Journal of Electronic Imaging

JElectronicImaging.org

# Hierarchical abstract semantic model for image classification

Zhipeng Ye
Peng Liu
Wei Zhao
Xianglong Tang

SPIE. IS&T imaging.org

# Hierarchical abstract semantic model for image classification

**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.** Semantic gap limits the performance of bag-of-visual-words. To deal with this problem, a hierarchical abstract semantics method that builds abstract semantic layers, generates semantic visual vocabularies, measures semantic gap, and constructs classifiers using the Adaboost strategy is proposed. First, abstract semantic layers are proposed to narrow the semantic gap between visual features and their interpretation. Then semantic visual words are extracted as features to train semantic classifiers. One popular form of measurement is used to quantify the semantic gap. The Adaboost training strategy is used to combine weak classifiers into strong ones to further improve performance. For a testing image, the category is estimated layer-by-layer. Corresponding abstract hierarchical structures for popular datasets, including Caltech-101 and MSRC, are proposed for evaluation. The experimental results show that the proposed method is capable of narrowing semantic gaps effectively and performs better than other categorization methods. *© 2015 SPIE and IS&T* [DOI: 10.1117/1.JEI.24.5.053022]

Keywords: image classification; bag-of-visual-words; semantic abstraction; hierarchical structure; semantic gaps.

Paper 15406 received May 20, 2015; accepted for publication Aug. 21, 2015; published online Oct. 6, 2015.

## 1 Introduction

Due to the explosive growth of digital techniques, especially the proliferation of smart phones with high-quality image sensors, visual datasets are created and stored as often as text data. To store and retrieve visual information more efficiently, it is necessary to develop automatic image annotation and object categorization techniques. Automatic object categorization is a developing field in computer vision. It is also the precondition of scene interaction in artificial intelligence and has become a goal of important value in image collection. Studies in object classification have reached considerable levels of performance. Among all the methods, the bag-of-visual-words (BoVW) method[1] is one of the approaches most commonly used in image retrieval (IR) and scenario classification, whose simplicity and effectiveness have been tested throughout the years. However, the performance of methods based on low-level features, such as BoVW, is affected by the so-called semantic gap between higher-level ordinary human concepts and their lower-level image representations.[2] Semantic compression has been proposed to narrow the semantic gap and avoid a supersized visual word vocabulary.[3–6] In addition, several hierarchy-based studies have been performed.[7–11] Hierarchy-based methods tend to build up higher-level semantic vocabularies to narrow the semantic gap. Li[7] proposed a Bayesian hierarchical model denoting a training set as codewords produced with unsupervised learning. Bannour and Hudelot[8,9] proposed a hierarchy-based classifier training method by decomposing the problem into several independent tasks to estimate the semantic similarity between the concepts, by incorporating visual, conceptual, and contextual factors information to provide a more expressive form of semantic measurement for images. Li-Jia et al.[10] proposed a method of automatically determining semantic features of image hierarchy by incorporating both image and tag information. Katsurai et al.[11] presented a cross-modal approach for extracting semantic relationships between concepts that was suitable for concept clustering and image annotation.

A hierarchical abstract semantics (HAS) model is proposed for object categorization in this paper. The HAS model is different from other hierarchical methods in three ways that further improve the performance. The first is the strategy by which semantic features are selected. Existing methods select different parts of images separately, but HAS treats semantic visual words from the whole image as one semantic feature to establish a higher-quality visual vocabulary. The proposed model also has a hierarchical structure with an upper abstract semantic layer, additional middle abstract semantic layer, and concrete layer to further narrow the semantic gaps. Previous studies have ignored the middle layer. Some previous works have also straightforwardly trained classifiers, but here, Adaboost is used to iteratively combine weak classifiers into strong ones to improve performance. This method is tested on popular computer vision datasets using corresponding hierarchical structures to quantify semantic gap and categorization performance.

The rest of the paper is organized as follows: In Sec. 2, relevant previous works are discussed. In Sec. 3, the HAS method is discussed in detail. The experimental results are presented in Sec. 4. Conclusions are presented in Sec. 5.

## 2 Related Work

The current work is related to two types of research, object recognition and image annotation, and measurement of semantic gaps. Relevant previous works are reviewed in the following subsections.

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

### 2.1 *Object Recognition and Image Annotation*

In the past decade, numerous studies have been carried out on automatic object categorization as a part of IR. In general, studies on IR can be divided into three types. The first approach is the traditional text-based approach to annotation. It uses human power to annotate images, and images are retrieved by text.[12] The second focuses on content-based IR. The images are commonly retrieved using low-level features such as color, shape, and texture.[13–17] The third approach is based on automatic image annotation techniques and involves learning models trained by a large number of sample images and uses the models to label new images. The field of automatic image annotation can be further divided into three subcategories. The first is based upon global features.[18] Supervised classification techniques can be used to solve the categorization tasks. The second uses regional features that represent an image as a set of visual blobs,[19,20] converting the categorization task to a problem of learning keywords from visual regions. It is reasonably common to use bag-of-features[1] (BoF) for annotation and categorization. Many effective categorization methods are vocabulary based.[21–23] A BoF model is built to represent an image as a histogram of local features. This is the basis of the BoVW. In BoVW, a codebook is constructed by clustering all the local features in the training data and an image is treated as a collection of unordered "visual words," which are obtained by *k*-means clustering local features. Then the image is represented using BoF to train the classifier. Jiang et al.[24] evaluated various factors that affect the performance of BoF for object categorization including selection of detector, kernel, vocabulary size, and weighting scheme. However, there are several drawbacks to BoVW, including the following: (1) spatial relationships between image patches are ignored during the construction of visual vocabulary;[25] (2) the hard-assignment strategy used by *k*-means does not necessarily generate optimized visual vocabulary;[26] and (3) semantic concepts are ignored during the clustering process.[27] These shortcomings have a significant effect on performance. Many studies have been carried out to solve these problems. Chai et al.[28] utilized foreground segmentation to improve classification performance on weakly annotated datasets. In order to address the problem that the BoF model ignores spatial information among local features, spatial pyramid matching (SPM)[29] was proposed to make use of the spatial information for object and scene categorization. There are several ways to generate highly descriptive visual vocabulary. Wang et al.[30] presented a simple but effective coding scheme called locality-constrained linear coding (LLC) in place of the vector quantization coding in traditional SPM to improve the categorization performance. van Gemert et al.[31] stated that one way to improve the system's ability to describe visual words is to introduce ambiguity into visual words. Semantic layers were constructed to narrow semantic gaps to generate better visual vocabulary.[32] Semantic visual vocabulary is established to increase its quality. Deng et al.[33] proposed a similarity-based learning approach, which was able to exploit hierarchical relationships between semantic labels at the training stage to improve the performance of IR. Wu et al.[34] developed a semantic-preserving bag-of-words (SPBoW) scheme to produce optimized BoW models by generating a semantic preserving codebook.

