Journal of Electronic Imaging

JElectronicImaging.org

Cognition inspired framework for indoor scene annotation

Zhipeng Ye
Peng Liu
Wei Zhao
Xianglong Tang

SPIE.
IS&T
imaging.org

Journal of Electronic Imaging 24(5), 053013 (Sep/Oct 2015)

Cognition inspired framework for indoor scene annotation

Zhipeng Ye, Peng Liu, Wei Zhao,* and Xianglong Tang
Harbin Institute of Technology, Pattern Recognition and Intelligent System Research Center, School of Computer Science and Technology,
92 West Dazhi Street, Harbin 150001, China

Abstract. We present a simple yet effective scene annotation framework based on a combination of bag-of-visual words (BoVW), three-dimensional scene structure estimation, scene context, and cognitive theory. From a macroperspective, the proposed cognition-based hybrid motivation framework divides the annotation problem into empirical inference and real-time classification. Inspired by the inference ability of human beings, common objects of indoor scenes are defined for experience-based inference, while in the real-time classification stage, an improved BoVW-based multilayer abstract semantics labeling method is proposed by introducing abstract semantic hierarchies to narrow the semantic gap and improve the performance of object categorization. The proposed framework was evaluated on a variety of common data sets and experimental results proved its effectiveness. *© 2015 SPIE and IS&T* [DOI: 10.1117/1.JEI.24.5.053013]

Keywords: image processing and analysis; scene annotation; object classification; bag-of-visual-words; multiple-layer abstract semantics.

Paper 15259 received Apr. 6, 2015; accepted for publication Aug. 20, 2015; published online Sep. 21, 2015.

1 Introduction

Scene annotation, as a primary goal of computer vision and robotic techniques involving many subtasks, such as depth estimation, saliency detection, and object annotation, has been intensely studied during the past few decades.[1] Within this research field, Markov random field (MRF) is considered as a natural model for exploiting spatial priors.[2,3] Recently, conditional random field[4–6] has been widely utilized and has brought significant improvement in MRF. Another state-of-the-art method, max-margin Markov networks, effectively incorporates large margin mechanisms into MRF.[7–9] Although scene analysis has been studied extensively during the past decade, improvement to it remains a critical challenge,[10] largely because existing models do not have sufficient variability to describe the variable world, which restricts the application, field, and performance of existing methods.

To solve this problem, based on traditional object classification method, we present a novel framework for scene annotation by incorporating three-dimensional (3-D) scene structure estimation, scene context, and cognitive theory. Studies on scene structure estimation aim to recover spatial information of a scene and estimate the position of objects. Context in an image encapsulates rich information on how natural scenes and objects relate to one another. Using contextual information has become popular in object recognition, as it enforces coherent scene interpretation and eliminates false positives to improve the accuracy of image classification.[11–14] Most approaches can be classified into two general categories: (i) context inference based on statistical summary of the scene (scene-based context models) and (ii) context representation in terms of relationships among objects in the image (object-based context). Cognition is the brain faculty for processing information and applying knowledge in humans.[15] Existing research shows that using biometric theory assists immensely in classification tasks.[16]

In our work, we study how humans attempt to comprehend a scene from the perspective of cognitive psychology and propose a flexible cognition-based hybrid motivation (CHM) framework, encompassing reasonable experience and assumption-based inference (EAI), and best-effort object labeling (BEL). EAI describes the common object appearing in an indoor scene, such as table, bed, and so on. In BEL, other objects are modeled and categorized in an abstract and hierarchical way, according to their context. A bag-of-visual words (BoVW)-based multilayer abstract semantics labeling (MASL) method is proposed to achieve this goal. Our approach is highly modularized, with no restrictions on its operation other than requiring the ability to train on data, making our method easy to extend and applicable to many other tasks with similar outputs.

The paper is organized as follows. Relative research is concisely introduced in Sec. 2. Our CHM framework and MASL classification method are proposed in Sec. 3. Experimental results and analysis are provided in Sec. 4. Finally, we summarized our current and future work in Sec. 5.

2 Related Research

In general, object annotation is a process of learning to answer the "what" question from given images, and includes geometry recovery and object categorization. We will introduce them separately.

*Address all correspondence to: Wei Zhao, E-mail: zhaowei@hit.edu.cn

1017-9909/2015/$25.00 © 2015 SPIE and IS&T

2.1 *Room Geometry Recovery*

A major challenge for indoor scene annotation is that most indoor scenes are cluttered by furniture and decorations, the appearances of which vary drastically across scenes. It is hard to model this characteristic consistently. The Manhattan world assumption states that there exist three dominant vanishing points that are orthogonal;[17] this assumption has significantly reduced the difficulty of recovering space layout, and has benefited research in overcoming the challenge of indoor scene annotation immensely. The commonly used method is to recover the geometry of indoor scenes and to estimate the position of each object. The input samples include both single still image and video sequences.[18–24] With the development of computer vision techniques, RGB-D images captured by devices such as Kinect can greatly reduce the difficulty and cost of generating 3-D scenes. Understandably, research studies are showing increasing interest in scanning and building 3-D scenes[25–27] with RGB-D cameras. We will utilize the object detection method described by Hedau et al.[28] to achieve the goal of object detection of an indoor scene. After this, the positions of most objects are established and will be utilized as input to the annotation method.

2.2 *Object Classification*

Object classification has been an intensively studied field in computer vision for the past decades. Many outstanding studies and corresponding improvements have been carried out to solve this problem, such as k-nearest neighbors (KNN), decision tree, Bayes model, support vector machine, linear discriminant analysis, graph-based methods, and multiple-instance-based learning methods.[29–32]

Among these methods, BoVW[33] is one of the most widely used methods due to its simplicity and effectiveness. In the learning model of BoVW, visual words are first obtained by k-means clustering local features. Then the image is represented by bag-of-features to train the classifier. However, it has four major drawbacks: (i) the quality of visual vocabulary is sensitive to data set size;[34] (ii) spatial relationships of image patches are ignored during the construction of visual vocabulary;[35] (iii) hard-assignment k-means clustering affects the generation of semantically optimized visual words;[36] and (iv) performance of annotation is affected by semantic gap.[37]

In the previous work, there are four types of strategies for improving BoVW: segmentation, coding, ambiguity, and semantic compression. Segmentation-based methods utilize ROI to remove background that is irrelevant to visual word generation.[38] Improvements to coding strategy[39] and the introduction of ambiguity increase the descriptive ability of visual words.[40] Local information is introduced to effectively generate complementary features.[41] Semantic compression is proposed to improve efficiency and performance.[42,43]

Here, we target the semantic gap between visual features and semantic concepts to improve the performance of BoVW by dividing semantic concepts into several hierarchies to narrow semantic gaps. To train each concept classifier, visual vocabularies are extracted from samples of each inherited object class to train the abstract classes from bottom to top, and the model runs from top to bottom for classification. We will show this process in detail in Sec. 3.

3 CHM Scene Annotation Framework

One fascinating human characteristic is the ability to categorize objects with only a few labeled training instances and to infer the categories of an indoor scene despite its inner decoration. Humans are born with the ability to perform adaptive object categorization. Is it possible for a computer to achieve this goal with the support of machine learning techniques? To solve this problem, we will describe our framework in the following subsections. The framework can be roughly divided into two stages: first, the spatial layout of a scene is estimated and objects in the scene are detected; then the objects will be annotated by BoVW and the proposed MASL classification methods, respectively, according to the inference processes of humans.

