Journal of
Electronic Imaging

JElectronicImaging.org

Hierarchy-associated semantic-rule inference framework for classifying indoor scenes

Dan Yu
Peng Liu
Zhipeng Ye
Xianglong Tang
Wei Zhao

SPIE. IS&T imaging.org

Dan Yu, Peng Liu, Zhipeng Ye, Xianglong Tang, Wei Zhao, "Hierarchy-associated semantic-rule inference framework for classifying indoor scenes," *J. Electron. Imaging* **25**(2), 023008 (2016), doi: 10.1117/1.JEI.25.2.023008.

Journal of Electronic Imaging 25(2), 023008 (Mar/Apr 2016)

Hierarchy-associated semantic-rule inference framework for classifying indoor scenes

Dan Yu, Peng Liu, Zhipeng Ye, Xianglong Tang, and Wei Zhao*
Harbin Institute of Technology, School of Computer Science and Technology, 92 West DaZhi Street, Harbin 150001, China

Abstract. Typically, the initial task of classifying indoor scenes is challenging, because the spatial layout and decoration of a scene can vary considerably. Recent efforts at classifying object relationships commonly depend on the results of scene annotation and predefined rules, making classification inflexible. Furthermore, annotation results are easily affected by external factors. Inspired by human cognition, a scene-classification framework was proposed using the empirically based annotation (EBA) and a match-over rule-based (MRB) inference system. The semantic hierarchy of images is exploited by EBA to construct rules empirically for MRB classification. The problem of scene classification is divided into low-level annotation and high-level inference from a macro perspective. Low-level annotation involves detecting the semantic hierarchy and annotating the scene with a deformable-parts model and a bag-of-visual-words model. In high-level inference, hierarchical rules are extracted to train the decision tree for classification. The categories of testing samples are generated from the parts to the whole. Compared with traditional classification strategies, the proposed semantic hierarchy and corresponding rules reduce the effect of a variable background and improve the classification performance. The proposed framework was evaluated on a popular indoor scene dataset, and the experimental results demonstrate its effectiveness. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.2.023008]

Keywords: indoor scene classification; bioinspired framework; semantic hierarchical structure; rule-based inference; bag-of-visual-words.

Paper 15778 received Oct. 28, 2015; accepted for publication Mar. 2, 2016; published online Mar. 28, 2016.

1 Introduction

Alongside 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, retrieval, and classification.[1,2] Manually categorizing these visual data has thus become an almost impossible mission. 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 subtasks, such as depth estimation and object detection and recognition. These subtasks have been studied intensely over the past few decades, and there is still ample room for improvement.[3] 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.[4,5] 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 more confusing, and they are often difficult even for human to classify.

Algorithms for scene classification can be roughly divided into two types: traditional and bioinspired methods.[6] Traditional methods use visual features to classify a scene, and this strategy can be further divided into three strategies. The first strategy is based on low-level features for classification, such as color, texture, and shape.[7] This strategy is effective, provided that there are only a small number of categories. The second strategy is devoted to the development of high-level features from a global perspective. This is done by treating the image as a collection of image blobs, and by introducing more descriptive features for precise scene classification.[5,8,9] This strategy 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.[10,11] In addition to low-level-based models, researchers have applied existing cognitive models such as the human visual system to computer-vision applications to further improve performance, and this have been proved effective.[12–14] This strategy is popular for visual tasks, including field-of-action recognition, image processing, and scene classification.[12,15–22] Last but not least, rule-based systems can also be used to solve the problem of classification.[23]

Previous research in scene classification commonly used the entire image and a predefined knowledge base for classification, restricting performance, and flexibility. Inspired by the human visual system (HVS), the hierarchical structure of scenes, and rule-based inference for determining the category of a scene were investigated and a hierarchy-associated semantic-rule inference (HASRI) framework was proposed in this paper. With the proposed framework, semantic hierarchies are extracted by deformable-parts model[24] and bag-of-visual-words (BoVW)[25] in order to construct the rules used to train a decision tree. This decision tree is responsible for inferring the general category of the scene according to hierarchical semantics of testing samples. Our approach

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

1017-9909/2016/$25.00 © 2016 SPIE and IS&T

is highly modularized and 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 Sec. 2. Then the bioinspired HASRI indoor scene-classification framework is proposed in Sec. 3. Experimental results are provided in Sec. 4. Finally, ongoing and future work is summarized in Sec. 5.

2 Related Work

2.1 *Object Detection and Classification*

Object detection and classification are both active research topics in computer vision. The discriminative part-based model (DPM) proposed by Felzenszwalb et al.[24] is a famous object detection approach that models unknown part positions as latent variables in a support vector machine (SVM) framework. The model contains three parts: histogram of gradients (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 as an 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 proposed the method for building cascade classifiers for the DPM model to significantly improve its detection speed.[26]

Object classification is a rapidly developing subfield of computer vision and machine learning. Existing work can be roughly categorized into two types:[27] low-level visual features and semantics. Low-level features include color, shape, texture, and so on.[28–32] One of the most common and effective methods for object classification using low-level visual features is BoVW[33] model. An image is treated as a collection of unordered appearance descriptors extracted from local patches. These are then quantified with discrete visual words by applying k-means clustering to the local features in order to construct a histogram of the bag-of-feature (BoF) that represents the image. Classifiers are then trained with the BoF for categorization. The combination of a BoVW and a scale-invariant feature transform (SIFT)[34] algorithm is one classic method for classifying images. Methods based on low-level features cannot suitably bridge the semantic gap between low-level features and manually supplied semantic concepts.[35] As a result, semantic approaches were proposed to solve the problem caused by the semantic gap. With this type of approach, semantic layers are constructed to narrow semantic gap and to generate an improved visual vocabulary.[36] In one previous study of Jaimes,[37] descriptive semantics can be classified into several levels: type, global distribution, local structure, global composition, generic objects, generic scene, specific objects, specific scene, abstract objects, and abstract scene. Because it is natural and human-like to represent semantic concepts abstractly,[38] it has become increasingly popular to organize and express semantics in a hierarchical manner.[39–44]

Combining multiple high-level knowledge representations in classification tasks is one approach that is receiving growing interest. However, there are two main drawbacks. The first is that there are no effective universal models for describing the knowledge of a scene with traditional classification tasks. Thus, effectively and uniformly modeling abstract semantics requires further investigation. The second drawback is that, since the knowledge of a scene cannot be directly extracted from images, it is commonly predefined, leading to an inflexible classification process, because knowledge rules cannot be updated during the learning process.

2.2 *Human Visual System*

Roughly speaking, HVS is built upon the combination of receiver (the eyes) and a processor (the brain). The cognitive process of HVS can be shown as shown in Fig. 1, where long-term memory (LTM) function as a huge knowledge warehouse that serializes all kinds of information, and short-term memory (STM) is a much smaller volatile storage space, acting as the initial location when handling short-term knowledge learned from environmental stimulations.[45] The HVS can be roughly divided into three successive stages: encoding, representation, and interpretation.[15] Encoding is the lowest-level stage in vision, and it involves converting light into electric signals. During the second stage, the representation of the encoded image is tuned to the specific characteristics of the visual signal. Enlightened by these biometric

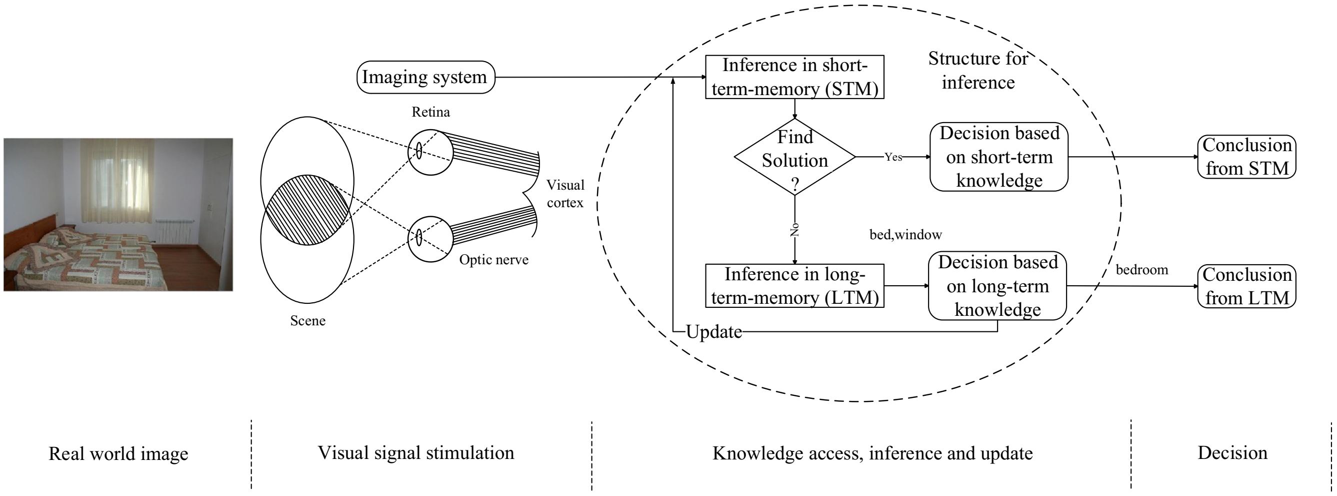

Imaging system
Retina
Visual cortex
Optic nerve
Scene
Inference in short-term-memory (STM)
Structure for inference
Find Solution ?
Yes
Decision based on short-term knowledge
Conclusion from STM
No
bed,window
Inference in long-term-memory (LTM)
Decision based on long-term knowledge
bedroom
Conclusion from LTM
Update
Real world image
Visual signal stimulation
Knowledge access, inference and update
Decision

Fig. 1 The human inferential process.

characteristics, bioinspired visual models have been investigated in computer vision[14,46] over the past few decades. Huang et al.[14] proposed a robust and efficient method for scene classification based on human visual cognition. Rebai[46] designed robotic visual memory and spatial cognition by exploiting characteristics of the human brain.

