



ELSEVIER

Contents lists available at ScienceDirect

Information Sciences

journal homepage: [www.elsevier.com/locate/ins](http://www.elsevier.com/locate/ins)

# Spatial adjacent bag of features with multiple superpixels for object segmentation and classification



Wenbing Tao <sup>a,e,\*</sup>, Yicong Zhou <sup>b</sup>, Liman Liu <sup>c,d</sup>, Kunqian Li <sup>a</sup>, Kun Sun <sup>a</sup>, Zhiguo Zhang <sup>a</sup>

<sup>a</sup> School of Automation, National Key Laboratory of Science & Technology on Multi-Spectral Information Processing, Ministry Key Laboratory for Image Processing and Intelligence Control, Huazhong University of Science and Technology, Wuhan 430074, China

<sup>b</sup> Department of Computer and Information Science, University of Macau, Macau, China

<sup>c</sup> School of Biomedical Engineering, South-Central University for Nationalities, Wuhan 430074, China

<sup>d</sup> National Engineering Research Center for E-learning, Central China Normal University, Wuhan 430073, China

<sup>e</sup> State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

## ARTICLE INFO

### Article history:

Received 18 September 2013

Received in revised form 9 May 2014

Accepted 19 May 2014

Available online 4 June 2014

### Keywords:

Spatial adjacent bag of features

Superpixel adjacent histogram

Object recognition

Multiple segmentations

## ABSTRACT

In the paper we present a new Spatial Adjacent Bag of Features (SABOF) model, in which the spatial information is effectively integrated into the traditional BOF model to enhance the scene and object recognition performance. The SABOF model chooses the frequency of each keyword and the largest frequency of its neighboring pairs to construct the feature histogram. Using the feature histogram whose dimension is only twice larger than that of the original BOF model, the SABOF model drastically enhances the discrimination performance. Combining the Superpixel Adjacent Histogram (SAH) Fulkerson et al., 2009 [12] with multiple segmentations Pantofaru et al., 2008 [33] and Russell et al., 2006 [36], the SABOF method effectively deals with the segmentation and classification of objects with different sizes. Changing the segmentation scale parameter to obtain multiple superpixel segmentations and correspondingly adjusting the neighbor parameters of the SAH method multiple classifiers are trained so that, the SABOF method can fuse multiple results of these classifiers to obtain better classification performance than any single classifier. The superpixel-based conditional random field (CRF) is used to further improve the classification performance. The experimental results of scene classification and of object recognition and localization on classical data sets demonstrate the performance of the proposed model and algorithm.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

Object localization and classification are both challenging tasks in the computer vision society, and are extremely important for image understanding. Recently, a number of attentions have been paid to solving these two problems in a unified framework. Sliding window approaches are of the successful object localization techniques [4,8,26,42]. Considering a sliding window around each pixel, these approaches apply a classifier function to find the best classification to fit the sliding window. They have been extensively used to detect the location of an object in an image. In [26] Blaschko et al. used the branch and bound method to search all possible subwindows in an image. In [47] Wei and Tao proposed an efficient

\* Corresponding author at: School of Automation, National Key Laboratory of Science & Technology on Multi-Spectral Information Processing, Ministry Key Laboratory for Image Processing and Intelligence Control, Huazhong University of Science and Technology, Wuhan 430074, China. Tel.: +86 27 87541924.

E-mail address: [wenbingtao@hust.edu.cn](mailto:wenbingtao@hust.edu.cn) (W. Tao).

histogram-based sliding window method that utilizes the spatial coherence of natural images and computes the objective function in an incremental manner. However in many cases, we want to perform the pixel-level object segmentation. Recently, some joint segmentation and classification methods [12,33,38] have been developed to integrate the segmentation and recognition into a unified framework and to automatically segment the image into several semantically meaningful regions where each region is labeled as a specific object class. Most of these approaches are based on the bottom-up local feature representation and often use the conditional random field [17,37] model to constrain the spatial consistency.

The classical Bag of Features (BOF) method represents an image with an orderless collection of local features and has been demonstrated to have impressive performance in object segmentation and classification [5,18,38]. However, due to the lack of information about the spatial structure of features, its descriptive ability is extremely limited. To overcome this, this paper proposes the Spatial Adjacent Bag of Features (SABOF) method to effectively integrate the spatial information for the pixel-level object segmentation and classification. To construct the feature histogram, the SABOF method considers not only the frequency of each keyword but also the frequency of every pair of keywords which are spatially neighboring. The frequency of each keyword and the largest frequency of its neighboring pairs are used to represent the feature histogram. Thus, using the feature histogram with the dimension just twice larger than that of the original BOF model, the SABOF method drastically enhances the discriminative power. Our experiments are provided to demonstrate that including more neighboring pairs to construct the feature histogram is not necessary and does not benefit the classification performance of the SABOF method.