3.1 *Experience and Assumption-Based Inference*

The commonly accepted inference process of humans is shown in Fig. 1, according to the descriptions of cognitive theory.[44] Long-term memory (LTM) operates like a huge knowledge warehouse serializing all kinds of information, whereas short-term memory (STM) contains a much smaller volatile storage space, and is the first point of handling of short-term knowledge learned from environmental stimulation. After knowledge in STM is serialized into LTM, key information is extracted to form or update experience for future use. If we encounter a novel object and cannot come to a conclusion about it from STM, LTM will be referred to if we do not immediately dismiss it. For example, when we see a cat for the first time, we will remember its key patterns and save them into both STM and LTM. Knowledge of the cat in STM may be overwritten by other knowledge since the space is limited. However, when we see a different cat, we will refer to LTM and recall its category. This is a very important ability in cognition.

Generalization is an inherited ability for humans, consisting not only of the ability to extract patterns and learn from a limited number of samples, but also the ability to store common key information to ensure similar objects can be recognized in the future. It is a highly advanced skill compared with the ability of a computer; however, as we understand it, the inference process of a modern computer vision algorithm is a similar process as shown in Fig. 2. As demonstrated by the results of image categorization contests such as Pascal VOC challenge and the ImageNet large scale visual recognition challenge, the performance of object recognition algorithms has greatly improved in the last two decades of development. In some circumstances, the performance is

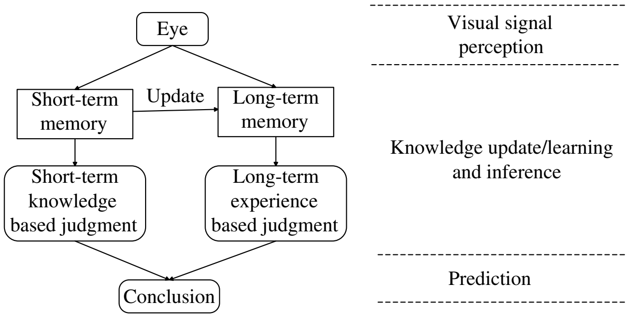

Eye
Short-term memory
Update
Long-term memory
Short-term knowledge based judgment
Long-term experience based judgment
Conclusion
Visual signal perception
Knowledge update/learning and inference
Prediction

Fig. 1 The human comprehension process.

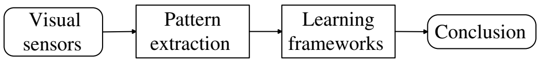


Fig. 2 Inference process of modern computer vision algorithms.

good enough for practical purposes. However, compared with the natural human inference process, most methods in computer vision are STM-like, i.e., the LTM process is ignored, meaning object categorization methods are nonadaptive and inflexible. This is the main reason that humans, despite having one brain, can effectively recall and recognize many types of objects,[44] while computers require training and adjusting many times over with different testing samples.

In this paper, we propose a CHM object categorization framework for adaptive scene classification. Inspired by Porway et al.,[45] we divide the problem of scene annotation into two steps for simplification: EAI and BEL. EAI simulates the experience-based inference ability of humans with manually set rules. For example, given the concept of "indoor scene," despite the variation in detail of each indoor image, there are common objects that appear in our mind, such as wall, windows, tables, chairs, and so on, as shown in Fig. 3. First, we detect and categorize objects using EAI. The remaining objects are then detected and categorized by BEL with reasonable context to achieve higher-level adaptation. In EAI, objects are detected by Hedau et al.,[28] then classified with BoVW. In BEL, the proposed MASL is adopted for classification and the model is ready to be extended for different situations. The introduction of EAI makes our approach a two-step flexible inference process, differing from other methods of indoor scene analysis.[47–50]

3.2 *BEL Object Labeling*

To deal with objects exceeding the empirical field of EAI, inspired by Luo et al.,[51] we propose a BoVW-based object classification method for BEL, known as MASL, by introducing semantic hierarchies with different levels of abstraction. It is modularly designed according to the organization of human memory,[44] making it convenient to add knowledge from different object categories without affecting existing knowledge. The motivation for proposing the MASL categorization method is that the real world is powered by hierarchical structures, and relevant research has been proven effective.[52] The difference between the previous method and MASL is the structure of abstraction. In MASL, the abstract layers are expandable instead of being a fixed frame. Furthermore, the abstraction techniques are introduced to generate layers with different abstract levels to describe the semantic concepts more clearly.

Semantic hierarchies are helpful for image classification as they supply a hierarchical framework for image classification and provide extra information in both learning and representation.[52] Three types of semantic hierarchy for image annotation have been recently explored: (1) language-based hierarchies based on textual information,[53] (2) visual hierarchies based on low-level image features,[54] and (3) semantic hierarchies based on both textual and visual features.[55] Here, we extend BoVW by introducing middle and upper hierarchies of abstract semantics, which are constructed by semantics assigned visual words extracted from concrete categories (CCs). The hierarchical structure of BoVW and MASL is shown in Fig. 4. According to the principle, levels of abstraction increase from bottom to top. Consequently, descriptive ability increases while the difference between each CC is dimmed and common attributes are preserved. From the figure, we can see that MASL is a superset of BoVW. If the abstract hierarchies are omitted, MASL degrades into BoVW.

The whole process of MASL consists of two parts: bottom-up semantic classifier learning and top-down classification. In the learning process, each concrete classifier BoVW_j is first trained with concrete semantic visual vocabulary. Then abstract semantic classifiers including U-SVM and every M-SVMs will be trained. Learning is from bottom to top because the abstraction level rises from bottom to top. At the stage of classification, the inference of the category of a testing image is from top to bottom, reflecting the decrease in the level of abstraction.

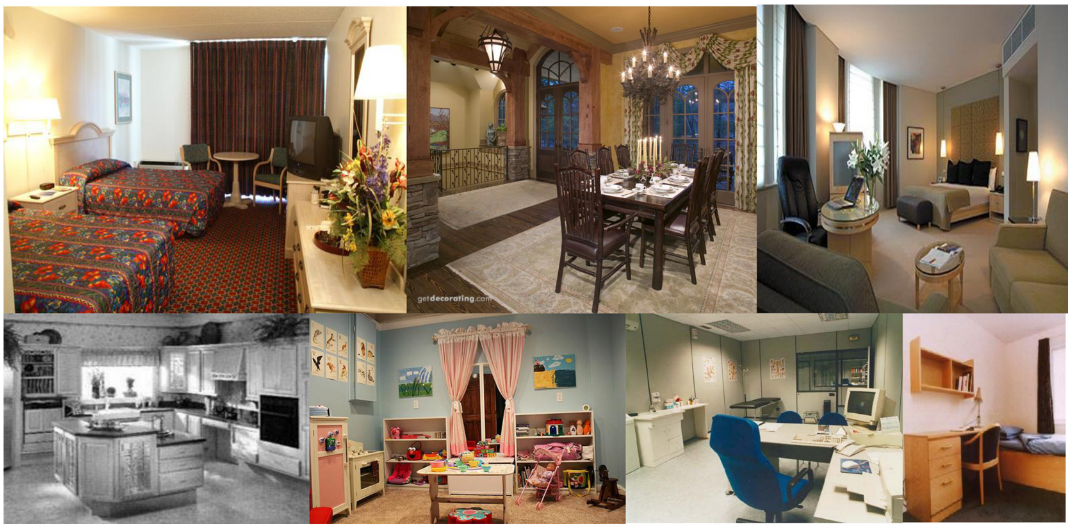


Fig. 3 A common structure of indoor scenes provided by Quattoni and Torralba.[46] We can see that although the decoration of each room is quite different, they share a similar structure and some common objects, e.g., wall, table, chair, window, etc.