Typically, when more objects are detected, there is more information that can be provided during the learning process to improve the performance. To understand a scene, information at multiple levels should be integrated, and the interactions between scene elements should be analyzed.[47] However, not all objects can be detected with existing methods, because the quality of images can be affected by several factors, such as the angle and the intensity. This problem is commonly ignored by most previous research. In this paper, semantic hierarchical structure is used to detect more objects in a scene. Different hierarchies are detected, and their semantics are used to construct the decision tree to classify scenes. Here, the hierarchies are used to represent objects located in different positions, and the categories of the objects are called semantics. Combination of different hierarchies forms the hierarchical structure. This is beneficial for scene classification, reducing environmental constrains and simplifying the implementation to improve the performance. These phenomena are shown in detail in Fig. 2.

2.3 *Rule Establishment and Inference*

Rule mining is a field driven by strong interests. The purpose of rule mining is to recognize the strong database rules with different measures of so-called "interestingness." By analyzing symbolic data, comprehensible patterns or models in data are discovered. After decades of study, several mining algorithms have been proposed, and these can be typically categorized as two techniques:[48] predictive and descriptive induction. For predictive induction, models are trained to predict unseen examples. The aim of descriptive induction, by contrast, is to find comprehensible patterns in unlabeled data. These two techniques are commonly investigated by different research communities: the machine-learning community targets predictive induction, and the data-mining community deals with descriptive induction.

With the development of image analysis, there has been an increasing interest in knowledge-based approaches to interpreting and understanding image sequences. To overcome problems such as the semantic gaps, knowledge-based classification methods have been developed with amplified descriptive abilities. Visual knowledge comprises terms that describe labeled instances in scenes, objects, actions, and attributes, along with the contextual relationships between them.[49] Knowledge representation and reasoning belong to

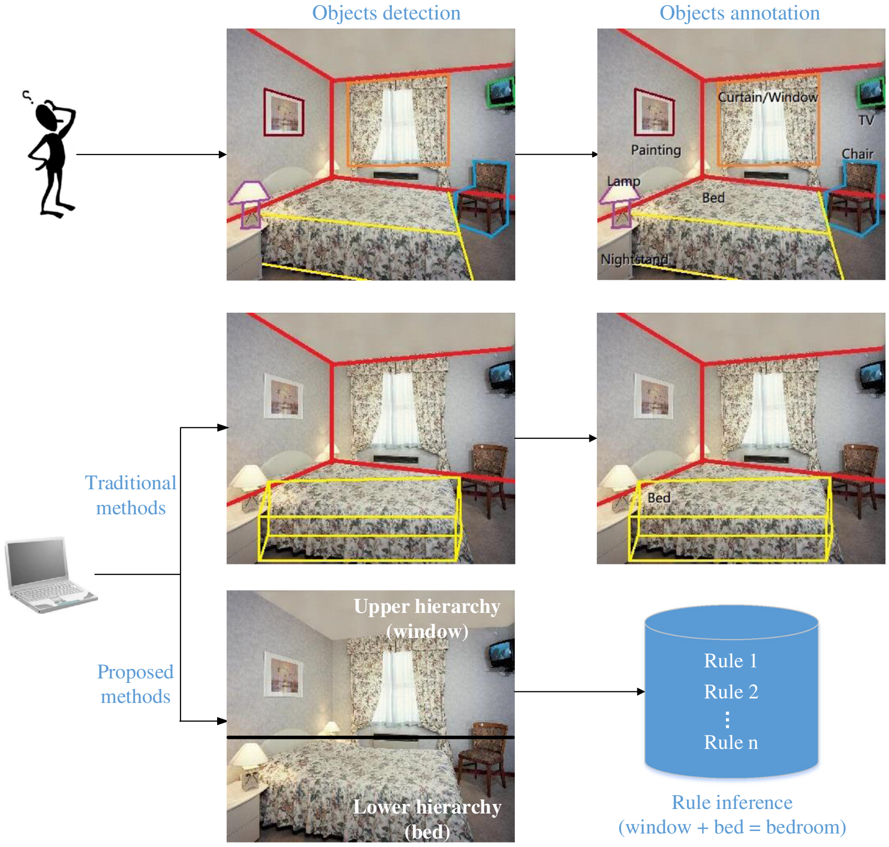


Fig. 2 A simple example of scene annotation achieved by human and computer. The upper and lower results, respectively, represent results of human and computer. It is obvious that one of the natural advantages of humans over computers is the ability of object detection. It is easy for humans to detect and annotate all objects in a scene even from a single image, while due to the environmental factors such as the position of camera, some objects cannot be detected by existing algorithms. A hierarchical structure based on DPM for scene classification was introduced to overcome this disadvantage. Scene classification is achieved by rule inference constructed dynamically from the hierarchical structures of training data.

the field of artificial intelligence and are used to solve complex tasks by analyzing how knowledge can be represented symbolically, and automatically manipulated with reasoning programs. This field is at the very core of a radical idea about how to understand intelligence.[50] For visual problems, expert systems are designed to solve problems by reasoning about knowledge that is represented primarily as if-then rules [51] or first-order production rules.[52] An expert system is usually divided into two subsystems: an inference engine and a knowledge base. The inference engine utilizes known rules generated by extracted features for deducing new facts.[53] Bischof and Caelli[54] first investigated a knowledge-based object-recognition problem using machine learning techniques. Zhou et al.[55] proposed a supervised rule-based video-classification system using low-level features for information browsing and updating. Amato and Di Lecce[56] studied the problem of automatic knowledge generation in a content-based image-retrieval system, for which a knowledge base was constructed with fuzzy clustering algorithms and used to exhibit the organization of an image during the retrieving process. Ang et al.[57] proposed an evolutionary algorithm for extracting rules from a local intensity search scheme that complemented the global search capability of evolutionary algorithms. Xu and Petrou[58] explored a hierarchical knowledge system by designing logical rules for defining an object by answering "why" and "how" questions to decide object characteristics. Chen proposed[49] a system called never ending image learner to automatically extract and enrich visual knowledge bases from the internet without interruption.[59] Porway et al.[60] proposed a knowledge architecture that introduced an illuminative nonrecursive hierarchical grammar tree to predict the categories of objects in an image, significantly reducing the difficulty involved in constructing a knowledge base.

The goal of knowledge-based systems is to amass the information needed for inferences, and these systems benefit significantly from advances in knowledge representation. The main drawback to such systems is that it is difficult to extract and add knowledge to the base. The knowledge base should only contain correct items, and this must be verified, restricting the performance of the overall system.[61] Meanwhile, it is challenging and tedious to construct a knowledge base using existing rules exclusively. Moreover, it is inevitably problematic to construct a dynamic knowledge base that adaptively recognize scenes. In this paper, dynamic rule base is constructed from visual data with a series of preprocessing steps, including detection and annotation. The framework for solving this problem is presented later.

3 Hierarchy-Associated Semantic-Rule Inference Classification Framework

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 considered 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 studies of 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 subscenes are common among images. Indeed, similarity is double-edged. On the one hand, it is important for training classifiers. On the other hand, it risks confusing the classifier and resulting in misclassifications. One reason that the problem of indoor scene classification is challenging is that both similarity and dissimilarity exists among different scenes. Samples of similarities of images are shown in Fig. 3.

Inspired by the human inference process, a hierarchical framework is proposed for scene classification. The process of scene classification is divided into two steps: empirically based annotation (EBA), and match-over rule-based (MRB) inference. The function of EBA is analogous to human vision, which is responsible for detecting and analyzing the hierarchies of indoor scenes. The hierarchical structures are used to train the classifiers. MRB inference is similar to human decision-making when classifying an entire scene according to the annotated results of EBA. The rules are constructed as knowledge in order to determine the category of the scene. In short, EBA corresponds to human vision, and MRB corresponds to decision-making, including both the STM and LTM. Moreover, the environmental context is considered, encapsulating rich information about how natural scenes and objects are related to each other. The co-occurrence of objects within a scene is also considered, facilitating coherence to an interpretation of a scene.[62] The proposed HASRI indoor scene-classification framework is shown in Fig. 4. Details of the hierarchical structure of indoor scenes are provided in the following subsection. Here, hierarchies and their corresponding semantics are explicit knowledge that can help to reduce uncertainty for classification.[41] When talked about hierarchical semantics, it is the category of each hierarchy assigned by categorization method we refer to. For semantics hierarchies, the main body is the hierarchies with corresponding categories. The term, HASRI, indicates that images are described by detected hierarchies. The categories

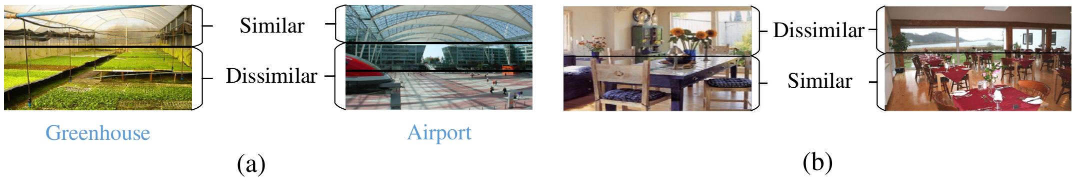


Fig. 3 Samples of indoor scenes with similar and dissimilar hierarchies. We can see that due to the context of indoor scenes, despite the number of hierarchies between different images, similar areas are on the same hierarchies. Thus, although there exist similarities between different images, it is still able to distinguish them by the areas that are not similar. Thus the similarities between different scenes are related to corresponding hierarchies.