Moreover, based on the proposed SABOF model, we integrate multiple segmentations with the Superpixel Adjacent Histogram (SAH) framework [12]. The SAH [12] was proposed to avoid the sparse features of each single superpixel and to provide context information learning. The quick shift algorithm [44] was used to extract superpixels from images with a fixed scale parameter. However, it was not explicitly stated how many adjacent superpixels would lead to the best performance in [12]. In our observation, the effect of the neighbor parameter  $N$  in SAH is closely related to the scale of superpixels. If each superpixel includes a small number of pixels, larger  $N$  would lead to better performance. On the other hand, if each superpixel includes a large number of pixels, a small value should be given to  $N$ . Therefore in this paper, considering the varied sizes of objects in images, we propose to fuse the changed parameter  $N$  with the multiple superpixel segmentations to enhance the adaptability and robustness of the SAH framework. Furthermore, more structure information of objects is obtained to enhance the performance of object segmentation and classification.

Multiple superpixel segmentations can be obtained by changing the scale parameter of the quick shift algorithm. See Fig. 1, from left to right, the scale parameter  $\sigma$  gradually becomes larger. For each superpixel result, a suitable neighbor parameter  $N$  is used in the SAH framework to construct a SABOF classifier. Thus, multiple SABOF classifiers are obtained combining multiple segmentation scale parameters with neighbor parameters of SAH. For testing images, multiple segmentations are provided and classified by the multiple SABOF classifiers. The multiple classifications of one image are combined to obtain the final segmentation and recognition result. Moreover, the superpixel-based conditional random field (CRF) is incorporated to further improve the segmentation performance.

The outline of the paper is as follows. In the next section we overview the related object segmentation and classification methods. In Section 3 we present the SABOF model. The SAH method with multiple segmentations is presented in Section 4. We validate our algorithm in Section 5 to show the advantages of our proposed model and algorithm, followed by a brief conclusion in Section 6.

## 2. Related works

Joint segmentation and classification have been studied in [6,9,24,32,38,48], where a global shape model is usually exploited and a unified framework integrates segmentation and recognition. They can efficiently classify only the highly structured objects but difficultly address the cases of severe occlusion and arbitrary viewpoints. Local features like textons [38,40] were usually applied for class segmentation algorithms to obtain pixel-level results. Moreover, a conditional random field or other spatial coherence constraint [37,43] was exploited to refine the results. Considering computational costs, some class segmentation algorithms operate on a reduced grid of the image to achieve a fast speed while sacrificing pixel accuracy. Other methods used superpixels [2,21,31,32,36] to increase the computational efficiency. The superpixels correspond to small regions obtained from an over-segmentation. Gould et al. [19] proposed a CRF to learn relative location offsets of categories based on superpixels. Fulkerson et al. [12] developed a classifier using histograms of local features based on superpixels. Multiple superpixel segmentations are often exploited to assist classification. In [36], Russell et al. used multiple segmentations to build a BOF model to discover and label object categories automatically. Similarly, Galleguillos et al. [21] localized objects using the multiple-instance learning in the weakly labeled data. Pantofaru et al. [33] integrated multiple

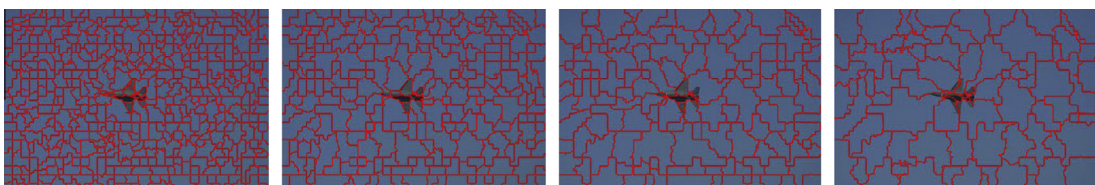


Fig. 1. Four segmentations using the quick shift algorithm with the scale parameter  $\sigma = 2, 4, 6$  and  $8$ , respectively.

segmentations to improve the robustness of object recognition within the BOF framework. All these methods assume that there should be at least a correct segmentation separating objects from their background.