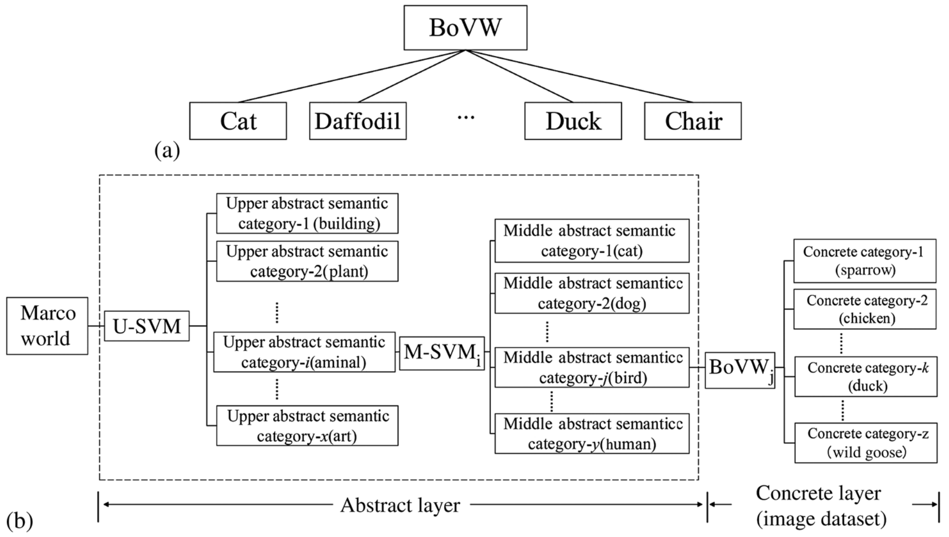

BoVW
Cat
Daffodil
...
Duck
Chair
(a)
Marco world
U-SVM
Upper abstract semantic category-1 (building)
Upper abstract semantic category-2(plant)
Upper abstract semantic category-i(aminal)
Upper abstract semantic category-x(art)
M-SVMi
Middle abstract semantic category-1(cat)
Middle abstract semanticc category-2(dog)
Middle abstract semanticc category-j(bird)
Middle abstract semanticc category-y(human)
BoVWj
Concrete category-1 (sparrow)
Concrete category-2 (chicken)
Concrete category-k (duck)
Concrete category-z (wild goose)
Abstract layer
Concrete layer (image dataset)
(b)

Fig. 4 Structure of bag-of-visual words (BoVW) and multilayer abstract semantics labeling (MASL) models. (a) The flat structure of BoVW. (b) Hierarchical structure of MASL.

3.2.1 *Bottom-up semantic classifier learning*

The purpose of bottom-up semantic learning is to train the semantic classifiers with a semantic visual vocabulary of corresponding layers.[56] This process can be divided into three steps: first, each concrete classifier BoVW_j in Fig. 4(b) is trained by a visual semantic attribute (VSA) composed by semantic visual words from semantic preserving BoW (SPBoW).[56] The input of BoVW_j from the concrete layer is images from data sets; then to train the middle abstract classifier M-SVM_i of middle abstract semantic category (MASC), samples from every CC of the MASC are randomly selected with equal probability to ensure every category has the same chance of being selected to construct a visual vocabulary to improve the descriptive ability. The semantic visual vocabulary is generated from selected samples by SPBoW to train M-SVM_i. The U-SVM classifier for the upper abstract semantic category (UASC) is trained in the same way to complete the learning process.

3.2.2 *Top-down classification*

Following completion of the bottom-up training stage, to find the category of an input image, UASC u is first generated by U-SVM. Then the corresponding MASC m is calculated by M-SVMs. Finally, the CC c is concluded. The processes of classification are described as

$$u = \operatorname{argmin}\,[D(F_t, F_i^u)], \quad m = \operatorname{argmin}\,[D(F_t, F_j^m)],$$
$$c = \operatorname{argmin}\,[D(F_t, F_k)], \tag{1}$$

where F_i^u and F_j^m are the visual attributes of the i'th upper and j'th middle abstract categories, F_k^m is the visual vocabulary, and D is the measurement function utilized by classifiers. The visual semantic vocabulary is generated by SPBoW.

Since the decision processes are sequential, if u was incorrect, the remaining inference processes would be meaningless. Thus, we adopt a two-step verification strategy to decrease the dependence on upper layers and reduce error rates. First, the testing image I is passed through U-SVM, then before the final decision is made, I is passed to all M-SVMs for further verification. Let the output values of U-SVM and M-SVMs be $P^{(u)}$ and $P^{(m)}$, respectively. The middle abstract category of the corresponding layer is finally decided by the following criterion:

$$C_{\text{middle}} = \sum_{i=1}^{U} \sum_{j=1}^{M} \operatorname{argmax}[p_i^{(u)} + p_j^{(m)}], \tag{2}$$

where U and M are the number of upper and middle classifiers, and $p_i^{(u)}$ and $p_j^{(m)}$ are the output values of i'th U-SVM and j'th M-SVM, respectively, $p_i^{(u)}$, $p_j^{(m)} \in [-1,1]$. At last, the traditional strategy of BoVW is utilized to get n outputs $p_1, p_2, \ldots, p_n$, where n is the number of categories under each classifier. Image I is classified according to the following criterion:

$$C = \operatorname*{argmax}_{t=1,\ldots,n}(p_t). \tag{3}$$

3.2.3 *Summary of multilayer abstract semantics labeling*

Compared with original BoVW, the proposed model makes its improvement from the perspective of abstraction by introducing abstract semantic layers to narrow semantic gaps. Semantic visual vocabulary is utilized as a training feature and strategy to improve the performance of classifiers. The proposed learning algorithm is described in Algorithm 1, where $k = 1, \ldots, m$, m is the size of the codebook, CC_k is short for k'th CC, MAC_j is the abbreviation of the j'th middle abstract category, UAC_i is short for the i'th upper abstract category, SVVS stands for semantic visual vocabulary set, and m is the number of CC under j'th MAC. Inh_k^j is the generated visual words by SPBoW from CC_k under MASC_j.

