated KB is tested with exactly the same category we used for learning. The purpose of this test is that it is unable to guess what query it may receive from a user from the perspective of the KB in practice, and dealing with this kind of query is also important to evaluate the quality of the generated KB.

For the experiments of partial observation of the CNN, the network is trained with a dataset of which some inherited categories are randomly removed, and the network is tested by the whole categories. The performance is measured from the viewpoint of all the five abstract categories. For example, for the categories "bedroom" and "kitchen" under the abstract category "home", we remove "bedroom" and use "kitchen" to train the whole network. Then we will test the network with both bedroom and "kitchen". If "bedroom" is categorized as any other inherited categories under "home", it

Table 2. Detailed experimental results on the MIT indoor dataset.

Abstract categories	Categories	Number of samples	Accuracy
Store	Bakery	405	0.493
	Grocery store	213	0.614
	Clothing store	106	0.503
	Deli	258	0.491
	Laundromat	276	0.546
	Jewellery shop	157	0.341
	Bookstore	380	0.464
	Video store	110	0.527
	Florist	103	0.608
	Shoe shop	116	0.382
	Mall	176	0.315
	Toystore	347	0.446
Home	Bedroom	662	0.488
	Nursery	144	0.593
	Closet	135	0.652
	Pantry	384	0.566
	Children room	112	0.451
	Lobby	101	0.486
	Dining room	274	0.532
	Corridor	346	0.674
	Livingroom	706	0.531
	Bathroom	197	0.596
	Kitchen	734	0.562
	Stairscase	155	0.587
	Winecellar	269	0.546
	garage	103	0.585
Public space	Prison cell	103	0.468
	Library	107	0.663
	Cloister	120	0.314
	Church	180	0.832
	Waiting room	151	0.523
	Museum	168	0.412
	Elevator	101	0.804
	Pool inside	174	0.564
	Inside bus	102	0.678
	Inside subway	457	0.564
	Subway	539	0.447
	Locker room	249	0.645
	Trainstation	153	0.458
	Airport inside	608	0.497
	Auditorium	176	0.787
	Hospital room	101	0.614
	Kinder garden	127	0.399
	Restaurant kitchen	107	0.401
Working place	Artstudio	140	0.373
	Classroom	113	0.708
	Laboratory wet	125	0.348
	Studio music	108	0.603
	Operating room	135	0.474
	Office	109	0.298
	Computer room	114	0.659
	Warehouse	506	0.439
	Green house	101	0.728
	Dental office	131	0.651
	Tv studio	166	0.548
	Meeting room	233	0.458
Leisure	Buffet	111	0.719
	Fastfood	116	0.503
	Concert hall	103	0.646
	Restaurant	513	0.388
	Bar	604	0.501
	Movie theater	175	0.457
	Gameroom	127	0.539
	Casino	515	0.508
	Bowling	213	0.723
	Gym	231	0.548
	Hairsalon	239	0.427