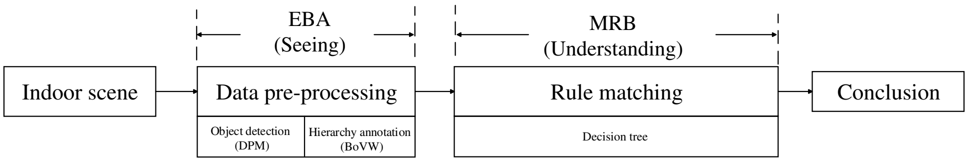
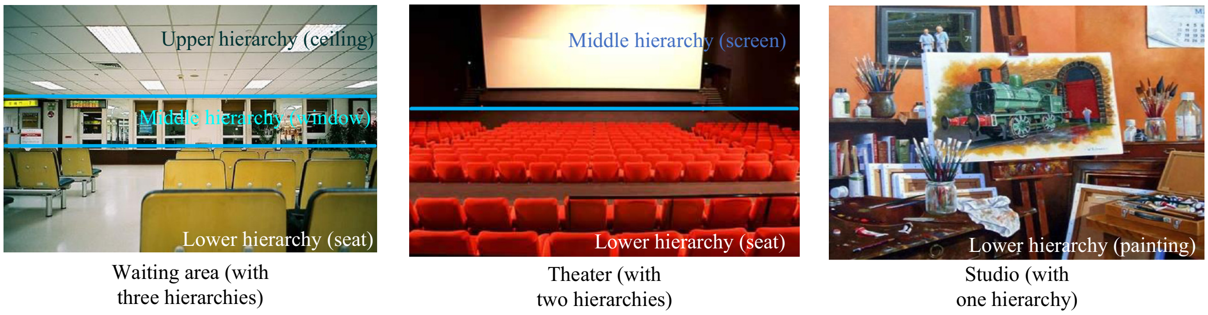


Fig. 4 The framework of HASRI for indoor scene classification.


Upper hierarchy (ceiling)
Middle hierarchy (window)
Lower hierarchy (seat)
Middle hierarchy (screen)
Lower hierarchy (seat)
Lower hierarchy (painting)

Waiting area (with three hierarchies)

Theater (with two hierarchies)

Studio (with one hierarchy)

Fig. 5 Sample of indoor scenes with different hierarchical structures.

of the images are inferred by rules, which is constructed from the hierarchical semantics.

3.1 *Constructing the Hierarchical Structure of an Indoor Scene*

Before classifying an indoor scene, samples are preprocessed, and the hierarchies are annotated for the subsequent learning process. Hierarchical methods can be roughly divided into two types: building hierarchical semantics,[36,39–41,63] and developing hierarchical models.[33,60,64] 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 abstract and common hierarchy of a scene is seldom explored, limiting the ability to further improve the accuracy. Thus, semantic information from the hierarchical structure is introduced to the learning process in order to improve the performance of scene classification. The scene is inferred by combining the hierarchical semantics of different levels. By exploiting the structures of semantic hierarchies, rules can be flexibly established. Three hierarchies were defined for an indoor scene according to the context of the image: the upper, middle, and lower hierarchies. These hierarchies represent their corresponding objects: the ceiling, wall, and floor. There is an obvious distinction between these three hierarchies. Each hierarchy chiefly contains a single dominant semantics that constitutes the rules for training the classifier for indoor scenes. This observation is beneficial for rule construction. Samples of the hierarchical structure of indoor scenes with different number of hierarchies are shown in Fig. 5. For each image, the number of hierarchies is determined by the detection results. The motivation for proposing this hierarchical structure 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 corresponding hierarchies caused by similarity, two images can be distinguished insofar as their hierarchy is different.

The process of automatically hierarchy detection is shown in Fig. 6 by using the discriminative-parts model (DPM).[24] The detected blocks are extended according to their spatial context in the image in order to acquire the hierarchical structures. First, the objects marked by yellow boxes are detected. Here, the location relationships of detected areas are considered as the spatial context.[62] All boxes will extend to both left and right direction as best as possible. The extended areas are marked by black dashed boxes. The remaining areas are treated as corresponding hierarchies. The final hierarchical structure is shown in an image with solid colors. The number of hierarchies varies according to the results of

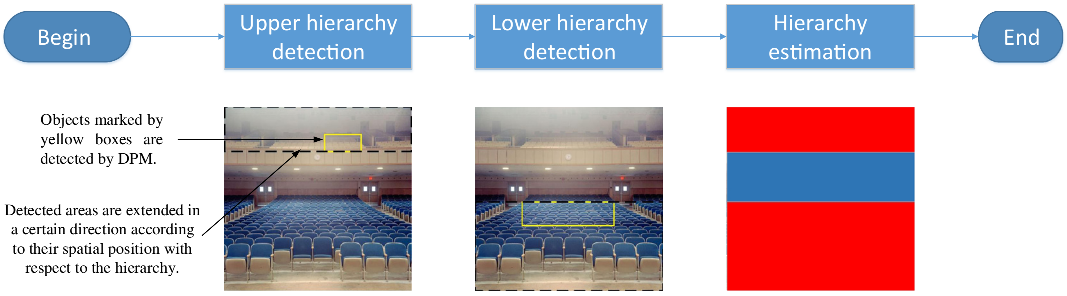

Begin
Upper hierarchy detection
Lower hierarchy detection
Hierarchy estimation
End
Objects marked by yellow boxes are detected by DPM.
Detected areas are extended in a certain direction according to their spatial position with respect to the hierarchy.

Fig. 6 Sample of hierarchical structure of an indoor scene.

object detection. After area extension, relative location information can be included in the corresponding hierarchies, improving the quality of hierarchies and simplifying the whole process.

3.2 *Rule Construction*

After annotating all of the hierarchies, the rules for inferring the category of an indoor scene are generated. This process will be described in the following subsections.

3.2.1 *Mining for knowledge-rules*

With the explosively growth of information, the scale of databases has growing rapidly. During the past two decades, research on knowledge mining has progressed from the problem of how to extract valuable information from large-scale datasets to the introduction of machine learning theories, expert systems, pattern recognition, and so on. Among this research, one of the most successful methods for rule learning is a decision tree. Decision trees construct rules for databases using mining techniques.[65,66] For our work, C4.5[67] was chosen as the decision-tree learning algorithm. In Fig. 5, inferential rules can be generated by combining the semantic information provided by their hierarchical structures as follows:

$$\begin{aligned}&\text{IF}\,(\text{upper Hierarchy}=\text{light})\vee(\text{middle Hierarchy}\\&\qquad\qquad=\text{window})\vee(\text{lower Hierarchy}=\text{seat})\\&\quad\text{THEN Category}=\text{airport}\\&\text{IF}\,(\text{upper Hierarchy}=\text{light})\vee(\text{middle Hierarchy}\\&\qquad\qquad=\text{board})\vee(\text{lower Hierarchy}=\text{computer})\\&\quad\text{THEN Category}=\text{computer\_room}.\end{aligned}\tag{1}$$

Missing hierarchies (not arbitrary values) are represented by a question mark (“?”). For example, if there are only upper and lower hierarchies for the categories “window” and “seat,” respectively, in an “airport,” then the rule can be described as follows:

$$\begin{aligned}&\text{IF}(\text{upper Hierarchy}=\text{window})\vee\,?\\&\ \vee(\text{low Hierarchy}=\text{seat})\\&\quad\text{THEN Category}=\text{airport}.\end{aligned}\tag{2}$$

The algorithm for generating a decision tree for indoor scene classification is provided in Algorithm 1. There are two advantages to our method that result in an improved classification performance compared to previous works that use a spatial pyramid[8] for the local geometric correspondence of subregions. First, images are marked as different hierarchies according to the features of indoor scenes in order to improve the quality of the visual words for better annotation results. Second, a decision tree trained with semantic hierarchies was used to combine local category information for classification. Here, the local category information is the semantics of hierarchies, i.e., the annotated categories of hierarchies, while the relative location information is the spatial location of detected objects. Take Fig. 5 as example. For the waiting area, there are three semantic hierarchies, upper, middle, and lower hierarchies. The upper hierarchy is assigned a local category information, ceiling.