Algorithm 1 The learning process of MASL.

```
Input: Training image set and image $I$ of unknown category.
Output: Category of $I$.
Preparation stage
For each CC$_k$ under MAC$_j$, generate SVVS, where $v_q$ and $s_q$ are visual words and the corresponding semantic information.
The SVVS of MAC$_j$ under UAC$_i$ is constructed by $M_j = \bigcup_{k=1}^{z} \mathrm{Inh}_k^j$ and M-A$_i = \bigcup_{j=1}^{y} M_j$.
For each UAC$_i$, randomly select SVVS with equal probability from each Inh$_k^j$. Let U-ABS$_i = \bigcup_{j=1}^{y} \bigcup_{k=1}^{z} \mathrm{Inh}_k^j$, U-A $= \bigcup_{i=1}^{x}$ U-ABS$_i$.
Training stage:
For every MAC$_j$, TRAIN$(\mathrm{BoVW}_j, \mathrm{Inh}_k^j)$
For every UAC$_i$, TRAIN$(\text{M-SVM}_i, \text{M-A}_i)$
TRAIN(U-SVM,U-A)
Categorization stage:
For an input image, calculate its upper and middle abstract categories by Eq. (1). Then find the category $c'$ that satisfies Eq. (2) as the start for the next search.
Calculate category $c$ of $I$ with Eq. (3).
Return $c$.
```

3.3 *Summary of the Proposed Cognition-Based Hybrid Motivation Framework*

We will summarize the general framework of CHM in Fig. 5 and Eq. (4). For every object in the scene, context information[13] is utilized to assist in classification. Emp-Obj and NEmp-Obj represent the objects that frequently and infrequently occur in an indoor scene according to experience, respectively. According to the mechanism of cognition process described in Fig. 1, knowledge learned in EAI will be added to BEL to simulate the human learning process.

$$\begin{aligned}
&\text{SceneAnnotation} \rightarrow \text{Obj-Detection} \cup \text{Obj-Classification}\\
&\text{Obj-Classification} \rightarrow \text{EAI} \cup \text{BEL}\\
&\text{EAI} \rightarrow \text{INF-BY-BoVW}\\
&\text{BEL} \rightarrow \text{INF-BY-MASL}\\
&\text{INF-BY-BoVW} \rightarrow (\text{Emp-Obj}_1^*) \cup (\text{Emp-Obj}_2^*) \cup \cdots\\
&\qquad\qquad \cup (\text{Emp-Obj}_n^*)\\
&\text{INF-BY-MASL} \rightarrow (\text{NEmp-Obj}_1^*) \cup (\text{NEmp-Obj}_2^*) \cup \cdots\\
&\qquad\qquad \cup (\text{NEmp-Obj}_m^*).
\end{aligned} \tag{4}$$

Here, we are using an asterisk in the expression, representing an arbitrary number of objects.

$$\begin{aligned}
\text{Object}^* \rightarrow &\ \varphi|\text{Object}|(\text{Object})(\text{Object})|(\text{Object})(\text{Object})\\
&\times (\text{Object}).\ldots
\end{aligned} \tag{5}$$

4 Experiments and Analysis

In this section, we will show the experimental results of MASL and CHM in the following subsections, including horizontal and vertical benchmarks, and semantic quantification. MASL is first tested on multiple data sets, and then utilized in CHM for indoor scene annotation.

4.1 *Data Sets and Experimental Settings*

The proposed framework is validated on four popular data sets used in classification experiments. The details are listed below:

The Caltech-101 data set[57] contains 9197 images in 101 categories. The size of each image is roughly 300×200 pixels. The outline of each object is carefully annotated. For each category, at least 30 images are randomly selected for the learning process as described by Wang et al.,[58] and the rest of the images are used as testing samples.

The PASCAL VOC 2007 data set[59] contains 9963 images in 20 object categories. There is a bounding box for each positive example of an object. Compared with Caltech-101, the data set is more difficult since the number of instances in an image is not always one. For Caltech-101 and PASCAL VOC 2007, we randomly select additional 10 images from each category to form a subset for vertical benchmarks described in Sec. 4.2.

The Microsoft Research Cambridge (MSRC) data set[60] contains 591 images in 23 object classes. Each image is

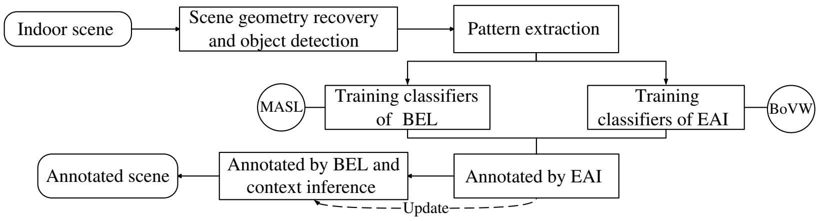

Indoor scene
Scene geometry recovery and object detection
Pattern extraction
MASL
Training classifiers of BEL
Training classifiers of EAI
BoVW
Annotated scene
Annotated by BEL and context inference
Annotated by EAI
Update

Fig. 5 Learning and annotating processes of the proposed cognition-based hybrid motivation framework. First, the objects are detected by the method proposed by Hedau et al.[28] After the modules are trained, images are labeled successively by experience- and assumption-based inference (EAI) and best-effort object labeling (BEL). To simulate the human inference process, the newly learned knowledge of objects in EAI will be added to BEL, including the category and corresponding classifiers.

labeled by pixels. The resolution of images is roughly 320 × 240 pixels. Two categories, "horse" and "mountain," were removed from evaluation due to their small number of positive samples, as suggested in the description page of the data set. MSRC and Caltech-101 are used for horizontal benchmarks described in Sec. 4.3. For MSRC, images are equally divided for training and testing.

The MIT indoor is a data set of 15,620 images over 67 indoor scenes.[46] There are at least 100 images per category. Here, we follow the settings described by Quattoni and Torralba.[46] The percentage of training and testing images of each category is 80% and 20%, respectively. All experiments were carried out on a workstation with quad-core 2.13 GHz CPU and 12 GB memory.

4.2 *Experiment I: Vertical Benchmarks*

As mentioned in the previous research,[36] the performance of BoVW-based methods is affected by the size of visual vocabulary. In vertical benchmarks, self-evaluation of MASI and BoVW under different parameters is utilized to determine the optimized size of visual vocabularies. Testing images are from Caltech-101[57] and PASCAL VOC 2007,[59] as shown in Fig. 6, with results given in Fig. 7. We can see from the result that a larger size of codebook is actually beneficial for reducing the error rate, but does not mean a consistently better performance. Complex visual vocabulary lowers the performance, and this is consistent with the previous research.[61] Meanwhile, the larger the codebook, the more computational cost is needed to build the visual vocabulary. Since the underlying techniques utilized by MASL are also based on BoVW, trends for both methods are similar. Compared with BoVW, the introduction of middle abstract layers in MASL leads to a better performance. Here, the size of visual vocabulary is set to 6000.

4.3 *Experiment II: Horizontal Benchmarks*

In this section of the experiments, we will compare MASL with other similar classification methods and then show the results of semantic gap quantification.

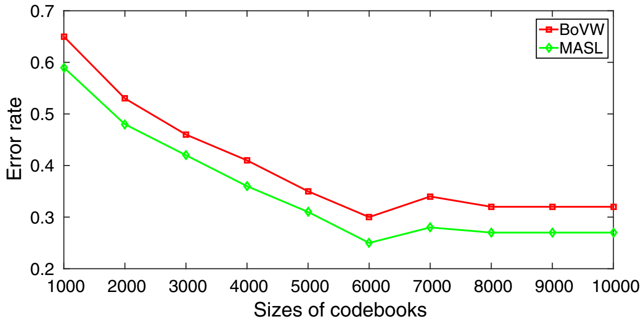


Fig. 7 Performance of BoVW and MASL with different-sized visual vocabularies.

4.3.1 *Comparison on classification*

As previously performed by Zhou et al.,[62] several approaches[62–64] are evaluated on selected data sets.[35,63,65,66] Results are given in Fig. 8(a) and classification performances of individual classes for MASL are reported in detail through confusion tables in Figs. 8(b) and 8(c). Names of data sets are abbreviated according to their providers.[62] Since OT[65] and FP[63] are part of the LS[35] data set, we only give the confusion table for LS and LF[66] (similar to Zhou et al.[62]) to show how the different categories are confused in Fig. 8.