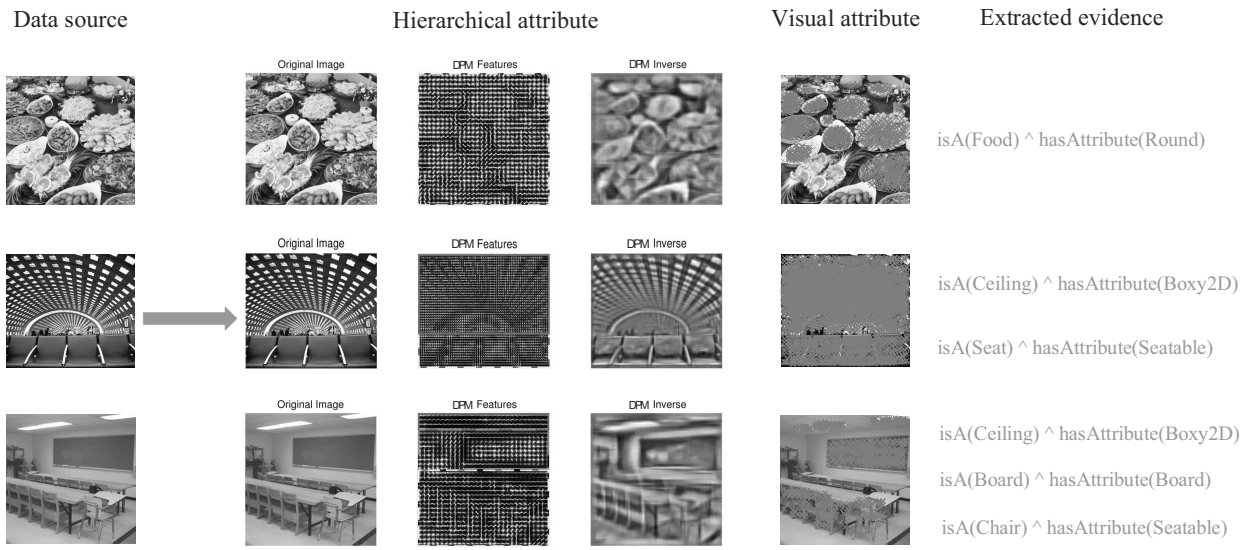


Fig. 5. Sample of a hierarchical structure and detected visual attributes of indoor scenes. The features of the DPM are visualized in the middle column. Detected visual attributes are marked in the image. Extracted evidence of scenes with different hierarchical structures is given in the last column. (The image is best viewed in color.)

Table 3. Comparative results with relevant methods on the MIT indoor dataset.

Methods	MAP
BoVW (Csurka *et al*., 2004)	12.7
ROI+GIST (Quattoni and Torralba, 2009)	26.1
DPM (Felzenszwalb *et al.*, 2010b)	30.4
Object bank (Li *et al*., 2014)	37.6
Multi-scale BoVW (Zhou *et al*., 2013)	46.5
ISPRs (Lin *et al*., 2014)	50.1
CNN-SVM (Sharif Razavian *et al*., 2014)	58.4
SRI	53.4

is considered correct, otherwise if it is categorized under other abstract categories, such as "leisure", the classification is considered incorrect.

The results of robustness and diverse query are respectively given in Figs. 8 and 9. The abstract categories are given by Quattoni and Torralba (2009). From the figures we can see that compared with the method based on classification (Zhou *et al*., 2013), the knowledge-based SRI model is more robust to the variation in evidence. Meanwhile, the KB can answer diverse queries although it is not specially trained for this kind of test. This is because combining visual attributes and semantic categories improves the quality and robustness of the knowledge base.

Although the CNN-based method achieved the best performance in horizontal comparison, when the evidence of unobserved raises, the performance of the CNN decreases significantly compared with that of SRI. This is because SRI combines both semantic and visual attributes in constructing the knowledge base, which is more descriptive than low-level visual features used by the CNN and other deep learning methods. Moreover, there exist several inevitable problems for CNN-based methods:

(i) The parameters of the CNN need to be carefully tuned to get satisfactory performance. It is heavily dependent on the experience of the experts, and is much easier for the proposed SRI framework.

(ii) The network needs a large amount of data; this will lead to the result that training the structure of the CNN is expensive in both purchasing special hardware and consuming huge computational resources. According to Sharif Razavian *et al.* (2014), the training of CNN consumes several weeks on Tesla K40 GPUs, while the proposed SRI framework took several days to be trained on a quad-core CPU;

(iii) The CNN-based method is not necessarily able to receive original images, which means the images should be pre-processed before the training process (Sharif Razavian *et al*., 2014; Dixit *et al*., 2015). This will introduce additional processing cost for large scale datasets. There is no such extra costs for the proposed SRI framework. Therefore, the proposed and other inference methods based on knowledge inference remains valuable in the era of deep learning.

2.876 $isA(Vehicle) \wedge hasAttribute(Mental) \Rightarrow isA(Garage)$
1.923 $isA(Ceiling) \wedge hasAttribute(Boxy2D) \wedge isA(Seat) \wedge hasAttribute(Seatable) \Rightarrow isA(Airport)$
0.951 $isA(Food) \wedge hasAttribute(Round) \Rightarrow isA(Buffet)$
0.752 $isA(Ceiling) \wedge hasAttribute(Boxy3D) \wedge isA(Window) \wedge hasAttribute(Boxy2D) \wedge isA(Bed) \wedge hasAttribute(Boxy3D) \Rightarrow isA(Bedroom)$
0.687 $isA(instrumentality) \wedge hasAttribute(Boxy3D) \Rightarrow isA(Studiomusic)$

(a)