3.2.2 *Updating rules*

Our proposed framework involves constructing a dynamic set of rules during the learning stage, rather than relying on a predefined set of rules. Figure 7 shows this process. For each training image in the leftmost column of Fig. 7, the hierarchies—marked by solid and dashed lines—are first constructed by detecting and annotating images with the DPM and the BoVW model, respectively. After the detection process, the corresponding rules with hierarchical semantics

Algorithm 1 The decision tree learning algorithm $DTGen(D)$.

```
Require: Rule set $D$.
Ensure: Decision tree $T$.
1:  Build attribute set from $D$ by the local and overall categories of different hierarchies and samples.   ▷ Preprocessing of samples
2:  Train the classifiers of the annotation module with $D$.
3:  for each attribute $a_i \in A$ do                                                                           ▷ Decision tree construction
4:      Compute information gain of $a_i$.
5:  end for
6:  Select the $a_{\text{best}}$ from $A$. Create a decision node $N_{\text{best}}$ in the root to test $a_{\text{best}}$ in 3.
7:  Iteratively construct the child tree $T_c$ under $N_{\text{best}}$ by subdatasets from $D$ based on $a_{\text{best}}$.
8:  Attach $T_c$ to the corresponding branch of $T$.
9: $A = A \setminus a_{\text{best}}$
10:  If $A = \phi$, then Return $T$. Otherwise goto 6.
```

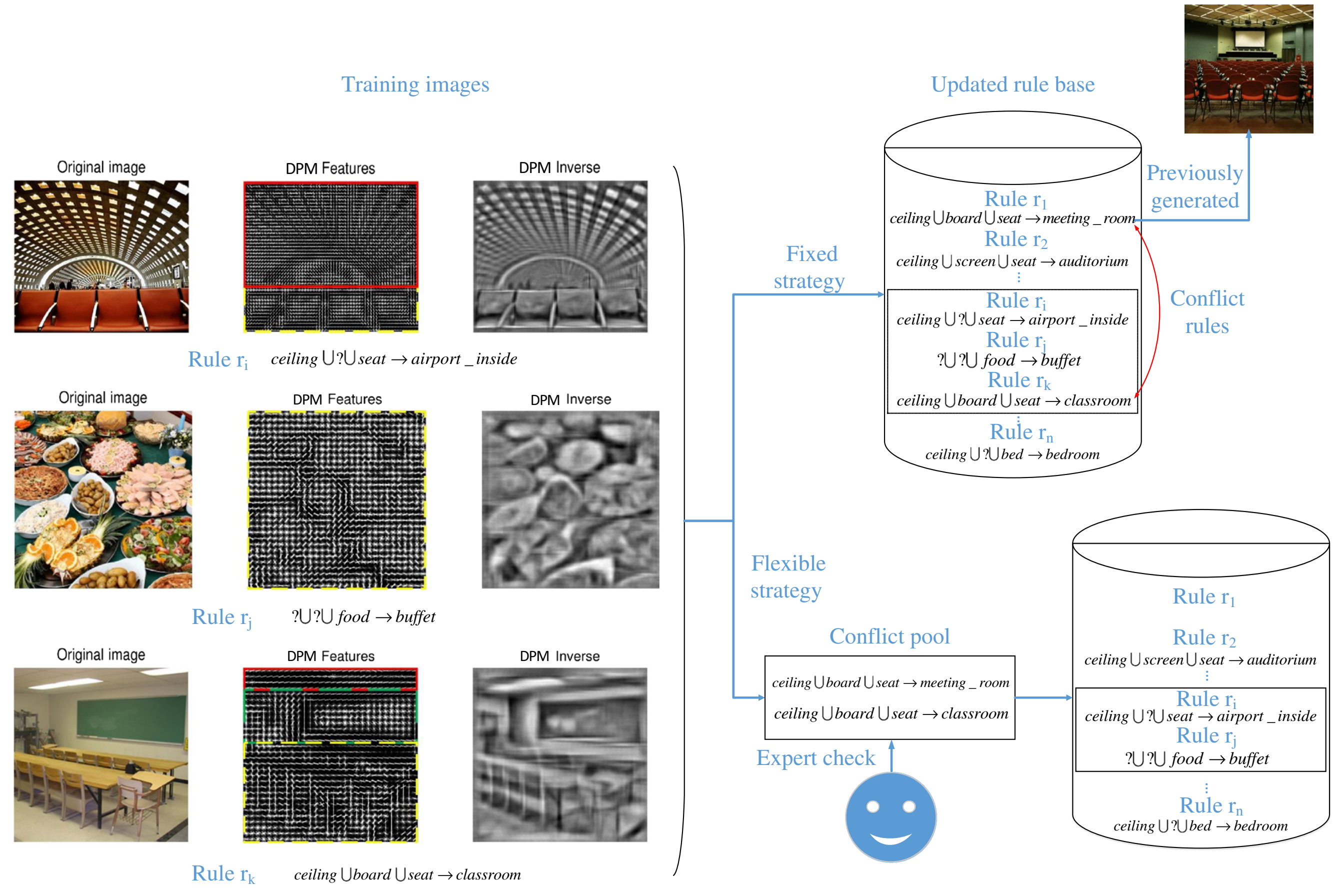

Training images
Original image
DPM Features
DPM Inverse
Rule r_i $ceiling \bigcup ? \bigcup seat \rightarrow airport\_inside$
Original image
DPM Features
DPM Inverse
Rule r_j $? \bigcup ? \bigcup food \rightarrow buffet$
Original image
DPM Features
DPM Inverse
Rule r_k $ceiling \bigcup board \bigcup seat \rightarrow classroom$
Fixed strategy
Updated rule base
Rule r_1
$ceiling \bigcup board \bigcup seat \rightarrow meeting\_room$
Rule r_2
$ceiling \bigcup screen \bigcup seat \rightarrow auditorium$
Rule r_i
$ceiling \bigcup ? \bigcup seat \rightarrow airport\_inside$
Rule r_j
$? \bigcup ? \bigcup food \rightarrow buffet$
Rule r_k
$ceiling \bigcup board \bigcup seat \rightarrow classroom$
Rule r_n
$ceiling \bigcup ? \bigcup bed \rightarrow bedroom$
Previously generated
Conflict rules
Flexible strategy
Conflict pool
$ceiling \bigcup board \bigcup seat \rightarrow meeting\_room$
$ceiling \bigcup board \bigcup seat \rightarrow classroom$
Expert check
Rule r_1
Rule r_2
$ceiling \bigcup screen \bigcup seat \rightarrow auditorium$
Rule r_i
$ceiling \bigcup ? \bigcup seat \rightarrow airport\_inside$
Rule r_j
$? \bigcup ? \bigcup food \rightarrow buffet$
Rule r_n
$ceiling \bigcup ? \bigcup bed \rightarrow bedroom$

Fig. 7 Example of detection results of DPM and the rule base update processes. Hierarchies are marked by different line styles: red solid line for upper hierarchy, green long dashed line for middle hierarchy, and yellow short dashed line for lower hierarchy.

are constructed from training images. Each newly generated rule is verified in order to prevent repeated and conflict rules. Repeated rules are simply discarded, and conflict rules are withdrawn and moved to a pool. The conflict rule set is empty before the training process. Then the verification set is used to gradually rich it. Finally, the rules that pass this verification process are added to the set of rules. The learning process is dynamic, because the updating process occurs during the learning step. There are two reasons for generating conflict rules. First, the conflict rules occur when the preprocessing module does not output the correct results. Thus, the human expert is asked for verification. Second, conflict rules arise when there is significant similarity between some indoor scenes that are difficult for even a human to distinguish. Thus, for conflict rules generated as a result of the first reason, the human annotator will correct the annotation and send the sample for retraining the corresponding module in EBA. For conflicts arising as a result of the second situation, the rules are maintained, and a category is randomly selected during the testing stage. The process mentioned above is summarized in Algorithm 2.

3.3 *Proposed Framework for Indoor Scene Classification*

Inspired by the human cognitive process, the proposed framework includes modules for EBA and MRB inference. The EBA's learning process involves annotating the semantic hierarchy and constructing the rules. The proposed framework differs from previous work[68] insofar as it exploits segmented hierarchical structures and a implementation of the classifier. Our method introduces a hierarchical structure for describing the scene according to the environmental context,

Algorithm 2 The $RCos(R)$ algorithm for rule base construction.

```
Require: Generated rule set $R$.
Ensure: Constructed rule base $R_b$, conflict rule set $C$.
for Every rule $r_i$ in $R$ do
    Check $r_i$ with every rules in $R$.
    if is Conflict ($r_i$, R) then
        Move ($r_i$, C)
    else
        Move ($r_i$, $R_b$)
    end if
end for
```

Algorithm 3 The whole learning process.

```
Require: Training data $D_t$, verification data $V$, testing data $D_m$.
Ensure: Constructed rule base $R_b$ and decision tree $T$.
Detect the hierarchical structure with DPM for $D_t$.            ▷ Preprocessing stage
Train the BoVW based classifier $C$ by $D_t$ for object annotation.
Annotate $V$ with $C$.
Extract and construct rule set $R$ from $V$.                      ▷ Postprocessing stage
$R_b = RCos(R)$.
$T = DTGen(R_b)$.
```

and it generates a set of inference rules for classification. With MRB inference, the test image is first passed to the annotation module to acquire information regarding the hierarchies. Then the overall category is inferred from the learned rules, indicating the fact that the category of an indoor scene is determined by what objects it contains. Training a decision tree with a dynamically constructed set of rules for inference allows it to adapt to variations in a scene. The knowledge base that is generated can be extended to different situations, because rules can be added or replaced for each corresponding application. Moreover, there is no need to modify the object-detection module. When advanced object-detection methods are utilized, existing knowledge is still useful for classification. The modularized design makes the overall system easy to maintain and saves additional costs. The whole framework is grammatically summarized in Eq. (3).

$$sc\_cl \rightarrow EBA \cup MRB \quad EBA \rightarrow obj\_an \quad MRB \rightarrow kb\_inf$$
$$obj\_an \rightarrow obj\_det \cup obj\_cl \quad kb\_inf \rightarrow conclusion. \tag{3}$$

Here sc_cl refers to the scene classification, obj_an refers to object annotation, $kb\_\inf$ denotes rule-based inferences, obj_det and obj_cl refer to object detection and classification, respectively. In EBA, the object-annotation process includes both object detection and classification, based on DPM and BoVW, respectively. At this stage, semantic hierarchies are constructed for the visual data. In MRB, the inference process is implemented with a decision tree to derive the category. Compared with previous works, there are several advantages in the proposed framework. First, introducing a hierarchical structure decreases the influence of varying backgrounds and highlights the differences between each category. Unlike traditional methods, such as the BoVW that simply extracts the features from the entire image, the proposed framework introduces semantic hierarchies to first extract the visual features from a locality; it then merges local information with a global category to improve the flexibility and performance of the classification. Second, compared with low-level features, constructing rules from the hierarchical semantics improves the ability of expression. Multiple categories can be subjected to the same set of rules on a reduced scale, providing flexible, human-like knowledge for classification. The entire process is summarized in Algorithm 3.