We can see from Fig. 8 that MASL achieves the best performance on all tests. This is because by introducing hierarchical semantics, the codebook generated by MASL is more discriminative and effective in reducing the semantic loss during codebook generation. From the confusion table shown in Fig. 8, we can see that although considerable confusion exists between man-made and natural scene categories (e.g., bedroom versus living room, kitchen versus living room), our MASL still outperforms other methods, including that of Zhou et al.,[62] indicating the effectiveness of the proposed MASL classification method.

4.3.2 *Semantic gap quantification*

Semantic gap measurement described by Tang et al.,[67] is utilized to quantify semantic gaps, which can be described as

Fig. 6 Sample images from PASCAL VOC 2007 and Caltech-101.

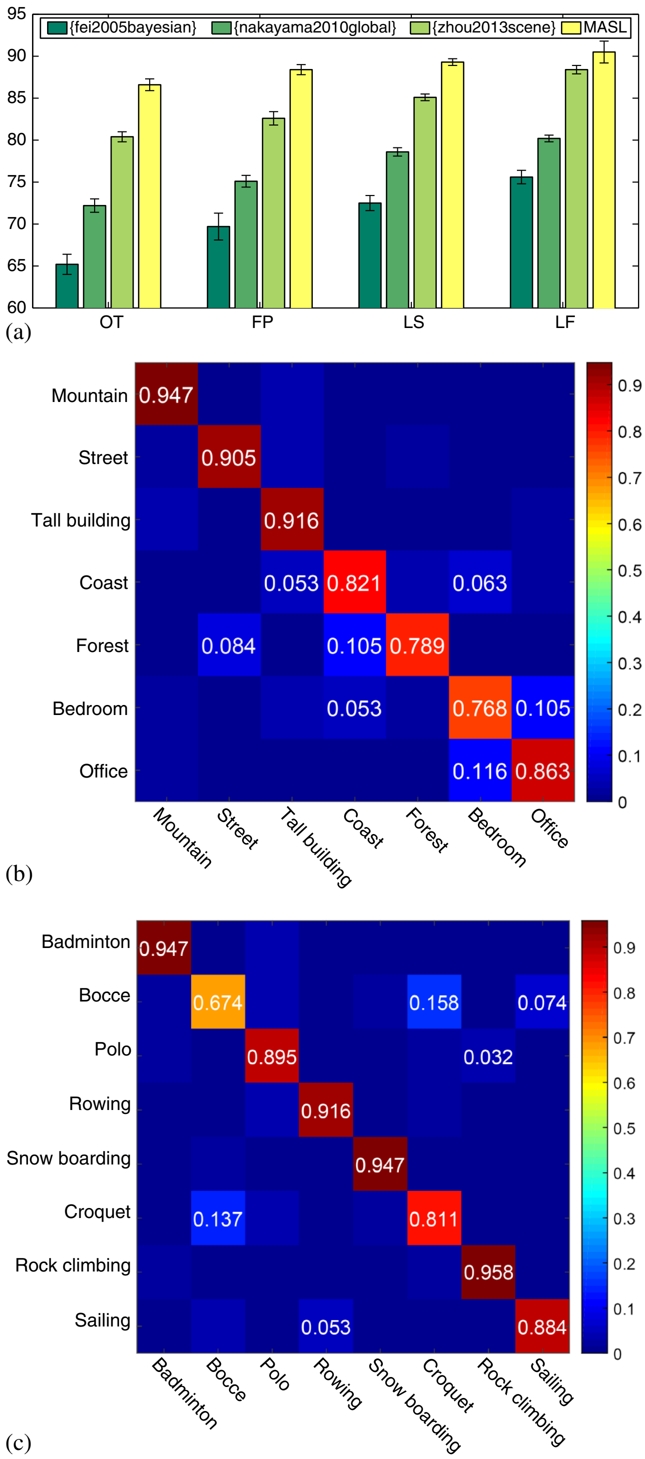


Fig. 8 Classification results and corresponding confusion tables of selected data sets. (a) Comparison with other classification methods. (b) Confusion table for LS data set. (c) Confusion table for LF data set.

$$\text{Im-SG}(x_i) = \frac{1}{k} \sum_{x_j \in N(x_i)} \text{dis-sim}(x_i, x_j), \tag{6}$$

where $N(x_i)$ represents the set of the k nearest neighbors of x_i in the visual space. Semantic distance dis-sim(x_i, x_j) between x_i and each of its neighbors x_j is measured by the cosine distance between the vectors of their tags. For MASL, semantic gap is quantified as

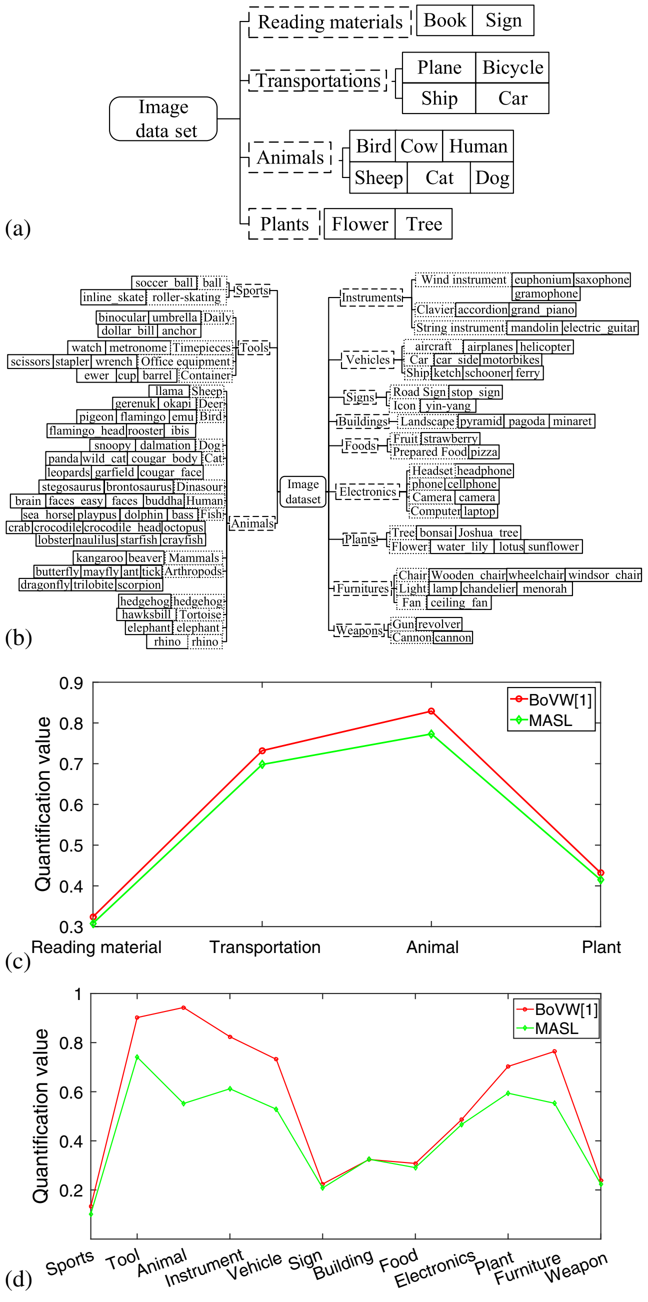


Fig. 9 Hierarchical structure of image data sets and the corresponding semantic gap quantification results. (a) The hierarchical structure of MSRC. (b) The hierarchical structure of Caltech-101. (c) Semantic gap quantization result of MSRC. (d) Semantic gap quantification result of Caltech-101.