−3.082 $isA(Chair) \wedge hasAttribute(Seatable) \Rightarrow isA(Bathroom)$
−2.875 $isA(Clothes) \wedge hasAttribute(Cloth) \Rightarrow isA(Cloister)$
−2.753 $isA(Ceiling) \wedge hasAttribute(Boxy3D) \wedge isA(Vehicle) \wedge hasAttribute(Mental) \Rightarrow isA(Kindergarden)$
−1.823 $isA(Ceiling) \wedge hasAttribute(Boxy2D) \wedge isA(Wardrobe) \wedge hasAttribute(Boxy3D) \Rightarrow isA(Diningroom)$
−1.742 $isA(Window) \wedge hasAttribute(Boxy2D) \wedge isA(Seat) \wedge hasAttribute(Seatable) \Rightarrow isA(Florist)$

(b)

Fig. 6. Top weighted rules. The rules can be well interpreted. For instance, the first rule in the second image means that a chair is less likely to appear in a bathroom: top positive rules (a), top negative rules (b).

5. Conclusion

In this paper, we investigated the indoor scene classification problem and proposed a novel semantic rule inference (SRI) framework. The structure of semantic hierarchies exists in indoor scenes, and the proposed hierarchical structure is able to distinguish both scenes of different categories with similar hierarchies and those of the same categories with different hierarchies. We proposed an inferential framework based on which the knowledge base is constructed with both the hierarchical structure and visual attributes. Experimental results demonstrated the effectiveness and robustness of the proposed SRI framework.

Acknowledgment

This work was supported by the National Science Foundation of China (grants no. 61171184 and 61201309).

References

Alleysson, D., Susstrunk, S. and Herault, J. (2005). Linear demosaicing inspired by the human visual system, *IEEE Transactions on Image Processing* **14**(4): 439–449.

Banerji, S., Sinha, A. and Liu, C. (2013). New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, *Neurocomputing* **117**(0): 173–185.

Bannour, H. and Hudelot, C. (2012a). *Building Semantic Hierarchies Faithful to Image Semantics*, Lecture Notes in Computer Science, Vol. 7131, Springer, Berlin/Heidelberg, pp. 4–15.

Bannour, H. and Hudelot, C. (2012b). Hierarchical image annotation using semantic hierarchies, *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA*, pp. 2431–2434.

Bell, S., Lawrence Zitnick, C., Bala, K. and Girshick, R. (2016). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 2874–2883.

Bottou, L. (2013). From machine learning to machine reasoning, *Machine Learning* **94**(2): 133–149.

Carneiro, G., Chan, A.B., Moreno, P.J. and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(3): 394–410.

Chaojie, W., Jun, Y. and Dapeng, T. (2013). High-level attributes modeling for indoor scenes classification, *Neurocomputing* **121**: 337–343.

Chaves, R., Ramrez, J., Grriz, J. and Illn, I. (2012). Functional brain image classification using association rules defined over discriminant regions, *Pattern Recognition Letters* **33**(12): 1666–1672.

Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C. (2004). Visual categorization with bags of keypoints, *Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic*, Vol. 1, pp. 1–2.

Delaigle, J., Devleeschouwer, C., Macq, B. and Langendijk, L. (2002). Human visual system features enabling watermarking, *2002 IEEE International Conference on Multimedia and Expo. ICME '02, Los Angeles, CA, USA*, Vol. 2, pp. 489–492.

Deng, J., Berg, A.C. and Fei-Fei, L. (2011). Hierarchical semantic indexing for large scale image retrieval, *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA*, pp. 785–792.

Dixit, M., Chen, S., Gao, D., Rasiwasia, N. and Vasconcelos, N. (2015). Scene classification with semantic fisher vectors,