4 Experimental Results

In this section, the overall performance of the proposed HASRI framework was evaluated on Massachusetts Institute of Technology (MIT's) indoor scene-recognition database. Tests were divided into two parts containing vertical and horizontal comparison to demonstrate the effectiveness of our work. First, we focus on the performance of the HASRI with different module settings. Then, HASRI was compared with other relevant methods for a horizontal comparison. Corresponding results are provided in the following subsections.

4.1 *Experimental Settings*

4.1.1 *Dataset*

MIT's indoor scene-recognition dataset[69] 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 other datasets such as Caltech-101, the distance between different categories is not significant. With MIT's dataset, it is sometimes confusing even for humans to distinguish between pairs of samples. The dataset was divided into three subsets: a training subset, a verification, and a testing subsets. Modules in EBA were trained using the training dataset, the rules for MRB inference were constructed using the verification dataset, and the evaluation was conducted using the testing dataset. The training images comprised 10% of the dataset. All categories were utilized, and a one-versus-all strategy was used to train the SVM classifiers. Gaussian kernel is used for SVM classifiers.

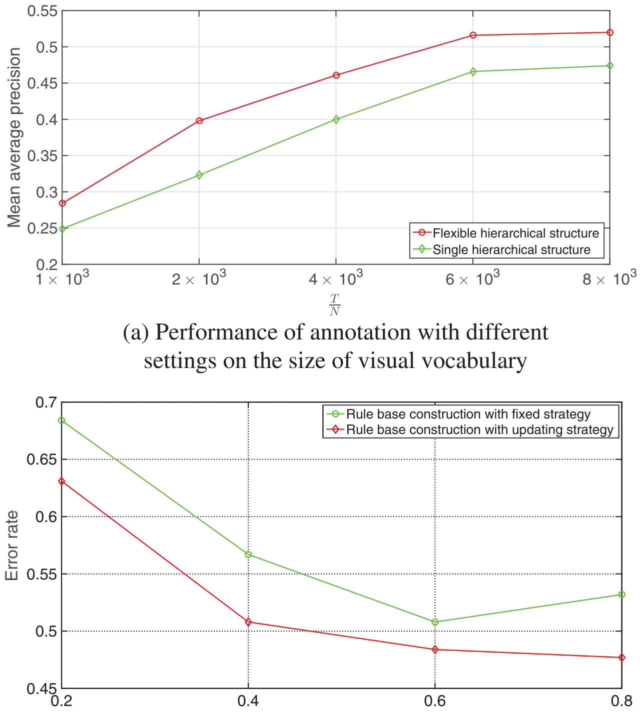
4.1.2 *Experimental settings*

In this paper, DPM[24] was used to extract the hierarchies of images, which can be understood as an extension of the

HoG, because it contains trained detectors to speed up the process. The method implemented by Vondrick[70] was used to visualize the results. A BoVW[25] was used for annotation. The mean-average precision (MAP) was used as a metric to evaluate the performance of the approaches. To provide a fair comparison of the difference between traditional and flexible rule-construction strategies, a fixed set of rules was constructed such that it can accept new rules without additional checks during the training stage. By contrast, flexible rules could be edited, and conflict rules could be addressed. The Waikato Environment for Knowledge Analysis (WEKA)[71] was used to implement the decision tree. We manually select the training dataset, then randomly and equally divided the remaining samples into verification and testing dataset. For testing samples that matched multiple rules in different categories, a category was randomly selected during the inferential process. Two-fold cross validation was used to compare our proposal with other methods, and the mean precision was reported.

4.2 *Vertical Experimental Results*

In the vertical experiment, different settings were used to evaluate the performance of the proposed framework. First, the results of classification with and without hierarchical structures were shown, and then the performances were discussed after constructing both fixed and flexible rules. In this paper, we referred "rule" as the knowledge terms generated from visual data to construct the knowledge base. As shown in Fig. 7, the word "strategy" is used to describe how we update the knowledge base.



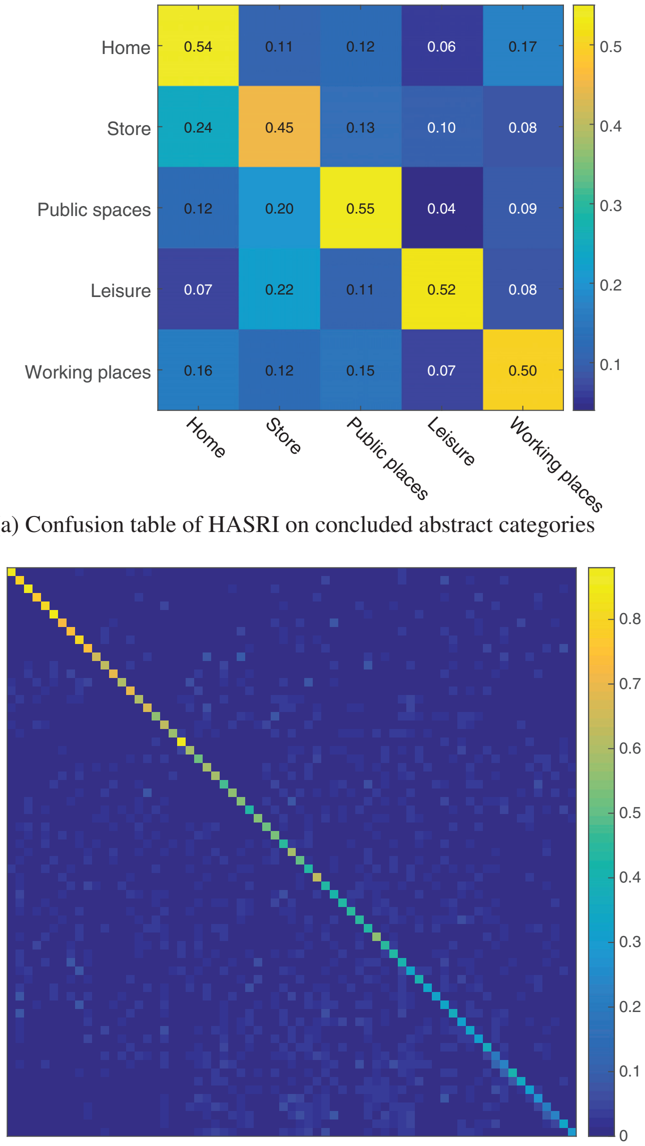
(a) Performance of annotation with different settings on the size of visual vocabulary

(b) Error rate of rule base with different number of verifying samples

Fig. 8 Self-evaluation of the proposed framework. (a) Performance of annotation with different settings on the size of visual vocabulary and (b) error rate of rule base with different number of verifying samples.

Figure 8(a) shows the annotation performance achieved by different hierarchical structures. For a single hierarchical structure, the model's performance degraded to that of a common annotation method. The results show an obvious improvement in the annotation performance with the proposed hierarchical structure. This is due to the fact that indoor scenes are typically complex. In such scenes, it is common to find multiple objects in the same scene, and annotating 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 annotation.

The performance of the proposed framework using different rule-construction strategies is shown in Fig. 8(b). The horizontal axis shows the ratio of the sizes of the verification samples to the testing samples. The purpose of this



(a) Confusion table of HASRI on concluded abstract categories

(b) Confusion table of HASRI on all categories

Fig. 9 Results of vertical tests of HASRI. (a) Confusion table of HASRI on concluded abstract categories and (b) confusion table of HASRI on all categories.

Table 1 Detailed results measured by average precision of the proposed HASRI framework.