$$\text{Im-SG}(x_i) = {}_{I}^{M}\text{Im-SG}(x_i) + {}_{M}^{U}\text{Im-SG}(x_i), \tag{7}$$

where ${}_{I}^{M}\text{Im-SG}(x_i)$ represents the image semantic gap between the concrete layer and MASC, and ${}_{M}^{U}\text{Im-SG}(x_i)$ denotes the image semantic gap between MASC and UASC. To fully evaluate the semantic gaps of MASL, we constructed multiple semantic hierarchies for MSRC[60] and Caltech-101,[57] as shown in Figs. 9(a) and 9(b). Four abstract categories were constructed by selected 14 CCs with sufficient and unambiguous training/testing images. Experimental results are provided in Figs. 9(c) and 9(d). We can see that MASL was more effective in narrowing the semantic gap between concepts and visual data. For abstract categories with few CCs, the difference between BoVW and MASL was not significant. The performance of MASL was slightly better than that of BoVW. However, for larger abstract categories such as animal on Caltech-101, substantial improvement was observed (0.943 for BoVW versus 0.552 for MASL on Caltech-101; 0.829 for BoVW versus 0.773 for MASL on MSRC). This is because, when constructing visual data, the introduction of upper and middle abstract layers prevents interference from completely irrelevant categories. Conversely, when the scale of abstract category is small, the disturbance between each CC is relatively small, thus narrowing the semantic gap between the two methods is approximately coincident.

4.4 *Experiment III: Benchmarks on Scene Annotation*

In this experiment, we fully evaluate the overall performance of the proposed CHM framework on the MIT-indoor data set. Here, we initially define three basic objects for the indoor scene: table, chair, and bed, and the inference process is given in

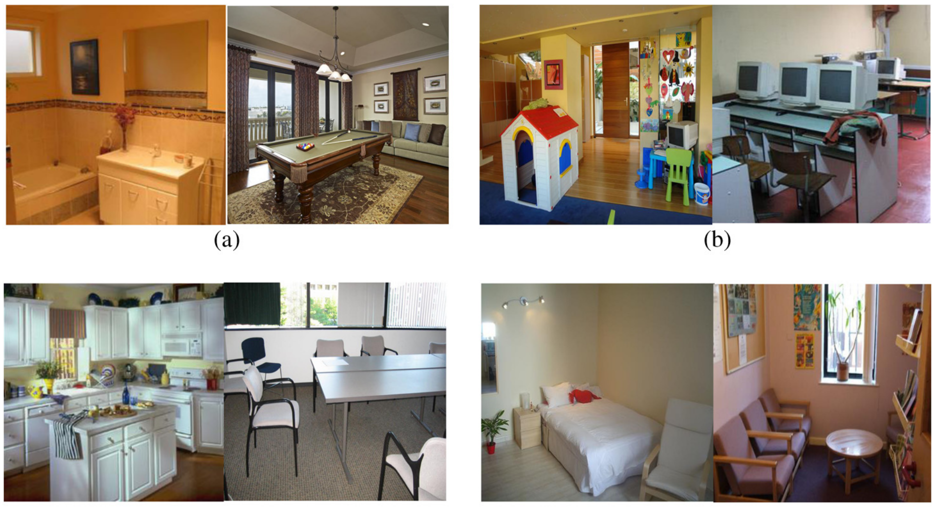
$$\text{INF-BY-BoVW} \rightarrow (\text{table}^*) \cup (\text{chair}^*) \cup (\text{bed}^*) \, . \tag{8}$$

Similar to Quattoni and Torralba,[46] the tests are divided into four parts containing different categories. Sample images are shown in Fig. 10.

The methods proposed by Hossein et al.[68] and Gong et al.[69] are based on a convolutional neural network (CNN). CNNs have become prominent in machine learning during the past decade due to their highly effective performance. Optimized parameters are set for each method, and a CNN is implemented by EBLearn.[70] To compare methods fairly across different applications, the performance of all methods is evaluated by the precision of classification on both manually labeled and automatically detected[28] objects in the data set. The results of average performances are given in Table 1. We can conclude that our CHM framework achieves a comparable performance with other methods. Both EAI and contextual information play important roles in boosting the performance of the framework and CHM outperforms other annotation methods[71–73] on all tests. Although CHM does not outperform CNN-based methods on the second and fourth tests, the performance gaps between them are not remarkable. Meanwhile, CHM outperforms the CNN-based methods on the first and third tests, and the average performance of CHM is better than all classification methods, proving its effectiveness, as most samples contain objects that fit the category definition well, given by Eq. (8). In the second and fourth tests, two CNN-based methods, respectively, achieved better performance, as differences between categories are relatively small, i.e., the testing sets are more confusing for other methods. The structural advantage of CNN, with multiple layers and neurons, ensures that CNN-based methods achieve a better performance in this situation.

A sample of the object detection and categorization results is shown in Fig. 11. Figures 11(a)–11(d) show the correct object detection and categorization results of indoor scene categories. We can see from Figs. 11(b) and 11(d) that the CHM framework deals with empirical and out-of-experience objects properly, showing self-adaption under the above assumption; Figs. 11(a) and 11(c) show the detection of the empirical objects we set; Fig. 11(b) shows the importance of environment context. When the empirical objects are detected [a table in Fig. 11(b)], objects around it are considered as out-of-experience objects and will be categorized by BEL. Figures 11(e) and 11(f) show the importance of EAI. With a table first detected, the object above would then be categorized as a chandelier according to context. All the results prove that our framework is useful for scene annotation.

Error examples are also shown. Figure 11(g) shows that some objects are incorrectly classified; in this case, the chest is classified as a chair. Figure 11(h) shows error detection results from the detection module. The chair and chest are detected with the table so that the whole object is classified as a table. Figure 11(i) shows both misdetection and misclassification. A shelf and table are detected as one object and classified as a chest. From these incorrect results, we can



(a)

(b)

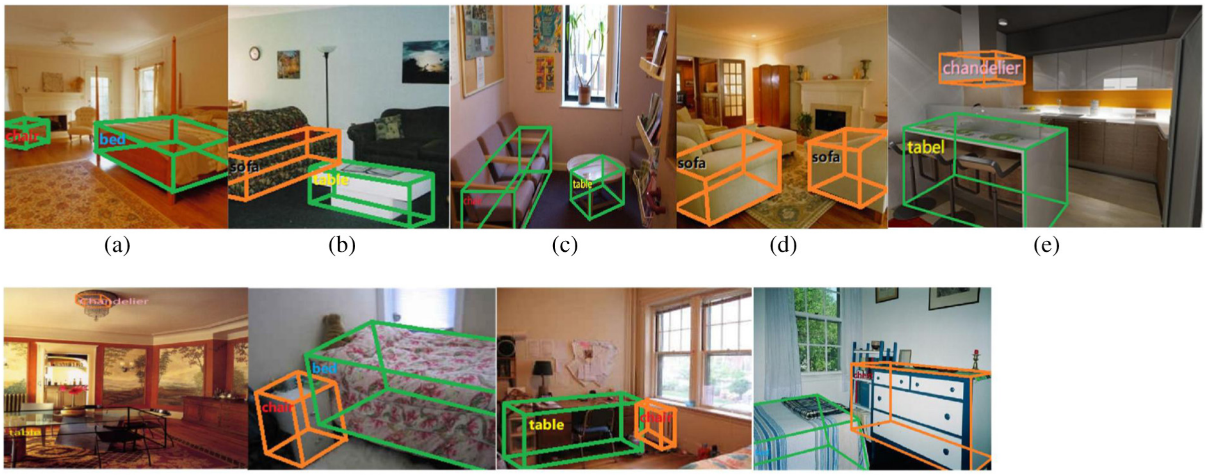
(c)

(d)

Fig. 10 Sample images from MIT-indoor data set. As described by Quattoni and Torralba, the whole data set is divided into four parts: (a) P1, (b) P2, (c) P3, and (d) P4.[46]

Table 1 Comparison of results between CHM and other methods on MIT-indoor data set.[46] The data set is divided into four groups for detailed tests.