### 2.2 *Semantic Gap Measurement*

The semantic gap is commonly defined as the lack of coincidence between the information extracted from the visual data and its interpretation.[35] Millard et al.[36] presented a vector-based model of the formality of semantics in text systems, which represents the translation of semantics between the system and humans. Each image is commonly considered relevant to more than one semantic concept, so there are also semantic gaps between each of the concepts. The purely automatic image annotation techniques are still far from satisfactory due to the well-known semantic gap. A few research efforts have been made to determine how to quantitatively measure the semantic gaps of concepts. According to the criterion that different concepts correspond to different semantic gaps, Lu et al.[37] proposed a method to quantitatively analyze semantic gaps and developed a framework to identify high-level concepts with small semantic gaps from a large-scale web image dataset. Due to the significance of measuring semantic gaps, a few scientific studies on the quantification of semantic concepts have been performed. Zhuang[38] proposed a measure for semantic gaps from the perspective of information quality. The semantic gap was treated as the cause of inefficacy of the transmission of information through representation of an information system, indicating the fact that the carrier of information was unable to transmit the information in question. Tang et al.[2] proposed a semantic gap quantification method during the study of a semantic-gap-oriented active learning method and incorporated the semantic gap measure into the sample selection strategy by minimizing corresponding information. In this paper, the quantification criterion proposed by Tang et al.,[2] which is one of the feasible methods to quantify semantic gap, is used to evaluate methods.

## 3 Hierarchical Abstract Semantics Method

Semantic hierarchies can improve the performance of image annotation by supplying a hierarchical framework for image classification and provide extra information in both learning and representation.[8] Three types of semantic hierarchies for image annotation have recently been explored: (1) language-based hierarchies based on textual information,[32] (2) visual hierarchies based on low-level image features,[39] and (3) semantic hierarchies based on both textual and visual features.[10] Here, the original BoVW is extended by introducing the inferential process of categories using a technique of abstraction to further narrow the semantic gap. The architecture of the HAS model is described in Sec. 3.1. Then the training method is described in Sec. 3.2.

### 3.1 *Hierarchical Abstract Semantics Model*

In this paper, a method of learning hierarchical semantic classifiers is proposed. This method relies on the structure of semantic hierarchies to train more accurate classifiers for classification. It is divided into two parts: a bottom-up training step and a top-down categorizing step. To build semantic hierarchies, there is a basic assumption that the objects in the real world that fit into the same category share a limited number of common attributes.[40] Here, original categories of each image dataset are called concrete categories (CCs). An abstract category is made up of CCs that share common features. This process is done manually, offline, and serves as prior information. For example, the middle abstract category

"bird" is made up of the concrete classes "sparrow," "chicken," and so on, while "bird" is part of the upper abstract category "animal." The structure of the original BoVW and the proposed HAS model is shown in Fig. 1. Unlike existing methods,[8,33] the proposed method has several abstract layers constructed using abstract semantics. The upper abstract layer is used to determine the general category of an image using an support vector machine (SVM)-based classifier named U-SVM. SVM is chosen as a basic classifier, since it is a classical and typical method in image classification with better performances.[41] Similarly, the middle abstract layer generates a refined category using M-SVM. Both layers can be extended flexibly according to the application. The purpose of using both U-SVM and M-SVM is to generate abstract semantics from top to bottom. The degree of abstraction increases from bottom to top. It increases the system's ability to describe the target and reduces the differences among CCs, while common attributes between each pair of CCs under the same abstract category are preserved. However, the number of categories, number of corresponding attributes, and quantity of information increase as the degree of abstraction decreases from top to bottom. For example, in the concrete layer, the category "chicken" is different from the category "sparrow." However, when the abstract level increases, they are merged into the same abstract category, "bird," which describes both categories by aggregating their common features such as feathers, wings, and beaks. The abstract layer works like middleware in semantics. It connects the real-world and image datasets. Information can be transmitted through the abstract layer: semantic visual words are transmitted bottom-up and the category of an image is transmitted top-down. Concrete classes under same abstract class share some common features and are similar to those under different abstract classes. Because of this, every concrete class under the same abstract class is similar but distinguishable from its fellows, indicating that the inner-class distance is small and the intraclass distance is large. It facilitates classification.

Unlike original BoVW, which uses CC training and categorizing in a flat way shown in Fig. 1(a), an abstract level, including one middle layer and one upper abstract layer is introduced, as shown in Fig. 1(b). The purpose of introducing abstract layers to the HAS model is to narrow the semantic gap. Both the middle and the upper sublayers of the abstract layers are constructed using semantic-preserved visual words[34] extracted from CC. As shown in the figure, HAS is a superset of BoVW. If the abstract layers were omitted, HAS would degrade into a standard BoVW.

### 3.2 *Bottom-Up Semantic Classifier Learning*

The training of HAS is a three-step, bottom-up process. First, each concrete classifier $\text{BoVW}_j$ is trained using a visual semantic attribute, which is composed of semantics' visual words generated through SPBoW.[34] The input of $\text{BoVW}_j$ from the concrete layer is the images collected from each dataset. Then to train classifier $\text{M-SVM}_i$ of the middle abstract semantic category (MASC), samples from every CC of the MASC were randomly selected with equal probability to make sure every category has a chance to be selected in establishing the visual vocabulary, which improves the ability of description. Then a semantic visual vocabulary is generated from selected samples using SPBoW training $\text{M-SVM}_i$. U-SVM classifiers of an upper abstract semantic category (UASC) are trained in the same way. Finally, the Adaboost training strategy[42] is used to combine weak classifiers into a strong one to complete the learning stage.