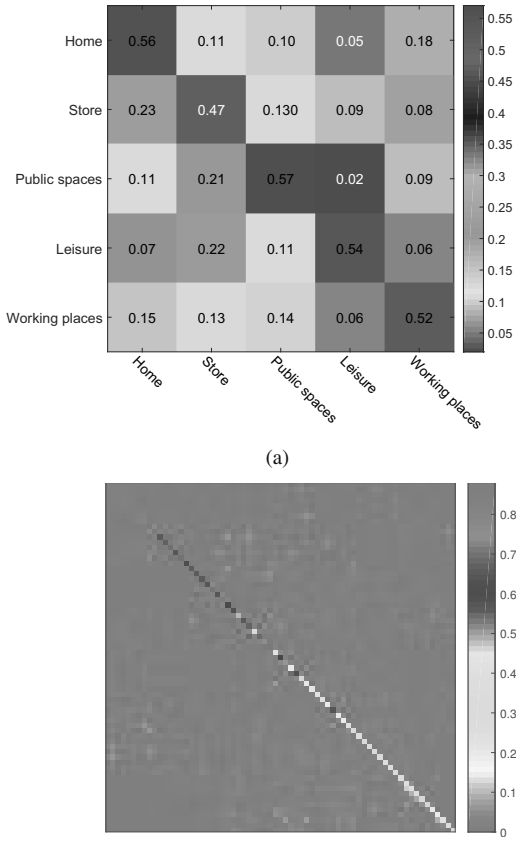


Fig. 7. Results of vertical tests of SRI: confusion table of SRI on 5 abstract categories (a), confusion table of SRI on all 67 categories (the five categories starting from the top left corner are greenhouse, classroom, church_inside, cloister, buffet; the five categories starting from the bottom right corner are hairsalon, subway, museum, children_room, prison_cell) (b).

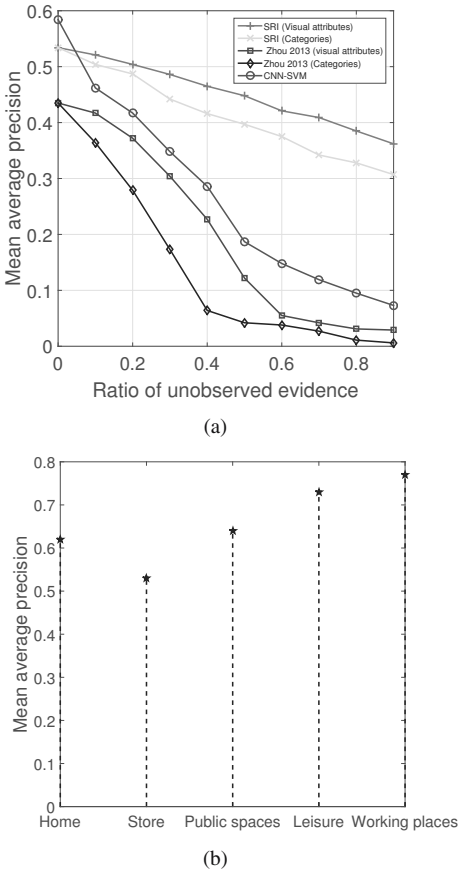


Fig. 8. Results of robustness and diverse query of SRI: performance variations of partial evidence for learning—the proposed SRI framework outperforms other methods in the situation of partial observation (a), overall results of diverse querying on the abstract categories (higher is better)—the proposed SRI framework based on the knowledge base is robust to the situation of diverse querying (b).

2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 2974–2983.

Escobar, M.-J. and Kornprobst, P. (2012). Action recognition via bio-inspired features: The richness of center–surround interaction, *Computer Vision and Image Understanding* **116**(5): 593–605.

Farhadi, A., Endres, I., Hoiem, D. and Forsyth, D. (2009). Describing objects by their attributes, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, FL, USA*, pp. 1778–1785.

Faria, D.R., Trindade, P., Lobo, J. and Dias, J. (2014). Knowledge-based reasoning from human grasp demonstrations for robot grasp synthesis, *Robotics and Autonomous Systems* **62**(6): 794–817.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA*, Vol. 2, pp. 524–531.

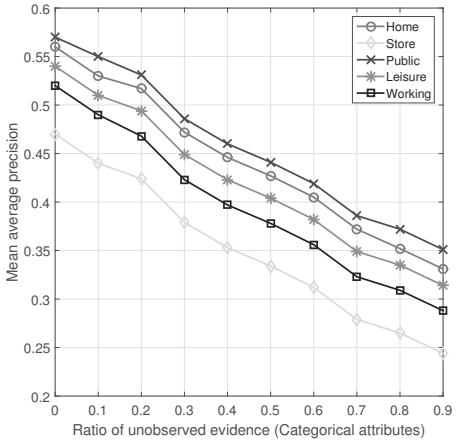
Felzenszwalb, P.F. and McAllester, D. (2007). The generalized a* architecture, *Journal of Artificial Intelligence Research* pp. 153–190.