	Group 1	Group 2	Group 3	Group 4	Average
Wang et al.[71]	34.56 ± 0.8	31.27 ± 0.4	44.21 ± 0.4	63.46 ± 2.6	43.38 ± 1.05
Tsai et al.[72]	32.56 ± 0.5	32.31 ± 0.4	45.43 ± 0.7	63.78 ± 2.2	43.52 ± 0.95
Xie et al.[73]	31.15 ± 0.25	32.92 ± 0.48	47.9 ± 0.64	56.46 ± 1.92	42.11 ± 0.82
Hossein et al.[68]	33.42 ± 0.37	33.67 ± 0.34	46.16 ± 0.21	**68.9** ± **1.8**	45.54 ± 0.68
Gong et al.[69]	32.56 ± 0.5	**34.38** ± **0.4**	42.17 ± 0.67	67.52 ± 1.6	44.16 ± 0.79
CHM					
Context	31.22 ± 1.3	26.54 ± 0.9	43.72 ± 0.8	64.78 ± 1.7	41.57 ± 1.18
EAI	34.14 ± 0.8	30.26 ± 0.4	44.24 ± 0.6	60.45 ± 1.1	42.27 ± 0.73
Context + EAI	**36.63** ± **0.7**	32.53 ± 0.3	**48.17** ± **0.5**	65.52 ± 1.3	**45.71** ± **0.7**



(a) (b) (c) (d) (e)

(f) (g) (h) (i)

Fig. 11 (a)–(i) Samples of the object detection and categorization results.

conclude that current object detection methods[28] remain unable to precisely restore the original 3-D structure for some situations, such as for low-resolution images with indistinct 3-D structure, and this needs further investigation. Additionally, much work still needs to be done to improve the performance of MASL.

The purpose of our study was to simulate the scene annotation process of humans to make the annotation process more appropriate and human-like. Although the improvement of the performance is not remarkable compared with existing methods, our work provides yet another method for indoor scene annotation, and a preliminary investigation into knowledge-based scene understanding by means of rule inference constructed from annotated objects. Differing from traditional image annotation methods, our method is bio-inspired, introducing cognitive models and inference rules to simulate the annotating process of humans. Our proposed CHM framework, as evidenced by the results shown in Table 1, outperforms the methods proposed by Wang et al.[71], Tsai et al.,[72] and Xie et al.[73] on all tests. In general, CHM performs better than CNN-based methods,[68,69] which are the state-of-the-art results. The performance of CHM is influenced by existing classification algorithms. Object classification methods based on BoVW have achieved significant improvement over the last few years; however, there remains significant room for improvement. Compared with the object detection and learning skills of humans, existing algorithms in computer vision are still far from satisfactory.

The progressive experimental results of classification, semantic gap quantification, and scene annotation have proven the effectiveness of the proposed MASL classification method and CHM scene annotation framework from multiple perspectives. Although far from perfect, the CHM annotation framework demonstrates the possibility of scene annotation combining cognition theory and computer vision. Finally, due to the modular design of our framework, its performance will improve with the development of existing object detection and recognition methods.

5 Conclusion

In this work, we address the scene annotation problem and propose a framework to simulate the human cognitive process. Compared with the previous works, our experimental results have proven the effectiveness of our framework on both narrowing semantic gaps and boosting the performance of classification. However, limitations remain and, although the performance of CHM is comparable with state-of-the-art methods, improvement is still required. We believe that the performance of CHM will improve with the development of object classification algorithms, due to its modular design. Following this preliminary study on indoor scene annotation, future research will target the interpretation of an indoor scene based on the annotated objects.

Acknowledgments

This research was supported by the National Science Foundation of China (Grant Nos. 61171184 and 61201309).