### 3.3 *Top-Down Categorization*

After the bottom-up training stage is completed, trained classifiers are ready for categorization. To determine the category of an input image, UASC $u$ is first generated by U-SVM. Then corresponding MASC $m$ is calculated by M-SVMs. Finally, the CC $c$ is decided upon. The categorizing process can be universally described as

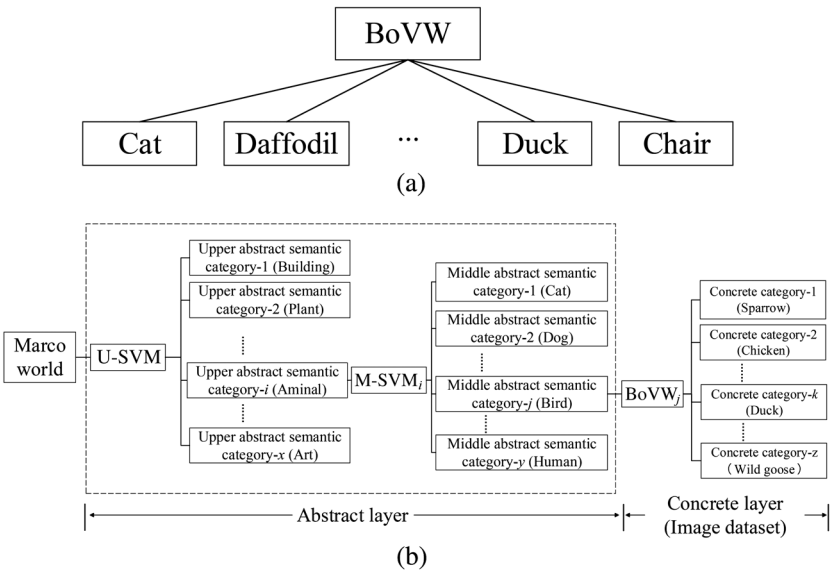



**Fig. 1** Architecture of flat bag-of-visual-word (BoVW) and hierarchical abstract semantics (HAS) model: (a) flat structure of BoVW, and (b) hierarchical structure of HAS.

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

Here, $F_t^u$ and $F_j^m$ are the visual features of the upper and middle abstract layers, $F_k$ is the visual feature of the concrete layer, and $D$ is the measurement provided by the classifier.

Since the decision processes are serial, if $u$ is incorrect, the rest of the decisions are meaningless. There are two strategies to decrease the dependence of the lower layers on decisions from the upper layers: (1) passing testing image $I$ through upper and middle abstract classifiers, the output of each upper and middle classifier is $P_u$ and $P_m$, respectively. The middle category is decided by the value of the $i$'th upper classifier $p_{ui}$ and the $j$'th middle classifier $p_{mj}$, which is described as follows:

$$C_{\text{middle}} = \sum_{i=1}^{U} \sum_{j=1}^{M} \text{argmax}(p_{ui} + p_{mj}). \quad (2)$$

Here, $U$ and $M$ are the numbers of upper and middle classifiers; (2) Adaboost training strategy is used to improve the performance of every classifier from each layer. Traditional BoVW is used to produce $n$ outputs $p_1, p_2, \ldots, p_n$, and $n$ depends on the number of categories under each classifier. $I$ is categorized according to the category classifier that outputs the largest value:

$$C = \underset{t=1,2,\ldots,n}{\text{argmax}} (p_t). \quad (3)$$

### 3.4 *Proposed Algorithm*

Unlike the original BoVW, the proposed model makes its improvement from the perspective of abstraction by introducing abstract semantic layers to narrow semantic gaps. Moreover, semantic visual vocabulary and Adaboost are utilized as the training feature and strategy, respectively, to improve the performance of classifiers. The preparing and learning algorithms of HAS are, respectively, described in Algorithms 1, 2, and 3, where $k = 1, \ldots, m$, $\text{CC}_k$ is the short for $k$'th CC, $\text{MAC}_j$ is the abbreviation of the $j$'th middle abstract category, $\text{UAC}_i$ 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.

For the preparation of HAS, at first, samples are randomly selected from each CC to generate the bottom semantic visual vocabulary. Then according to the hierarchical structure mentioned above, images of all categories under the same MASC are randomly selected to generate the visual vocabulary for the MASC. To build the visual vocabulary of UASC, samples are selected from all CCs under the same UASC. As shown from this process, the visual vocabulary of UASC covers more CCs than those of any MASC. Thus, the ability of the system to express itself increases from bottom to top, which is the same as the degree of abstraction. After the steps of generating visual vocabularies, classifiers on every level of the hierarchy are trained layer-by-layer with the strategy of Adaboost from bottom to top. According to the previous work,[42] SVM classifiers trained by a few samples are considered weak classifiers. For this reason, before the training

**Algorithm 1** The preparation stage of HAS.

**Input:** Training image set.

**Output:** Semantic visual vocabulary for each category.

1: For each $\text{CC}_k$ under $\text{MAC}_j$, generate SVVS, where $v_q$ and $s_q$ are visual words and their corresponding semantic information; $c$ is the size of the codebook.

2: The SVVS of $\text{MAC}_j$ under $\text{UAC}_i$ is constructed by $M_j = \bigcup_{k=1}^{z} \text{Inh}_k^j$ and $\text{M-A}_{-i} = \bigcup_{j=1}^{y} M_j$.

3: For each $\text{UAC}_i$, randomly select SVVS with equal probability from each $\text{Inh}_k^j$, forming $\text{U-ABS}_i = \bigcup_{j=1}^{y} \bigcup_{k=1}^{z} \text{Inh}_k^j$. Let $\text{U-A} = \bigcup_{i=1}^{x} \text{U-ABS}_i$.

4: **Return** U-A.

**Algorithm 2** The training processes of HAS.

**Input:** Visual vocabulary of concrete layer $\text{Inh}_k^j$, middle abstract layer $\text{M-A}_i$, and upper abstract layer U-A.

**Output:** Strong classifiers $\bigcup_{j=1}^{N_i} \text{BoVW}_j$ for concrete layer, $\bigcup_{i=1}^{N_m} \text{M-SVM}_i$ for middle abstract layer, and U-SVM for upper abstract layer.

1: For every $\text{MAC}_j$ ADABOOST-TRAIN$(\text{BoVW}_j, \text{Inh}_k^j)$.