Categories	Single hierarchy + flexible strategy	Multiple hierarchy + fixed strategy	Multiple hierarchy + flexible strategy
Church inside	0.677	0.726	0.819
Elevator	0.676	0.652	0.795
Auditorium	0.649	0.651	0.769
Buffet	0.645	0.634	0.748
Classroom	0.640	0.629	0.730
Greenhouse	0.618	0.628	0.728
Bowling	0.617	0.625	0.720
Concert hall	0.606	0.625	0.671
Computer room	0.604	0.607	0.671
Dental office	0.601	0.600	0.659
Library	0.594	0.600	0.649
Inside bus	0.583	0.600	0.644
Closet	0.581	0.595	0.637
Corridor	0.581	0.593	0.630
Grocery store	0.575	0.593	0.629
Locker room	0.571	0.590	0.625
Florist	0.563	0.584	0.616
Studio music	0.561	0.574	0.612
Hospital room	0.548	0.573	0.608
Nursery	0.546	0.570	0.586
Bathroom	0.544	0.560	0.574
Laundromat	0.536	0.549	0.569
Stairs case	0.535	0.548	0.567
Garage	0.532	0.545	0.563
Gym	0.522	0.533	0.559
TV studio	0.520	0.533	0.556
Video store	0.506	0.531	0.548
Game room	0.502	0.528	0.547
Pantry	0.496	0.526	0.543
Pool inside	0.496	0.524	0.541
Inside subway	0.485	0.523	0.537
Kitchen	0.478	0.520	0.536
Wine cellar	0.478	0.515	0.532
Fastfood restaurant	0.474	0.514	0.531
Bar	0.472	0.507	0.530
Clothing store	0.471	0.502	0.527
Casino	0.471	0.499	0.525
Deli	0.460	0.481	0.523
Bakery	0.444	0.480	0.522
Waiting room	0.431	0.479	0.497
Dining room	0.430	0.469	0.497
Bookstore	0.421	0.466	0.495
Living room	0.416	0.465	0.490
Movie theater	0.414	0.458	0.489
Bedroom	0.410	0.456	0.485
Toy store	0.406	0.451	0.483
Operating room	0.406	0.437	0.477
Airport inside	0.405	0.433	0.476
Art studio	0.405	0.431	0.474
Lobby	0.399	0.427	0.463
Prison cell	0.385	0.425	0.457
Train station	0.382	0.406	0.437
Hair salon	0.378	0.402	0.432
Subway	0.376	0.386	0.431
Warehouse	0.374	0.381	0.429
Meeting room	0.372	0.379	0.425
Children room	0.369	0.379	0.423
Shoe shop	0.358	0.374	0.418
Kindergarden	0.353	0.361	0.417
Restaurant	0.352	0.358	0.411
Museum	0.343	0.328	0.397
Restaurant kitchen	0.325	0.315	0.392
Jewelry shop	0.311	0.315	0.370
Laboratory wet	0.282	0.303	0.357
Mall	0.273	0.287	0.333
Office	0.268	0.231	0.306
Cloister	0.234	0.226	0.291

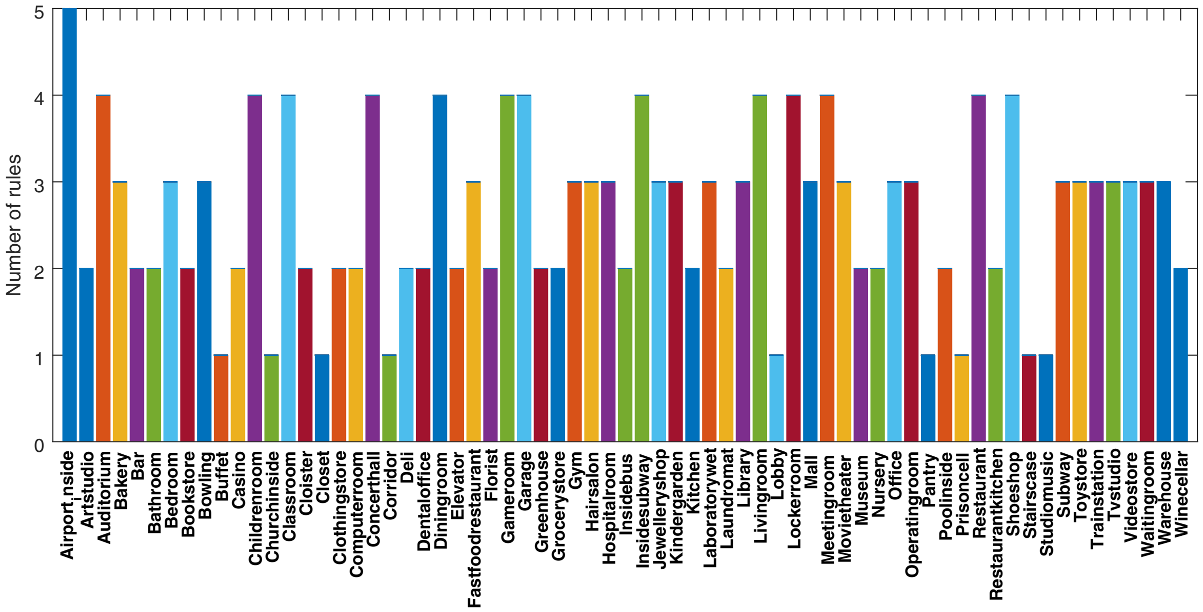


Fig. 10 Statistical results on the number of rules for MIT indoor dataset. The percentage of categories containing different number of rules are respectively 1.49% (five rules), 19.40% (four rules), 32.84% (three rules), 32.84% (two rules), 13.43% (one rule).

Table 2 Rules constructed from the MIT dataset for category inference.

Category	Rules
Airport	Ceiling ∪? ∪ floor
	Ceiling ∪? ∪ people
	Ceiling ∪ window ∪ seat
	? ∪ window ∪ floor
	Ceiling ∪? ∪ seat
Artstudio	? ∪ painting ∪?
	?∪ people ∪ painting
	? ∪ window ∪ drawing
Bar	Ceiling ∪ wine_counter ∪?
	? ∪ wine_counter ∪?
Bakery	Ceiling ∪ bread ∪?
	? ∪ bread ∪ people
	Ceiling ∪ bread∪ shelf
Auditorium	Ceiling ∪ seat ∪ stage
	Ceiling ∪? ∪ seat
	?∪ stage ∪?
	Ceiling ∪ screen ∪ seat

Table 2 (*Continued*).

Category	Rules
Bathroom	Ceiling ∪ sanitary_ware ∪?
	? ∪ sanitary_ware ∪?
Bedroom	Ceiling ∪ bed ∪ painting
	Ceiling ∪? ∪bed
	Ceiling ∪ bed ∪?
Bookstore	? ∪ bookshelf ∪?
	? ∪ people ∪?
Bowling	Ceiling ∪? ∪ bowling_road
	? ∪ people ∪ bowling_road
	? ∪? ∪ bowling_road
Buffet	? ∪ food ∪?
Casino	Ceiling ∪ gamble_facilities ∪?
	? ∪? ∪ gamble_facilities
Children_room	Ceiling ∪? ∪ bed
	? ∪ bed ∪?
	? ∪ toy ∪?
	? ∪? ∪ people

Table 2 (*Continued*).

Category	Rules
Church_inside	Gothic_ceiling ∪? ∪ seat
Classroom	Ceiling ∪ board ∪ desk
	? ∪ board ∪ desk
	? ∪? ∪ desk
	Ceiling ∪? ∪ desk
Closet	? ∪? ∪ closet
Cloister	Ceiling ∪? ∪ floor
	? ∪ pillar ∪?
Clothingstore	Ceiling ∪ clothes ∪?
	? ∪? ∪ clothes
Computerroom	Ceiling ∪ computer ∪?
	Ceiling ∪ board ∪ computer
Concert_hall	Ceiling ∪ platform ∪ seat
	Ceiling ∪? ∪ stage
	? ∪ stage ∪ seat
	Ceiling ∪? ∪ seat
Corridor	? ∪ corridor ∪?
Deli	Ceiling ∪? ∪ goods
	? ∪ people ∪ goods
Dental_office	? ∪ instrument ∪?
	? ∪ people ∪ instrument
Dining_room	Ceiling ∪? ∪ dining_table
	? ∪? ∪ dining_table
	Ceiling_lamp ∪ window ∪ dining_table
	? ∪ window ∪ dining_table
Elevator	? ∪ elevator ∪?
	Elevator ∪? ∪ people
Fastfood_restaurant	Ceiling ∪ counter ∪ table
	? ∪ counter ∪?
	? ∪ counter ∪ people
Florist	? ∪ flower ∪?
	? ∪ people ∪ flower

Table 2 (*Continued*).

Category	Rules
Gameroom	Ceiling ∪ billiard_table ∪?
	? ∪ billiard_table ∪?
	? ∪? ∪ billiard_table
	? ∪ people ∪ billiard_table
Garage	Ceiling ∪ vehicle ∪?
	? ∪ vehicle ∪?
	Ceiling ∪ tools ∪?
	? ∪ tools ∪?
Greenhouse	Ceiling ∪ plants ∪?
	? ∪ plants ∪?
Grocerystore	Ceiling ∪ goods ∪?
	? ∪ goods ∪?
Gym	Ceiling ∪ fitness_facilities ∪?
	? ∪ fitness_facilities ∪?
	? ∪ human ∪ fitness_facilities
Inside_bus	Bus_ceiling ∪ seat ∪?
	? ∪ people ∪ seat
Hairsalon	Ceiling ∪ hairdressing_facilities ∪?
	? ∪ hairdressing_facilities ∪?
	? ∪? ∪ hairdressing_facilities
Hospital_room	? ∪ hospital_bed ∪?
	Ceiling ∪ hospital_bed ∪?
	? ∪ patient ∪ hospital_bed
Inside_subway	Carriage_ceiling ∪ carriage_seat ∪?
	Carriage_ceiling ∪ people ∪?
	? ∪? ∪ carriage_seat
	Carriage_ceiling ∪? ∪ people
Jewelryshop	Ceiling ∪ Jewelry_counter ∪?
	? ∪ Jewelry ∪?
	? ∪ Jewelry_counter ∪?
Kitchen	Ceiling ∪ cupboard ∪?
	? ∪? ∪ operating_desk

Table 2 (*Continued*).