References

1. D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.* **45**(1), 346–362 (2012).
2. S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer Science & Business Media, London (2009).
3. C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: a survey," *Comput. Vision Image Understanding* **117**(11), 1610–1627 (2013).
4. X. He, R. S. Zemel, and M. Carreira-Perpindn, "Multiscale conditional random fields for image labeling," in *Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2004)*, Vol. 2, pp. II-695, IEEE (2004).
5. S. Gould et al., "Multi-class segmentation with relative location prior," *Int. J. Comput. Vision* **80**(3), 300–316 (2008).
6. N. Payet and S. Todorovic, "Hough forest random field for object recognition and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(5), 1066–1079 (2013).
7. B. T. C. G. D. Roller, "Max-margin Markov networks," *Adv. Neural Inf. Process. Syst.* **16**, 25 (2003).
8. J. Fan et al., "Structured max-margin learning for inter-related classifier training and multilabel image annotation," *IEEE Trans. Image Process.* **20**(3), 837–854 (2011).
9. W. Zhang et al., "Multi-kernel multi-label learning with max-margin concept network," in *IJCAI Proc. Int. Joint Conf. on Artificial Intelligence*, Vol. 22, No. 1, p. 1615 (2011).
10. R. Mottaghi et al., "Analyzing semantic segmentation using hybrid human-machine CRFs," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3143–3150, IEEE (2013).
11. C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, IEEE (2008).
12. L. Ladicky et al., "Graph cut based inference with co-occurrence statistics," in *Computer Vision-ECCV 2010*, pp. 239–253, Springer (2010).
13. M. J. Choi, A. Torralba, and A. S. Willsky, "Context models and out-of-context objects," *Pattern Recognit. Lett.* **33**(7), 853–862 (2012).
14. Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3D scenes," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2993–3000, IEEE (2013).
15. O. Blomberg, "Conceptions of cognition for cognitive engineering," *Int. J. Aviat. Psychol.* **21**(1), 85–104 (2011).
16. T. Tang and H. Qiao, "Improving invariance in visual classification with biologically inspired mechanism," *Neurocomputing* **133**, 328–341 (2014).
17. J. M. Coughlan and A. L. Yuille, "The Manhattan world assumption: regularities in scene statistics which enable Bayesian inference," in *NIPS*, pp. 845–851 (2000).
18. A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-D scene structure from a single still image," in *IEEE 11th Int. Conf. on Computer Vision, 2007 (ICCV 2007)*, pp. 1–8, IEEE (2007).
19. V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *IEEE 12th Int. Conf. on Computer vision*, pp. 1849–1856, IEEE (2009).
20. G. Tsai et al., "Real-time indoor scene understanding using Bayesian filtering with motion cues," in *2011 IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 121–128, IEEE (2011).
21. A. G. Schwing et al., "Efficient structured prediction for 3D indoor scene understanding," in *2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2815–2822, IEEE (2012).
22. V. Hedau, D. Hoiem, and D. Forsyth, "Recovering free space of indoor scenes from a single image," in *2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2807–2814, IEEE (2012).
23. H. Wang, S. Gould, and D. Roller, "Discriminative learning with latent variables for cluttered indoor scene understanding," *Commun. ACM* **56**(4), 92–99 (2013).
24. S. Ramalingam et al., "Manhattan junction catalogue for spatial reasoning of indoor scenes," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3065–3072, IEEE (2013).
25. T. Shao et al., "An interactive approach to semantic modeling of indoor scenes with an RGBD camera," *ACM Trans. Graph.* **31**(6), 136 (2012).
26. X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: features and algorithms," in *2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2759–2766, IEEE (2012).
27. J. P. Valentin et al., "Mesh based semantic modelling for indoor and outdoor scenes," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2067–2074, IEEE (2013).
28. V. Hedau, D. Hoiem, and D. Forsyth, "Thinking inside the box: using appearance models and context based on room geometry," in *Computer Vision-ECCV 2010*, pp. 224–237, Springer (2010).
29. M. Wang, X. Liu, and X. Wu, "Visual classification by l1-hypergraph modeling," *IEEE Trans. Knowl. Data Eng.* **27**, 2564–2574 (2015).
30. M. Wang et al., "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.* **19**, 733–746 (2009).
31. J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Process.* **21**, 3262–3272 (2012).
32. R. Hong et al., "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.* **44**, 669–680 (2014).
33. G. Csurka et al., "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision ECCV*, Prague, Vol. 1, No. 1–22, pp. 1–2 (2004).
34. M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vision* **62**(1–2), 61–81 (2005).
35. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. 2, pp. 2169–2178, IEEE (2006).
36. J. C. Van Gemert et al., "Comparing compact codebooks for visual categorization," *Comput. Vision Image Understanding* **114**(4), 450–462 (2010).
37. R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *2013 20th IEEE Int. Conf. on Image Processing (ICIP)*, pp. 4377–4381, IEEE (2013).
38. R. Du et al., "Object categorization based on a supervised mean shift algorithm," in *Workshops and Demonstrations Computer Vision-ECCV 2012*, pp. 611–614, Springer (2012).
39. J. Snchez, F. Perronnin, and T. De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognit. Lett.* **33**(16), 2216–2223 (2012).
40. J. C. van Gemert et al., "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(7), 1271–1283 (2010).
41. L. Xie et al., "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.* **23**, 1994–2008 (2014).
42. J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 461–468, IEEE (2009).
43. B. Fernando et al., "Supervised learning of Gaussian mixture models for visual vocabulary generation," *Pattern Recognit.* **45**(2), 897–907 (2012).
44. R. Sternberg, *Cognitive Psychology*, Wadsworth Publishing, Boston (2011).
45. J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vision* **88**(2), 254–283 (2010).
46. A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 413–420 (2009).
47. S. Gupta et al., "Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation," *Int. J. Comput. Vision* **112**(2), 133–149 (2015).
48. X. Li and Y. Guo, "Multi-level adaptive active learning for scene classification," *Lect. Notes Comput. Sci.* **8695**, 234–249 (2014).
49. M. Zang et al., "A novel topic feature for image scene classification," *Neurocomputing* **148**, 467–476 (2015).
50. Z. Zuo et al., "Learning discriminative and shareable features for scene classification," *Lect. Notes Comput. Sci.* **8689**, 552–568 (2014).

51. J. Luo, A. E. Savakis, and A. Singhal, “A Bayesian network-based framework for semantic image understanding,” *Pattern Recognit.* **38**(6), 919–934 (2005).
52. H. Bannour and C. Hudelot, “Building semantic hierarchies faithful to image semantics,” *Lect. Notes Comput. Sci.* **7131**, 4–15 (2012).
53. M. Marszalek and C. Schmid, “Semantic hierarchies for visual object recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 07)*, pp. 1–7 (2007).
54. G. Griffin and P. Perona, “Learning and using taxonomies for fast visual categorization,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8 (2008).
55. L. Li-Jia et al., “Building and using a semantic visual image hierarchy,” in *2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3336–3343 (2010).
56. L. Wu, S. C. Hoi, and N. Yu, “Semantics-preserving bag-of-words models and applications,” *IEEE Trans. Image Process.* **19**(7), 1908–1920 (2010).
57. L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories,” *Comput. Vision Image Understanding* **106**(1), 59–70 (2007).
58. J. Wang et al., “Locality-constrained linear coding for image classification,” in *2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3367 (2010).
59. M. Everingham et al., “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vision* **88**(2), 303–338 (2010).
60. S. Savarese, J. Winn, and A. Criminisi, “Discriminative object class models of appearance and shape by correlations,” in *2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. 2, pp. 2033–2040, IEEE (2006).
61. A. Bosch, A. Zisserman, and X. Munoz, “Scene classification via plsa,” in *Computer Vision-ECCV 2006*, pp. 517–530, Springer (2006).
62. L. Zhou, Z. Zhou, and D. Hu, “Scene classification using a multiresolution bag-of-features model,” *Pattern Recognit.* **46**(1), 424–433 (2013).
63. L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Vol. 2, pp. 524–531, IEEE (2005).
64. H. Nakayama, T. Harada, and Y. Kuniyoshi, “Global Gaussian approach for scene categorization using information geometry,” in *2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2336–2343, IEEE (2010).
65. A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *Int. J. Comput. Vision* **42**(3), 145–175 (2001).
66. L.-J. Li and L. Fei-Fei, “What, where and who? Classifying events by scene and object recognition,” in *IEEE 11th Int. Conf. on Computer Vision (ICCV 2007)*, pp. 1–8, IEEE (2007).
67. J. Tang et al., “Semantic-gap-oriented active learning for multi-label image annotation semantic-gap-oriented active learning for multi label image annotation,” *IEEE Trans. Image Process.* **21**(4), 2354–2360 (2012).
68. A. Hossein et al., “From generic to specific deep representations for visual recognition,” in *2015 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE (2015).
69. Y. Gong et al., “Multi-scale orderless pooling of deep convolutional activation features,” in *Computer Vision-ECCV 2014*, pp. 392–407, Springer (2014).
70. P. Sermanet, K. Kavukcuoglu, and Y. LeCun, “Eblearn: open-source energy-based learning in C++,” in *21st Int. Conf. on Tools with Artificial Intelligence (ICTAI’09)*, pp. 693–697, IEEE (2009).
71. X. J. Wang et al., “Annotating images by mining image search results,” *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(11), 1919–1932 (2008).
72. D. Tsai et al., “Large-scale image annotation using visual synset,” in *2011 IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 611–618, IEEE (2011).
73. L. Xie et al., “Orientational pyramid matching for recognizing indoor scenes,” in *2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3734–3741, IEEE (2014).

Zhipeng Ye is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology (HIT). He received his master’s degree in computer application technology from Harbin Institute of Technology in 2013. His research interests cover image processing 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 from HIT in 2007. His research interests cover image processing, video processing, pattern recognition, and design of very large scale integration circuit.

Wei Zhao is an associate professor at the School of Computer Science and Technology. She received her doctoral degree in computer application technology from 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, HIT. He received his doctoral degree in computer application technology from HIT in 1995. His research interests cover pattern recognition, aerospace image processing, medical image processing, and machine learning.