2: For every $\text{UAC}_i$ ADABOOST-TRAIN$(\text{M-SVM}_i, \text{M-A}_i)$.

3: ADABOOST-TRAIN(U-SVM, U-A).

4: **Return** Trained classifiers for each layer, including $\bigcup_{j=1}^{N_i} \text{BoVW}_j$, $\bigcup_{i=1}^{N_m} \text{M-SVM}_i$, and U-SVM.

**Algorithm 3** The processes of ADABOOST-TRAIN.

**Input:** Visual vocabulary set $\text{U-A} = (x_1, y_1), \ldots, (x_N, y_N)$ with labels and the weights of training samples: $w_i^1 = 1/N$, $i = 1, \ldots, N$; the initial minimum error rate $\varepsilon_{\min}$.

**Output:** Strong classifiers for each category.

1: Train a linear SVM component classifier, $h_t$, on the weighted training set.

2: Calculate the training error of $h_t$: $\varepsilon_t = \sum_{i=1}^{N} w_i^t$, $y_i \neq h_t(x_i)$.

3: Update the weights of training samples: $w_i^{t+1} = \{w_i^t \exp[-\alpha_t y_i h_t(x_i)]\}/C_t$.

4: Update the weights of component classifier $h_t$: $\alpha_t = \ln[2(1 - \varepsilon_t)]$, $C_t$ is a normalization constant, and $\sum_{i=1}^{N} w_i^{t+1} = 1$.

5: If $(\varepsilon_t > \varepsilon_{\min})$ go to (1).

6: **Return** $f(x) = \text{sign}[\sum_{t=1}^{T} \alpha_t h_t(x)]$.

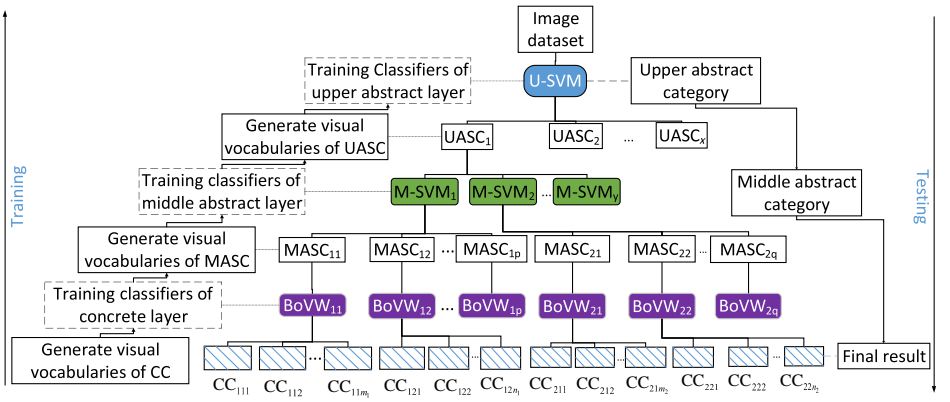



**Fig. 2** The whole procedure and framework of the proposed HAS.

starts, the weak classifiers are initially trained with vocabularies of different sizes, and the weights are adjusted according to the corresponding error rates. Finally, to classify an image, that image is passed through the classification framework from top to bottom. The procedure outlined above and the framework of the proposed HAS method are shown in Fig. 2.

## 4 Experimental Results

Two popular CV datasets were used to evaluate the performance of the proposed model on image categorization: MSRC[43] and Caltech-101.[44] MSRC contains 23 object classes with 591 images, which are labeled by pixel. The size of each image is roughly $320 \times 240$ 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 dataset. Four abstract categories are constructed by selecting 14 CCs with sufficient and unambiguous training and testing images. Caltech-101 contains 101 categories with 9197 images. The size of each image is roughly $300 \times 200$ pixels. Outlines of each object are carefully annotated. Most images contain only one object, centered, which renders object recognition less difficult. Some of the samples are shown in Fig. 3. For MSRC, at least 15 images for each abstract category are selected for training. For Caltech-101, the strategy reported by Wang et al.[30] was used to train the classifier. The remaining images in the datasets are used or testing.

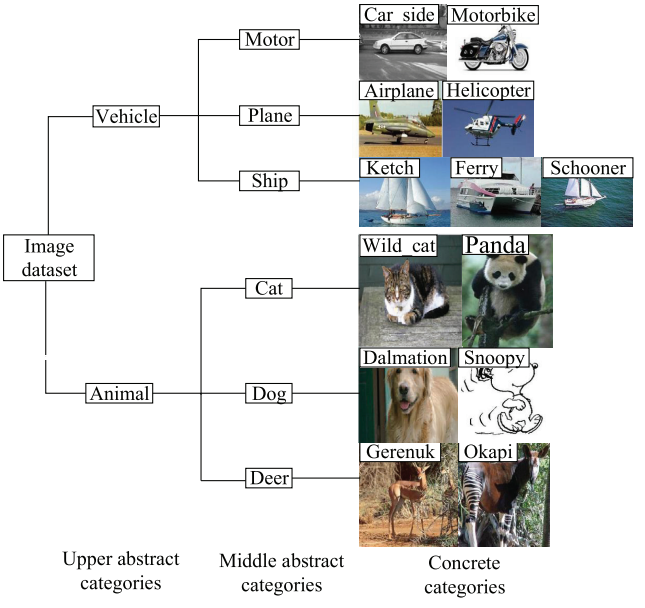



**Fig. 3** Sample images and corresponding hierarchical structure of Caltech-101 dataset.

### 4.1 *Setup*

The hierarchical structure of MSRC and Caltech-101 used in this paper, which was inspired by the works of Bannour and Hudelot[8] and Li-Jia et al.,[10] is shown in Fig. 4. The structures of each dataset were organized and presented for the first time. The proposed method is compared with BoVW,[1] LLC,[30] and one-versus-opposite-nodes (OVON).[8] Mean average precision and area under curve were used to evaluate the experimental results. Lowe's scale-invariant feature transform (SIFT) descriptor[45] was used to detect keypoints. The sizes of visual vocabularies are 1000 per dataset generated by a $k$-means algorithm, which means that each image is represented by a histogram of 1000 visual words and each bin in the histogram corresponds to the number of occurrences of a visual word in that image. The LIBLINEAR open source library was chosen to implement linear SVM due to its excellent speed and performance on large-scale datasets.[46] The one-versus-all training strategy was used to train classifiers. $\varepsilon_{\min}$ was set to be 0.3. To fairly reflect the difference between each method, we use SIFT as an image descriptor instead of the histogram of oriented gradients utilized in LLC.[30]