Felzenszwalb, P., Girshick, R. and McAllester, D. (2010a). Cascade object detection with deformable part models, *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA,* pp. 2241–2248.

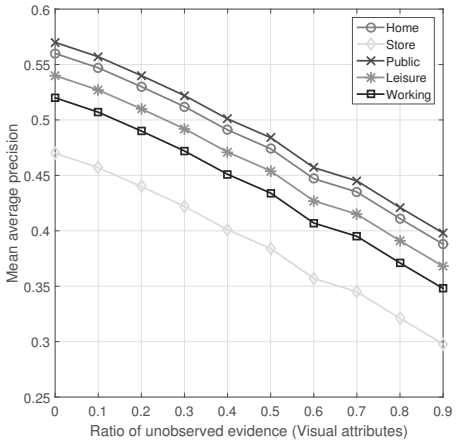
Felzenszwalb, P., Girshick, R., McAllester, D. and Ramanan, D. (2010b). Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9): 1627–1645.

Felzenszwalb, P., McAllester, D. and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, AK, USA*, pp. 1–8.

Feng, Q., Yuan, C., Pan, J.S., Yang, J.F., Chou, Y.T., Zhou, Y. and Li, W. (2017). Superimposed sparse parameter



(a)



(b)

Fig. 9. Detailed results of partial observation of SRI: details of the performance variations in partial categorical attributes (a), details of the performance variations in partial visual attributes (b).

classifiers for face recognition, *IEEE Transactions on Cybernetics* **47**(2): 378–390.

Feng, Q. and Zhou, Y. (2016). Kernel regularized data uncertainty for action recognition, *IEEE Transactions on Circuits and Systems for Video Technology* **PP**(99): 1–1.

Feng, Q., Zhou, Y. and Lan, R. (2016). Pairwise linear regression classification for image set retrieval, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 4865–4872.

Girshick, R.B., Felzenszwalb, P.F. and McAllester, D.A. (2011). Object detection with grammar models, *in* J. Shawe-Taylor *et al.* (Eds.), *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., Granada, pp. 442–450.

Gupta, P., Arrabolu, S.S., Brown, M. and Savarese, S. (2009). Video scene categorization by 3D hierarchical histogram matching, *IEEE 12th International Conference on Computer Vision, Kyoto, Japan*, pp. 1655–1662.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The Weka data mining software: An update, *ACM SIGKDD Explorations Newsletter* **11**(1): 10–18.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 770–778.

Hoiem, D., Efros, A.A. and Hebert, M. (2005). Automatic photo pop-up, *ACM SIGGRAPH 2005, Los Angeles, CA, USA*, pp. 577–584.

Hosang, J., Benenson, R., Dollár, P. and Schiele, B. (2016). What makes for effective detection proposals?, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(4): 814–830.

Huang, K., Tao, D., Yuan, Y., Li, X. and Tan, T. (2011). Biologically inspired features for scene classification in video surveillance, *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics* **41**(1): 307–313.

jia Li, L., Su, H., Fei-fei, L. and Xing, E.P. (2010). Object bank: A high-level image representation for scene classification and semantic feature sparsification, *in* J. Lafferty *et al.* (Eds.), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., Cambridge, pp. 1378–1386.

Kembhavi, A., Yeh, T. and Davis, L.S. (2010). Why did the person cross the road (there)? Scene understanding using probabilistic logic models and common sense reasoning, *in* K. Daniilidis *et al.* (Eds.), *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Part II*, Springer, Berlin/Heidelberg, pp. 693–706.

Khan, S., Bennamoun, M., Sohel, F. and Togneri, R. (2014). *Geometry Driven Semantic Labeling of Indoor Scenes*, Lecture Notes in Computer Science, Vol. 8689, Springer International Publishing, Berlin, pp. 679–694.

Kong, T., Yao, A., Chen, Y. and Sun, F. (2016). Hypernet: Towards accurate region proposal generation and joint object detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 845–853.

Lazebnik, S., Schmid, C. and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA,* Vol. 2, pp. 2169–2178.

Li-Jia, L., Chong, W., Yongwhan, L., Blei, D.M. and Li, F.-F. (2010). Building and using a semantivisual image hierarchy, *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA*, pp. 3336–3343.