Category	Rules
Kindergarten	Ceiling ∪ toy ∪?
	Ceiling ∪ table ∪?
	? ∪ toy ∪?
Laboratorywet	? ∪ reagent ∪?
	Ceiling ∪ reagent ∪?
	? ∪ equipment ∪?
Laundromat	Ceiling ∪ washing_machine ∪?
	? ∪ washing_machine ∪?
Library	? ∪ bookshelf ∪?
	Ceiling ∪ bookshelf ∪?
	? ∪ bookshelf ∪ table
Living_room	? ∪ sofa ∪?
	? ∪ window ∪ sofa
	Ceiling ∪ window ∪ sofa
	Ceiling ∪? ∪ sofa
Lobby	Ceiling ∪? ∪ floor
Locker_room	Ceiling ∪ locker
	? ∪ locker ∪?
	? ∪ locker ∪ bench
	? ∪ rack ∪?
Mall	Ceiling ∪ corridor ∪?
	? ∪ corridor ∪?
	Ceiling ∪ escalator ∪?
Meeting_room	Ceiling ∪ screen ∪ table
	? ∪ table ∪?
	Screen ∪? ∪ table
	? ∪ window ∪ table
Pantry	Food_shelf
Movietheater	? ∪ screen ∪ seat
	Ceiling ∪ screen ∪ seat
	Ceiling ∪? ∪ seat

Table 2 (*Continued*).

Category	Rules
Museum	Ceiling ∪? ∪ floor
	? ∪ painting ∪?
Nursery	? ∪ crib ∪?
	Ceiling ∪ crib ∪?
Office	Ceiling ∪ computer ∪ office_table
	? ∪ shelf ∪ table
	? ∪ board ∪ table
Operating_room	Ceiling ∪ operation_table ∪?
	Doctor ∪ operation_table
	? ∪? ∪ operation_table
Poolinside	Ceiling ∪ pool ∪?
	? ∪? ∪ pool
Prisoncell	? ∪ prisoncell ∪?
Restaurant	Ceiling ∪ window ∪ table
	? ∪ painting ∪ table
	Ceiling ∪ people ∪?
	Ceiling ∪ painting ∪ table
Restaurant_kitchen	Ceiling ∪? ∪ kitchen_facility
	? ∪ human ∪ food
Shoeshop	Ceiling ∪ shoe_shelf ∪?
	? ∪ shoe_shelf ∪?
	? ∪ human ∪ shoe_shelf
	Ceiling ∪ human ∪ shoe_shelf
Staircase	Ceiling ∪? ∪ stairs
Studiomusic	Ceiling ∪? ∪ instrument
Subway	Ceiling ∪ train ∪?
	Ceiling ∪? ∪ track
	? ∪? ∪ train
Toystore	Ceiling ∪? ∪ toy_shelf
	? ∪? ∪ toy_shelf
	? ∪ human ∪ toy

Table 2 (*Continued*).

Category	Rules
Trainstation	Ceiling ∪ train ∪ platform
	Ceiling ∪? ∪ human
	Ceiling ∪? ∪ train
Tv_studio	Ceiling ∪ human ∪?
	Ceiling ∪? ∪ instrument
	? ∪ human ∪ instrument
Videostore	Ceiling ∪? ∪ disk
	? ∪ human ∪ disk
	Ceiling ∪ human ∪ disk
Waitingroom	Ceiling ∪ window ∪ seat
	? ∪? ∪ seat
	Ceiling ∪ human ∪?
Warehouse	Ceiling ∪? ∪ cargo
	Ceiling ∪? ∪ shelf
	? ∪? ∪ shelf
Winecellar	Ceiling ∪? ∪ barrel
	? ∪ barrel ∪?

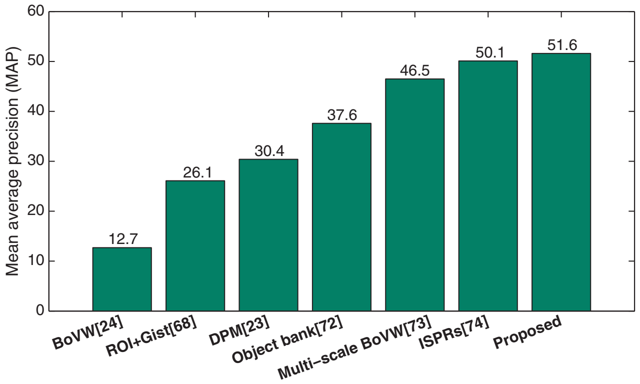


Fig. 11 Comparison results of corresponding methods on MIT-Indoor dataset. Performance is evaluated by MAP. Here SIFT was chosen for BoVW,[25] SIFT, and Gist were chosen for the prototype model,[68] HoG was chosen for DPM,[24] HoG, texture, location, and geometry for Object bank,[72] SIFT for multiscale BoVW,[73] HoG for ISPRs.[74] Discussion on the results was listed in Sec. 4.3.

Table 3 Comparative results of all methods with different features.

Methods	MAP
BoVW (HoG)	10.4
BoVW (SIFT)[25]	12.7
Prototype (HoG + Gist)	22.4
Prototype (SIFT + Gist)[68]	26.1
DPM (HoG)[24]	30.4
Object bank (HoG)[72]	37.6
Multiscale BoVW (HoG)	43.7
Multiscale BoVW (SIFT)[73]	46.5
ISPRs (HoG)[74]	50.1
Proposed	**51.6**

experiment is to show the relationship between the error rate and the number of training samples. With more training data, there is a higher probability of introducing conflict rules that affect the classification performance. We can see from these results that, compared with the traditional strategy, using a flexible rule-updating technique was beneficial for classification. With fixed rules, the error rate increased when the ratio reached 0.8, whereas by adopting a flexible strategy, the error rate continued to decrease as more samples were added. This is because the flexible strategy for constructing rules is more effective at reducing conflict rules than the fixed strategy. The confusion tables for the proposed HASRI with a hierarchical structure and a flexible strategy for categories of different abstract semantics are provided in Figs. 9(a) and 9(b). Here the size of the visual vocabulary was set to $k = 6 \times 10^3$, and $T/N = 0.6$ to achieve a balance between the performance and the computational cost. Details for the classification results with the aforementioned settings are provided in Table 1. From the results, we can see that the introduced hierarchical structure and the flexible rule updating strategy have positive effect on the performance. By modeling the images with the structure of multiple hierarchy, there is significant performance increment compared with those that treat the image as single hierarchy. This is consistent with our observation that there exists much similarities between different categories of indoor scenes, and universally modeling the images is not appropriate. Using the hierarchical structure, it is able to classify the categories more properly. Meanwhile, we attempt to learn the relationships of indoor scenes by using flexible rule base updating strategies to reduce the effect of conflict rules. The results have proven the effectiveness of the proposed framework.

The distribution of the generated rules for MIT's indoor dataset is shown in Fig. 10, and the generated rule base is given in Table 2. In this figure, it is clear that most categories can be summarized using a maximum of five simple rules, indicating that there are semantic similarities among objects in the same category. Furthermore, the combination of simple rules can effectively describe complex scenes. This is because semantic similarities are ubiquitous in the real world. Indoor scenes can thus be described effectively with rules for semantic hierarchies.

4.3 *Horizontal Experimental Results*

In this subsection, we compare the proposed framework with traditional and state-of-the-art methods, including the classic BoVW,[25] the prototype based model,[68] the DPM,[24] the object bank,[72] a method based on multiresolution classification method,[73] and the important spatial pooling regions (ISPRs).[74] The DPM is a successful object detector that directly improves the traditional HoG. The concept of an object bank proposed by Li,[72] and this method offers high-level encoding of an object's appearance and spatial location information for image recognition. Zhou et al.[73] presented a scene-classification framework by introducing multiscale information to the original BoVW. ISPRs[74] jointly learn spatial-pooling regions with discriminative part appearance in a unified framework for scene classification.

Figure 11 shows the results of the proposed HASRI compared to other scene-classification methods. Results of methods with HoG feature were also provided in Table 3. The BoVW was used as baseline. We can see that the proposed method achieved best performance compared with other methods. The proposed HASRI framework detected the objects in each hierarchy, exhibiting their spatial relationship according to the semantic hierarchical structure. Furthermore, the HASRI constructed and updated rules dynamically from the dataset in order to generate a knowledge base for the decision tree, rather than relying on predefined rules. Thus, the proposed HASRI framework is effective for indoor scene classification, and it consistently outperformed state-of-the-art methods, in which other features were used, including SIFT and GIST.

5 Conclusion

In this paper, we investigated the scene classification problem and proposed a novel HASRI framework modeled on the biological processes of human cognition. The performance of the indoor scene classification is substantially affected by the number of detected objects and their spatial relationship. The semantic hierarchical structure in the HASRI framework can detect more objects and better represent their spatial relationship. With the proposed framework, the rules for a decision tree are constructed using a flexible strategy based on the semantic hierarchical structure, and these rules are updated during the learning process. Experimental results demonstrated that the HASRI framework is effective, and that it outperforms other methods for indoor scene classification.