During the process of categorization, the performance of actual CC was evaluated as follows:

$$\text{correct\_rate} = \frac{\sum_{j=1}^{N} \delta(C_j - C_i)}{N} \times 100\%. \tag{4}$$

Here, $N$ is the number of testing images, and

$$\delta(C_j - C_i) = \begin{cases} 1 & C_j = C_i \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Semantic measurement[2] is used to quantitatively reflect semantic gaps of BoVW and HAS. Both UASC and MASC, defined above, were used to fairly quantize the semantic gap between different methods. For BoVW, the semantic gap is quantified as follows:

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

Here, $N(x_i)$ represents the set of the $k$-nearest neighbors of $x_i$ in the visual space. The semantic distance $\text{dis-sim}(x_i, x_j)$ between $x_i$ and each of its neighbors $x_j$ is measured by the cosine distance between the vectors of their tags. For HAS, the semantic gap is quantified as follows:

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

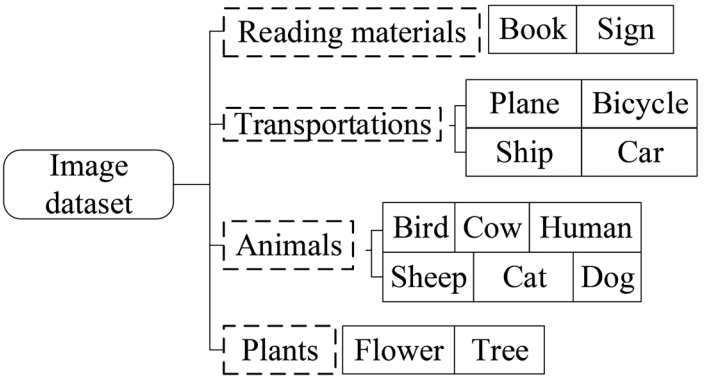
Here, ${}^{M}_{I}\text{Im-SG}(x_i)$ stands for the image semantic gap between the concrete layer and MASC, and ${}^{U}_{M}\text{Im-SG}(x_i)$ stands for the image semantic gap between MASC and UASC. HAS was prepared as described in Sec. 3.4, which means image semantic gaps were calculated as soon as the training data were ready. The experiments were performed on a workstation with quad-core 2.13 GHz CPU and 12 GB memory.

### 4.2 *Results and Analysis*

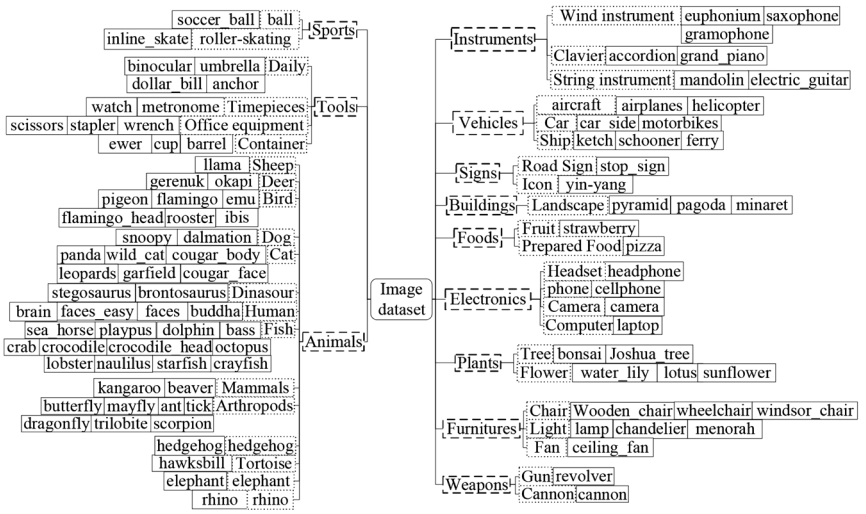
The results of these experiments are presented in this subsection, which is divided into two parts, including both horizontal and vertical comparisons on the proposed method,

**Table 1** Experimental results on performance and complexity tests on different numbers of hierarchies.

| Hierarchical number of HAS | MSRC | Caltech-101 |
|---|---|---|
| 1 | $0.327P_1$ (−67.3%) | $0.352P_2$ (−64.8%) |
| | $0.56T_1$ (−44%) | $0.431T_2$ (−56.9%) |
| 2 | $0.67P_1$ (−33%) | $0.634P_2$ (−36.6%) |
| | $0.85T_1$ (−15%) | $0.697T_2$ (−30.3%) |
| 3 | $P_1$ | $P_2$ |
| | $T_1$ | $T_2$ |
| 4 | $1.021P_1$ (+2.1%) | $1.01P_2$ (+1%) |
| | $1.8T_1$ (+80%) | $2.48T_2$ (+148%) |
| 5 | $1.028P_1$ (+2.8%) | $1.018P_2$ (+1.8%) |
| | $3.2T_1$ (+220%) | $3.74T_2$ (+274%) |




(a)




(b)

**Fig. 4** Details of the hierarchical structures for Caltech-101 and MSRC. Upper, middle abstract, and concrete categories (CCs) are represented by dashed, dotted, and solid boxes, respectively. For MSRC, the upper and middle abstracts are merged into single dashed boxes to make the chart clearer: (a) the semantic hierarchical structure of MSRC, and (b) hierarchical structure of Caltech-101.

and single-layer and multiple-layer image categorization methods.

#### 4.2.1 *Evaluation of parameters*

To determine the number of hierarchies for HAS, both average classification performance and runtime for a single image here served as agents of comprehensive measurement. The experimental results are given in Table 1. Here, we utilize symbols $P$ and $T$ to represent the performance of classification and consumption of time, where $P_1 = 0.813$, $T_1 = 0.357$ s, $P_2 = 0.763$, and $T_2 = 0.452$ s. These results show that the performance of HAS is relevant to the number of hierarchies. HAS with one or two hierarchies has a shorter runtime for the whole algorithm than HAS with three. However, at the cost of a significant drop in classification performance, this indicates that adding more detailed semantic layers to increase the performance of HAS is not effective, since it takes much more time to run. This shows that HAS with three layers achieves the best balance between performance and runtime. This setting of the number of hierarchies was addressed in the following experiments.