Li, L.-J., Su, H., Lim, Y. and Fei-Fei, L. (2014). Object bank: An object-level image representation for high-level visual recognition, *International Journal of Computer Vision* **107**(1): 20–39.

Lin, D., Lu, C., Liao, R. and Jia, J. (2014). Learning important spatial pooling regions for scene classification, *2014 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, pp. 3726–3733.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A.C. (2016). *SSD: Single Shot MultiBox Detector*, Springer International Publishing, Cham, pp. 21–37.

Liu, Z. and von Wichert, G. (2013). Applying rule-based context knowledge to build abstract semantic maps of indoor environments, *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan*, pp. 5141–5147.

Lorenza Saitta, J.-D.Z. (2013). *Abstraction in Artificial Intelligence and Complex Systems*, Springer, New York, NY.

Marszalek, M. and Schmid, C. (2007). Semantic hierarchies for visual object recognition, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, MN, USA*, pp. 1–7.

MIT (n.d.) Indoor scene recognition. Dataset, `http://web.mit.edu/torralba/www/indoor.html`.

Mottaghi, R., Fidler, S., Yao, J., Urtasun, R. and Parikh, D. (2013). Analyzing semantic segmentation using hybrid human-machine CRFS, *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA*, pp. 3143–3150.

Neville, J. and Jensen, D. (2007). Relational dependency networks, *Journal of Machine Learning Research* **8**: 653–692.

Nguyen, D.T., Ogunbona, P.O. and Li, W. (2013). A novel shape-based non-redundant local binary pattern descriptor for object detection, *Pattern Recognition* **46**(5): 1485–1500.

Penatti, O.A., Silva, F.B., Valle, E., Gouet-Brunet, V. and Torres, R.d.S. (2014). Visual word spatial arrangement for image retrieval and classification, *Pattern Recognition* **47**(2): 705–720.

Porway, J., Wang, Q. and Zhu, S.C. (2010). A hierarchical and contextual model for aerial image parsing, *International Journal of Computer Vision* **88**(2): 254–283.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, FL, USA*, pp. 413–420.

Ren, X. and Ramanan, D. (2013). Histograms of sparse codes for object detection, *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA*, pp. 3246–3253.

Ribeiro, M.X., Bugatti, P.H., Traina Jr, C., Marques, P.M.A., Rosa, N.A. and Traina, A.J.M. (2009). Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques, *Data and Knowledge Engineering* **68**(12): 1370–1382.

Richardson, M. and Domingos, P. (2006). Markov logic networks, *Machine Learning* **62**(1): 107–136.

Rigamonti, R., Brown, M.A. and Lepetit, V. (2011). Are sparse representations really relevant for image classification?, *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA*, pp. 1545–1552.

Rigamonti, R., Sironi, A., Lepetit, V. and Fua, P. (2013). Learning separable filters, *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA*, pp. 2754–2761.

Sadovnik, A. and Chen, T. (2011). Pictorial structures for object recognition and part labeling in drawings, *18th IEEE International Conference on Image Processing (ICIP), Brussels, Belgium*, pp. 3613–3616.

Sharif Razavian, A., Azizpour, H., Sullivan, J. and Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA*, pp. 806–813.

Shotton, J., Blake, A. and Cipolla, R. (2005). Contour-based learning for object detection, *10th IEEE International Conference on Computer Vision, ICCV 2005, Beijing, China*, Vol. 1, pp. 503–510.

Siagian, C. and Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(2): 300–312.

Singla, P. and Domingos, P. (2006). Entity resolution with Markov logic, *6th International Conference on Data Mining, ICDM'06, Hong Kong, China*, pp. 572–582.

Tang, J., Zha, Z.-J., Tao, D. and Chua, T.-S. (2012). Semantic-gap-oriented active learning for multilabel image annotation, *IEEE Transactions on Image Processing* **21**(4): 2354–2360.

Tang, T. and Qiao, H. (2014). Improving invariance in visual classification with biologically inspired mechanism, *Neurocomputing* **133**(8): 328–341.

Teo, C.L., Fermller, C. and Aloimonos, Y. (2015). A Gestaltist approach to contour-based object recognition: Combining bottom-up and top-down cues, *International Journal of Robotics Research* **34**(4-5): 627–652.