Acknowledgments

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

References

1. G. Carneiro et al., "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(3), 394–410 (2007).
2. O. A. Penatti et al., "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognit.* **47**(2), 705–720 (2014).
3. R. Mottaghi et al., "Analyzing semantic segmentation using hybrid human-machine CRFs," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3143–3150, IEEE (2013).
4. L. Xie et al., "Orientational pyramid matching for recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 3734–3741, IEEE (2014).
5. S. Khan et al., "Geometry driven semantic labeling of indoor scenes," *Lect. Notes Comput. Sci.* **8689**, 679–694 (2014).
6. Z. Ye et al., "Cognition inspired framework for indoor scene annotation," *J. Electron. Imaging* **24**(5), 053013 (2015).
7. S. Banerji, A. Sinha, and C. Liu, "New image descriptors based on color, texture, shape, and wavelets for object and scene image classification," *Neurocomputing* **117**(0), 173–185 (2013).
8. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. **2**, pp. 2169–2178, IEEE (2006).
9. L. J. Li et al., "Object bank: a high-level image representation for scene classification & semantic feature sparsification," in *Advances in Neural Information Processing Systems 23*, J. Lafferty et al., Eds., pp. 1378–1386, Curran Associates, Inc., New York, NY (2010).
10. J. Tang et al., "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Trans. Image Process.* **21**, 2354–2360 (2012).
11. C. Zhang et al., "Beyond visual features: a weak semantic image representation using exemplar classifiers for classification," *Neurocomputing* **120**(0), 318–324 (2013).
12. M.-J. Escobar and P. Kornprobst, "Action recognition via bio-inspired features: the richness of center surround interaction," *Comput. Vision Image Understanding* **116**(5), 593–605 (2012).
13. T. Tang and H. Qiao, "Improving invariance in visual classification with biologically inspired mechanism," *Neurocomputing* **133**, 328–341 (2014).
14. K. Huang et al., "Biologically inspired features for scene classification in video surveillance," *IEEE Trans. Syst. Man, Cybern. B* **41**, 307–313 (2011).
15. J. Delaigle et al., "Human visual system features enabling watermarking," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME '02)*, Vol. **2**, pp. 489–492, (2002).
16. D. Alleysson, S. Susstrunk, and J. Herault, "Linear demosaicing inspired by the human visual system," *IEEE Trans. Image Process.* **14**, 439–449 (2005).
17. C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 300–312 (2007).
18. M. Wang, X. Liu, and X. Wu, "Visual classification by l1-hypergraph modeling," *IEEE Trans. Knowl. Data Eng.* **27**, 2564–2574 (2015).
19. M. Wang et al., "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.* **19**, 733–746 (2009).
20. R. Hong et al., "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.* **44**, 669–680 (2014).
21. J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Process.* **21**, 3262–3272 (2012).
22. L. Xie et al., "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.* **23**, 1994–2008 (2014).
23. M. X. Ribeiro et al., "Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques," *Data Knowl. Eng.* **68**(12), 1370–1382 (2009).
24. P. Felzenszwalb et al., "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010).
25. G. Csurka et al., "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, Vol. **1**, pp. 1–2 (2004).
26. P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2241–2248 (2010).
27. Z. Ye et al., "Hierarchical abstract semantic model for image classification," *J. Electron. Imaging* **24**(5), 053022 (2015).
28. A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.* **29**(8), 1233–1244 (1996).
29. G. H. Liu and J. Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.* **46**(1), 188–198 (2013).
30. S. Liapis and G. Tziritas, "Color and texture image retrieval using chromaticity histograms and wavelet frames," *IEEE Trans. Multimedia* **6**(5), 676–686 (2004).
31. T. Hou et al., "Bag-of-feature-graphs: a new paradigm for non-rigid shape retrieval," in *21st Int. Conf. on Pattern Recognition (ICPR '12)*, pp. 1513–1516, IEEE (2012).
32. L. Nanni, M. Paci, and S. Brahnam, "Indirect immunofluorescence image classification using texture descriptors," *Expert Syst. Appl.* **41**(5), 2463–2471 (2014).
33. 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 '05)*, Vol. **2**, pp. 524–531 (2005).
34. G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 3506–3513, IEEE (2012).
35. M. Hao et al., "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia* **12**(5), 462–473 (2010).

36. 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).
37. A. Jaimes and S. F. Chang, "A conceptual framework for indexing visual information at multiple levels," *Proc. SPIE* **3964**, 2 (1999).
38. L. Saitta and J. D. Zucker, *Abstraction in Artificial Intelligence and Complex Systems*, Springer, New York (2013).
39. J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 785–792, IEEE (2011).
40. H. Bannour and C. Hudelot, "Hierarchical image annotation using semantic hierarchies," in *Proc. of the 21st ACM Int. Conf. on Information and Knowledge Management*, pp. 2431–2434, ACM (2012).
41. H. Bannour and C. Hudelot, "Building semantic hierarchies faithful to image semantics," *Lect. Notes Comput. Sci.* **7131**, 4–15 (2012).
42. Z. Gao et al., "Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset," *Multimedia Tools Appl.* **68**(3), 641–657 (2014).
43. Y. Lu et al., "Constructing concept lexica with small semantic gaps," *IEEE Trans. Multimedia* **12**(4), 288–299 (2010).
44. 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).
45. R. L. Solso, M. K. MacLin, and O. H. MacLin, *Cognitive Psychology*, Pearson/A and B, Boston, MA (2005).
46. K. Rebai, O. Azouaoui, and N. Achour, "Bio-inspired visual memory for robot cognitive map building and scene recognition," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS '12)*, pp. 2985–2990 (2012).
47. C. Wongun, C. Yu-Wei, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3D geometric phrases," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 33–40 (2013).
48. P. K. Novak, N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining," *J. Mach. Learn. Res.* **10**, 377–403 (2009).
49. X. Chen, A. Shrivastava, and A. Gupta, "NEIL: extracting visual knowledge from web data," in *IEEE Int. Conf. on Computer Vision (ICCV '13)*, pp. 1409–1416, IEEE (2013).
50. R. Brachman and H. Levesque, *Knowledge Representation and Reasoning*, Morgan Kaufmann Publishers Inc., San Francisco, CA (2004).
51. A. Gupta et al., "Parallel algorithms and architectures for rule-based systems," in *SIGARCH Comput. Archit. News*, Vol. **14**, pp. 28–37 (1986).
52. S. D. Tran and L. S. Davis, "Event modeling and recognition using Markov logic networks," in *Computer Vision–ECCV 2008*, pp. 610–623, Springer (2008).
53. F. Hayes-Roth, D. Waterman, and D. Lenat, *Building Expert Systems*, Addison-Wesley, Reading, MA (1984).
54. W. F. Bischof and T. Caelli, "Scene understanding by rule evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(11), 1284–1288 (1997).
55. W. S. Zhou, S. Dao, and C. C. J. Kuo, "On-line knowledge- and rule-based video classification system for video indexing and dissemination," *Inf. Syst.* **27**(8), 559–586 (2002).
56. A. Amato and V. Di Lecce, "A knowledge based approach for a fast image retrieval system," *Image Vision Comput.* **26**(11), 1466–1480 (2008).
57. J. H. Ang, K. Tan, and A. Mamun, "An evolutionary memetic algorithm for rule extraction," *Expert Syst. Appl.* **37**(2), 1302–1315 (2010).
58. M. Xu and M. Petrou, "3D scene interpretation by combining probability theory and logic: the tower of knowledge," *Comput. Vision Image Understanding* **115**(11), 1581–1596 (2011).
59. X. Chen, A. Shrivastava, and A. Gupta, "Enriching visual knowledge bases via object discovery and segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 2035–2042, IEEE (2014).
60. 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).
61. S. L. Kendal and M. Creen, *An Introduction to Knowledge Engineering*, Springer, London, United Kingdom (2007).
62. M. J. Choi, A. Torralba, and A. S. Willsky, "Context models and out-of-context objects," *Pattern Recognit. Lett.* **33**(7), 853–862 (2012).
63. L. Li-Jia et al., "Building and using a semantivisual image hierarchy," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3336–3343 (2010).
64. P. Gupta et al., "Video scene categorization by 3D hierarchical histogram matching," in *IEEE 12th Int. Conf. on Computer Vision*, pp. 1655–1662, IEEE (2009).
65. G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*, pp. 229–238, AAAI/MIT Press, Cambridge, MA (1991).
66. R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.* **22**, 207–216 (1993).
67. J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA (2014).
68. A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 413–420 (2009).
69. "The Indoor scene recognition dataset," http://web.mit.edu/torralba/www/indoor.html
70. C. Vondrick et al., "Hoggles: visualizing object detection features," in *ICCV* (2013).
71. M. Hall et al., "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.* **11**, 10–18 (2009).
72. L. J. Li et al., "Object bank: an object-level image representation for high-level visual recognition," *Int. J. Comput. Vision* **107**(1), 20–39 (2014).
73. L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.* **46**(1), 424–433 (2013).
74. D. Lin et al., "Learning important spatial pooling regions for scene classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 3726–3733 (2014).

Dan Yu is a PhD candidate at the Pattern Recognition Research Center of Harbin Institute of Technology (HIT). He received his BS and MS degrees in computer application from Harbin Institute of Technology and Nanjing University of Science and Technology in 2002 and 2011, respectively. His research interest covers image processing, pattern recognition, and visual control and navigation.

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

Zhipeng Ye is a PhD candidate at the School of Computer Science and Technology, HIT. He received his master's degree in computer application technology from HIT in 2013. His research interest covers image processing and machine learning.

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

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 interest covers pattern recognition, image processing, and deep space target visual analysis.