At the second part of self-evaluation, the HAS presented here was compared with BoVW on both datasets using the criterion of semantic gap quantification. For BoVW with flat structure, the semantic gap was calculated directly. For HAS, the total semantic gap was the sum of each semantic gap between the upper, middle, and concrete concepts in each layer. A comparison of BoVW with HAS regarding semantic gap quantification is shown in Fig. 5. As shown, HAS was more effective in narrowing the semantic gap between concepts and visual data. For abstract categories with few CCs, the difference between BoVW and HAS was not significant. When the scale of the abstract category is small, the disturbance between each CC is also relatively small, so the ability of narrowing the semantic gap between both methods is approximately coincident. But for larger abstract categories such as "animal" of Caltech-101, substantial improvement is observed. In general, HAS is more effective in narrowing the semantic gap. This is because the introduction of upper and middle abstract semantics reduces the inconsistency between the distributions of low-level visual features and the high-level semantic concepts. In the process of constructing visual

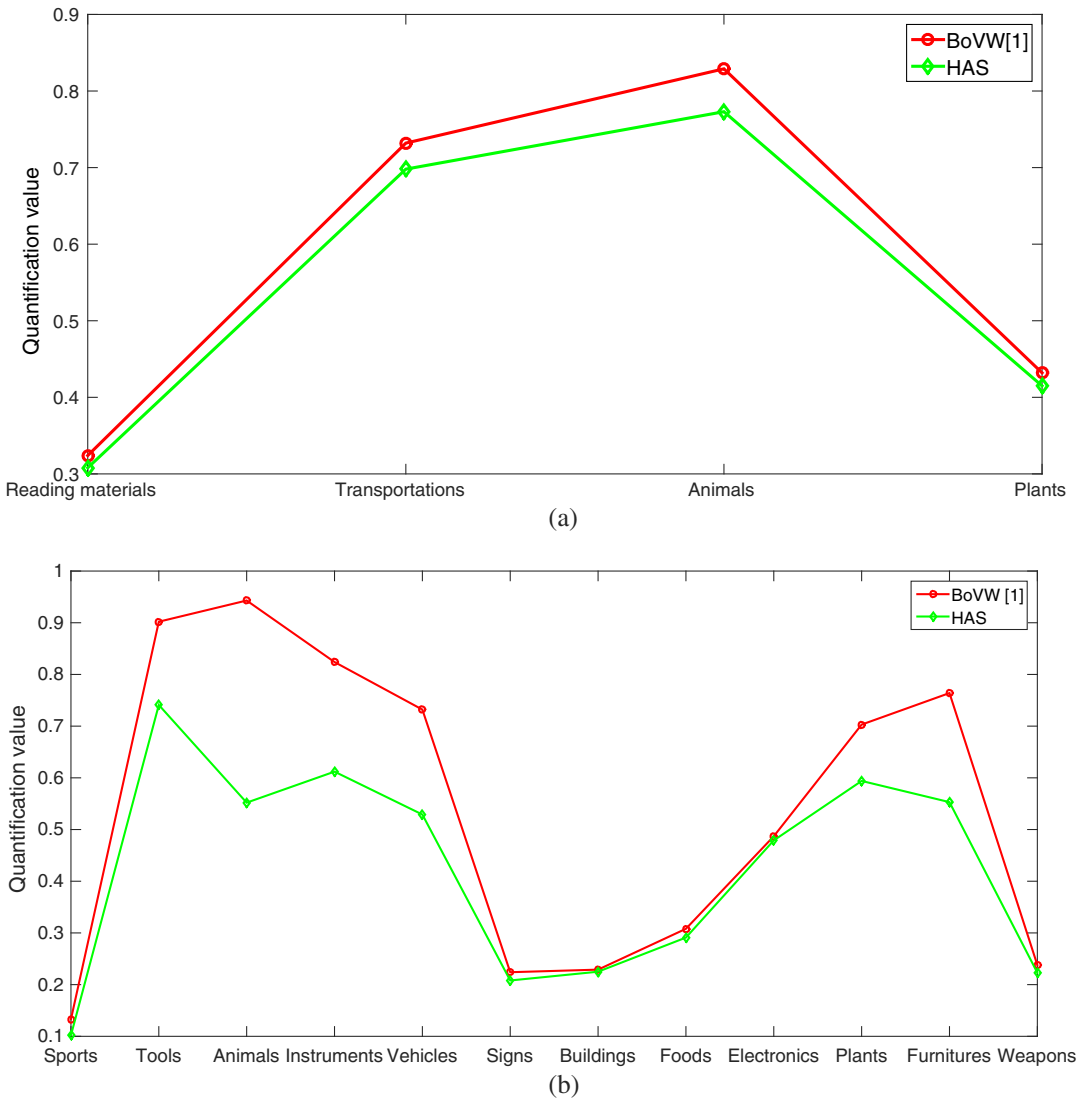



**Fig. 5** Semantic gap quantization result: (a) semantic gap quantization result of MSRC, and (b) semantic gap quantization result of Caltech-101.