Vondrick, C., Khosla, A., Malisiewicz, T. and Torralba, A. (2013). HOGgles: Visualizing object detection features, *IEEE International Conference on Computer Vision, Sydney, Australia*, pp. 1–8.

Welter, P., Riesmeier, J., Fischer, B., Grouls, C., Kuhl, C. and Deserno (né Lehmann), T.M. (2011). Bridging the integration gap between imaging and information systems: A uniform data concept for content-based image retrieval in computer-aided diagnosis, *Journal of the American Medical Informatics Association* **18**(4): 506–510.

Xie, L., Tian, Q., Wang, M. and Zhang, B. (2014a). Spatial pooling of heterogeneous features for image classification, *IEEE Transactions on Image Processing* **23**(5): 1994–2008.

Xie, L., Wang, J., Guo, B., Zhang, B. and Tian, Q. (2014b). Orientational pyramid matching for recognizing indoor scenes, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA*, pp. 3734–3741.

Xu, M. and Petrou, M. (2010). Learning logic rules for scene interpretation based on Markov logic networks, *ACCV 9th Asian Conference on Computer Vision, Xi'an, China*, pp. 341–350.

Xu, M., Petrou, M. and Lu, J. (2011). Learning logic rules for the tower of knowledge using Markov logic networks, *International Journal of Pattern Recognition and Artificial Intelligence* **25**(06): 889–907.

Ye, Z., Liu, P., Zhao, W. and Tang, X. (2015). Cognition inspired framework for indoor scene annotation, *Journal of Electronic Imaging* **24**(5): 053013.

Yu, J., Rui, Y., Tang, Y.Y. and Tao, D. (2014). High-order distance-based multiview stochastic learning in image classification, *IEEE Transactions on Cybernetics* **44**(12): 2431–2442.

Yu, J., Tao, D., Rui, Y. and Cheng, J. (2013). Pairwise constraints based multiview features fusion for scene classification, *Pattern Recognition* **46**(2): 483–496.

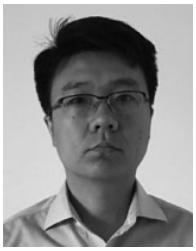
Yu, J., Tao, D. and Wang, M. (2012a). Adaptive hypergraph learning and its application in image classification, *IEEE Transactions on Image Processing* **21**(7): 3262–3272.

Yu, J., Wang, M. and Tao, D. (2012b). Semisupervised multiview distance metric learning for cartoon synthesis, *IEEE Transactions on Image Processing* **21**(11): 4636–4648.

Zhang, C., Liu, J., Tian, Q., Liang, C. and Huang, Q. (2013). Beyond visual features: A weak semantic image representation using exemplar classifiers for classification, *Neurocomputing* **120**(0): 318–324.

Zhou, L., Zhou, Z. and Hu, D. (2013). Scene classification using a multi-resolution bag-of-features model, *Pattern Recognition* **46**(1): 424–433.

Zhu, Y., Fathi, A. and Fei-Fei, L. (2014). Reasoning about object affordances in a knowledge base representation, *in* D. Fleet *et al.* (Eds.), *Computer Vision ECCV 2014*, Lecture Notes in Computer Science, Vol. 8690, Springer International Publishing, Zurich, pp. 408–424.



Jinzhe Jiang is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology. He received his BSc and MSc degrees in automation at the Nanjing University of Aeronautics and Astronautics in 2004 and 2008, respectively. His research interests cover pattern recognition and machine learning.

Peng Liu is an associate professor at the School of Computer Science and Technology, HIT. He received his doctoral degree in microelectronics and solid state electronics at the HIT in 2007. His research interests cover image processing, video processing, pattern recognition and design of VLSI circuits.

Zhipeng Ye is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology. He received his MSc degree in computer application technology at the Harbin Institute of Technology in 2013. His research interests cover image processing and machine learning.

Wei Zhao is an associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. She received her doctoral degree in computer application technology at the HIT in 2006. Her research interests cover pattern recognition, image processing, and deep space target visual analysis.

Xianglong Tang is a professor at the School of Computer Science and Technology, Harbin Institute of Technology. He received his doctoral degree in computer application technology at the HIT in 1995. His research interests cover pattern recognition, aerospace image processing, medical image processing and machine learning.

Received: 31 October 2016
Revised: 19 April 2017
Re-revised: 10 July 2017
Accepted: 16 July 2017