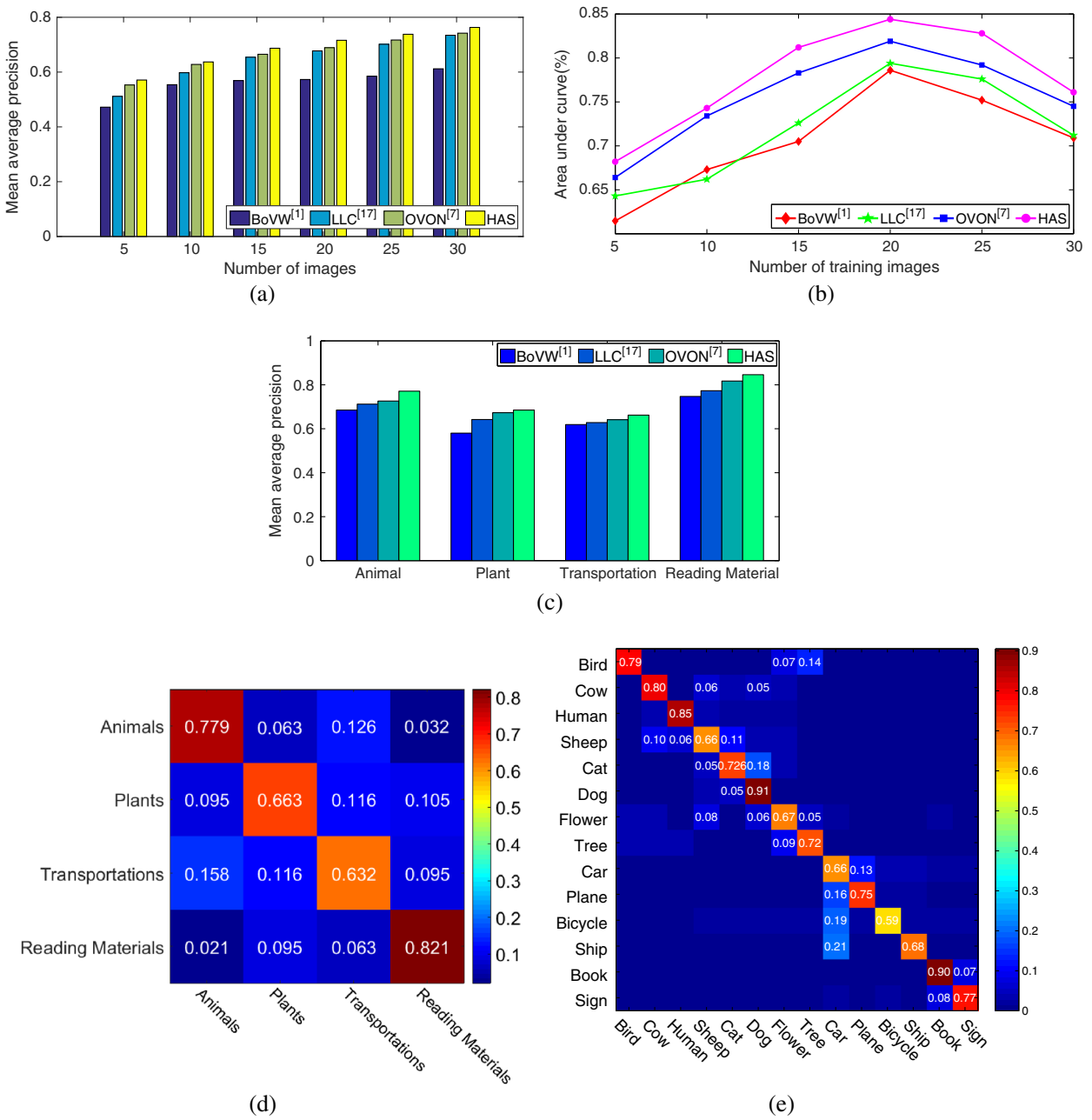



**Fig. 6** Categorization results for each dataset: (a) classification results on Caltech-101, (b) area under curve of Caltech-101, (c) classification results on MSRC, (d) confusion table of abstract categories on MSRC, and (e) confusion table of CCs on MSRC.

data, the introduction of upper and middle abstract layers ensures that the visual vocabularies are constructed from relevant categories. The improvement is also present because HAS classifiers trained with semantic visual vocabularies generated by SPBoW[34] are much more descriptive than the ordinary visual vocabularies of BoVW.

### 4.2.2 *Evaluation with different categorization methods*

The classification results of both MSRC and Caltech-101 are shown in Fig. 6. The setting of the testing processes of Caltech-101 is the same as those reported by Wang et al.[30] The experimental results given in Figs. 6(a) and 6(b) show that the performance of HAS is a substantial improvement over that of other methods, including classical BoVW, LLC, and hierarchical OVON methods on both datasets. There are reasons why the proposed HAS performs better than other methods. First, during construction of the visual vocabulary, HAS does not exclude the ambiguity of visual words and uses semantic visual words from the whole image as one semantic feature to construct a high-quality visual vocabulary during the establishment of abstract semantic visual vocabularies. This construction strategy is beneficial for improving the categorization performance.[31] HAS selects lower semantic visual words randomly with equal probability, constructing upper semantics to make sure that each term has a chance to be selected for training. The Adaboost

**Table 2** Comparative results with relative methods on MSRC dataset.

| Methods | Caltech-101 |
|---|---|
| Lazebnik et al.[29] | $0.646 \pm 0.008$ |
| Zhong[48] | 0.69 |
| Maji et al.[49] | $0.566 \pm 0.008$ |
| Bilen et al.[50] | $0.753 \pm 0.007$ |
| McCann and Lowe[51] | $0.76 \pm 0.009$ |
| Proposed | $0.763 \pm 0.012$ |

training strategy was used to construct strong classifiers to further improve the performance with respect to classification.

The quality of classification on both datasets is shown in Figs. 6(c)–6(e). For Caltech-101, area under curves are given to show the quality of classification because this dataset is much larger than MSRC, and it is difficult to list results for each category, as in previous works.[30] The confusion matrix of MSRC on every abstract and in every CC is listed in Figs. 6(d) and 6(e) to further show the details of the classification results. Values larger than 0.05 are listed to render the table clear, which is consistent with results reported by Zhou et al.[47] According to the experimental results, the introduction of multiple semantic layers can be used to distinguish one category from another. Most categorization errors occur under the same abstract category, proving that the semantic gap is efficiently arrowed by the introduction of abstract layers, meeting the quantification result listed above. The results of some popular and state-of-the-art categorization methods on Caltech-101 are reported in Table 2. The results show that the performance of the HAS presented here is comparable with that of traditional and state-of-the-art classification methods.

## 5 Conclusion

A multilayer abstract semantics inference model that is highly abstract and easy to extend was introduced here to deal with object categorization problems. Three techniques were used here to improve the performance of the categorization process: abstract layers, semantic vocabularies, and the Adaboost training strategy, which form the whole framework. The capabilities of this were demonstrated here on popular computer vision datasets, and it showed substantial improvements in semantic gap and categorization performance over existing methods. Abstract hierarchical structures were used on existing datasets for construction of semantic visual vocabularies. Future work will include building a generic abstract knowledge base to render the framework dataset independent.

### *Acknowledgments*

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

### *References*

1. G. Csurka et al., "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision (ECCV 2004)*, Vol. 1, pp. 1–2 (2004).
2. J. Tang et al., "Semantic-gap-oriented active learning for multi-label image annotation," *IEEE Trans. Image Process.* **21**(4), 2354–2360 (2012).
3. 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).
4. B. Fernando et al., "Supervised learning of Gaussian mixture models for visual vocabulary generation," *Pattern Recognit.* **45**(2), 897–907 (2012).
5. C. Ji et al., "Labeling images by integrating sparse multiple distance learning and semantic context modeling," in *Computer Vision-ECCV 2012*, pp. 688–701, Springer (2012).
6. O. A. Penatti et al., "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognit.* **47**(2), 705–720 (2014).
7. 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).
8. H. Bannour and C. Hudelot, "Hierarchical image annotation using semantic hierarchies," in *Proc. 21st ACM Int. Conf. on Information and Knowledge Management*, pp. 2431–2434, ACM (2012).
9. H. Bannour and C. Hudelot, "Building semantic hierarchies faithful to image semantics," in *Advances in Multimedia Modeling*, K. Schoeffmann et al., Eds., pp. 4–15, Springer Berlin, Heidelberg (2012).
10. L. Li-Jia et al., "Building and using a semantic visual image hierarchy," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3336–3343 (2010).
11. M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia* **16**(4), 1059–1074 (2014).
12. H. Tamura and N. Yokoya, "Image database systems: a survey," *Pattern Recognit.* **17**(1), 29–43 (1984).
13. A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.* **29**(8), 1233–1244 (1996).
14. G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.* **46**(1), 188–198 (2013).
15. S. Liapis and G. Tziritas, "Color and texture image retrieval using chromaticity histograms and wavelet frames," *IEEE Trans. Multimedia* **6**(5), 676–686 (2004).
16. T. Hou et al., "Bag-of-feature-graphs: a new paradigm for non-rigid shape retrieval," in *2012 21st Int. Conf. on Pattern Recognition (ICPR)*, pp. 1513–1516, IEEE (2012).
17. L. Nanni, M. Paci, and S. Brahnam, "Indirect immunofluorescence image classification using texture descriptors," *Expert Syst. Appl.* **41**(5), 2463–2471 (2014).
18. K.-S. Goh, E. Y. Chang, and B. Li, "Using one-class and two-class SVMs for multiclass image annotation," *IEEE Trans. Knowl. Data Eng.* **17**(10), 1333–1346 (2005).
19. P. Duygulu et al., "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary," in *Computer Vision—ECCV 2002*, A. Heyden et al., Eds., pp. 97–112, Springer Berlin, Heidelberg (2002).
20. 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).
21. J. Qin and N. H. Yung, "Scene categorization via contextual visual words," *Pattern Recognit.* **43**(5), 1874–1888 (2010).
22. N. M. Elfiky et al., "Discriminative compact pyramids for object and scene recognition," *Pattern Recognit.* **45**(4), 1627–1636 (2012).
23. F. Wang, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. 16th ACM Int. Conf. on Multimedia (MM'08)*, pp. 239–248, ACM, New York (2008).
24. Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. 6th ACM Int. Conf. on Image and Video Retrieval*, pp. 494–501, ACM (2007).
25. J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1487–1494, IEEE (2011).
26. J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'07)*, pp. 1–8, IEEE (2007).
27. S.-W. Choi, C. H. Lee, and I. K. Park, "Scene classification via hypergraph-based semantic attributes subnetworks identification," in *Computer Vision-ECCV*, pp. 361–376, Springer (2014).
28. Y. Chai et al., "Tricos: a tri-level class-discriminative co-segmentation method for image classification," in *Computer Vision-ECCV*, pp. 794–807, Springer (2012).
29. 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).
30. J. Wang et al., "Locality-constrained linear coding for image classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3367, IEEE (2010).
31. J. C. van Gemert et al., "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(7), 1271–1283 (2010).
32. M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–7 (2007).
33. J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 785–792, IEEE (2011).
34. 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).
35. A. W. Smeulders et al., "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), 1349–1380 (2000).
36. D. E. Millard et al., "Mind the semantic gap," in *Proc. Sixteenth ACM Conf. on Hypertext and Hypermedia (HYPERTEXT'05)*, pp. 54–62, ACM, New York (2005).
37. Y. Lu et al., "Constructing concept lexica with small semantic gaps," *IEEE Trans. Multimedia* **12**(4), 288–299 (2010).
38. Q. S. Zhuang, J. K. Feng, and H. Bao, "Measuring semantic gap: an information quantity perspective," in *5th IEEE Int. Conf. on Industrial Informatics*, pp. 669–674 (2007).
39. 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).
40. L. Saitta and J. D. Zucker, *Abstraction in Artificial Intelligence and Complex Systems*, 1st ed., Springer-Verlag, New York (2013).
41. A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(4), 712–727 (2008).
42. X. Li, L. Wang, and E. Sung, "Adaboost with SVM-based component classifiers," *Eng. Appl. Artif. Intell.* **21**(5), 785–795 (2008).
43. S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlations," *IEEE Computer Society Conf. in Computer Vision and Pattern Recognition*, Vol. 2, pp. 2033–2040, IEEE (2006).
44. 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).
45. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
46. R. E. Fan et al., "LIBLINEAR: a library for large linear classification," *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
47. L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.* **46**(1), 424–433 (2013).
48. Z. Ji et al., "Balance between object and background: object-enhanced features for scene image classification," *Neurocomputing* **120**, 15–23 (2013).
49. S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel SVMs," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 66–77 (2013).
50. H. Bilen, V. Namboodiri, and L. Van Gool, "Object and action classification with latent window parameters," *Int. J. Comput. Vision* **106**(3), 237–251 (2014).
51. S. McCann and D. G. Lowe, "Local naive Bayes nearest neighbor for image classification," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3650–3656 (2012).

**Zhipeng Ye** is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology. He received his master's degree in computer application technology from Harbin Institute of Technology in 2013. His research interest covers 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 interest covers image processing, video processing, pattern recognition, and design of VLSI 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 interest covers 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 interest covers pattern recognition, aerospace image processing, medical image processing, and machine learning.