However, the traditional BOF method has limitations in lack of the spatial structure information of features because the BOF method represents an image using an orderless collection of local features. Unfortunately, developing an effective feature description with the object structural information is always a challenging topic, especially in the presence of occlusion or large viewpoint changes. Many approaches have been developed to overcome the limitations of the traditional BOF model [25,39,46]. The generative part model-based method [14,23] and geometric correspondence search-based method [3,34,35,49] have achieved robust performance, but they suffer from the complexity of the models and low computational efficiency [1]. Spatial Pyramid Matching (SPM) [3,11,27] develops a kernel-based recognition method based on the pyramid matching scheme [22]. SPM computes local feature histograms at increasingly fine resolutions and has proved superior to the original BOF model. But SPM is difficult to be directly applied to the pixel-level object segmentation and classification. Zhang et al. [50] proposed a framework that encodes the spatial information into the inverted index by integrating the local adjacency of visual words. Zhang et al. [53] proposed the probabilistic graphlet for weakly supervised image segmentation. The graphlet as a graph with a small size of superpixels is used to capture the spatial structure of superpixels. In [54,55] the graphlet is also used for photo cropping and aerial image category recognition. Zheng et al. [51] proposed a high-level representation of visual words (called visual synset) by constructing an intermediate visual descriptor from a frequently co-occurring visual word-set. Herve and Boujmaa [20] used visual word pairs to describe the co-occurrence of neighboring visual words or other spatial relations. Zhang et al. [52] encoded more spatial information through the geometry-preserving visual phrases which incorporates the relative spatial locations of features with a visual phrase. Li et al. [29] proposed the contextual bag-of-words representation that integrates the semantic conceptual relation and spatial neighboring relation. Tirilly et al. [41] proposed a new image representation (called visual sentences) that considers the simple spatial relations between visual words and uses the probabilistic Latent Semantic Analysis (PLSA) to eliminate the noisiest visual words.

Although these methods have been demonstrated to have sufficiently improved performance compared with the traditional BOF model, other issues still remain such as overfitting, redundant features and low computational efficiency. In this paper, we propose a new SABOF model and integrate it with the multiple-segmentation SAH framework to achieve joint object localization and classification. The contributions are summarized here: 1) A new Spatial Adjacent Bag of Features (SABOF) model is proposed to effectively integrate the spatial information of features for object segmentation and classification. It uses both the frequency of each keyword and the largest frequency of the neighboring pairs to construct the feature histogram. 2) The multiple superpixels are integrated with the SAH framework [12] to enhance the adaptability and robustness of the algorithm to handle the objects with different scales.

### 3. Spatial adjacent bag of features

The dense scale-invariant feature transform (SIFT) descriptors [28] are extracted from each pixel in images using the “VLFeat toolbox” in [45]. Thus, each pixel corresponds to a SIFT descriptor. We combine the SIFT descriptors extracted from the training image set and use  $k$ -means to build the dictionary of the cluster centers, called keywords. Each SIFT descriptor is assigned to its closest keyword based on the Euclidean distance. Each superpixel in the training image or testing image is represented by a histogram of keywords,  $h(i)$ .

$$h(i) = \sum_{j \in S} \text{count}(S(j) = i), \quad i = 1, \dots, K. \quad (1)$$

where  $S$  denotes a set of the SIFT descriptors of a given object in the image, and  $K$  is the number of keywords.  $S(j)$  returns the keyword assigned to the SIFT descriptor  $j$ .

The traditional BOF method represents an image as an orderless collection of local features. Although the BOF method has achieved great success in object classification, it has been demonstrated that introducing the spatial correlation of keywords [3,49] is helpful to increase the recognition accuracy. Therefore, in this paper we propose a novel Spatial Adjacent Bag of Features (SABOF) to model the spatial configuration of keywords. In the new model, we count not only the number of occurrences of each keyword in the superpixel but also the number  $h(i, k)$  that the keyword  $k$  appears next to the keyword  $i$ .

$$h(i, k) = \sum_{j, l \in S, S(j)=1} \sum_{l \in N(j)} \varphi(S(l) - k). \quad (2)$$

where  $N(j)$  is the four-neighbor system of  $j$ , and  $\varphi(S(l) - k)$  returns 1 if  $S(l) = k$ , else returns 0. Especially, let  $h(i, i) = h(i)$ , where  $h(i)$  is defined in Formula (1). It is easy to find  $h(i, k)$  is a symmetric matrix. Therefore, we can use only the upper triangular data in the matrix  $h(i, k)$  to construct the feature histogram for classification.

Compared with the traditional BOF model, the SABOF model integrates the spatial relationship information among keywords and thus can represent objects more effectively. The cost is that the dimension of the feature histogram increases from  $K$  to  $K(K - 1)/2$ . If we set the number of keywords as 200, the dimension of the feature histogram using the SABOF model will be as high as 19,900. Therefore, we have to reduce the dimension of the feature histogram.

In this paper, we use a simple strategy to reduce the dimension of the histogram of the SABOF model without decreasing the number of keywords. Given a superpixel and the visual dictionary, in addition to the number of occurrences of each keyword  $i$  in the superpixel,  $h(i)$ , we count  $h_{\max}(i)$ , the maximum of  $h(i, k)$  which is defined as  $h_{\max}(i) = \max_{k \neq i} h(i, k)$ ,  $k = 1, \dots, K$  and  $k \neq i$ . Then the feature histogram can be written as

$$\{h(i), h_{\max}(i)\}, \quad i = 1, \dots, K. \tag{3}$$

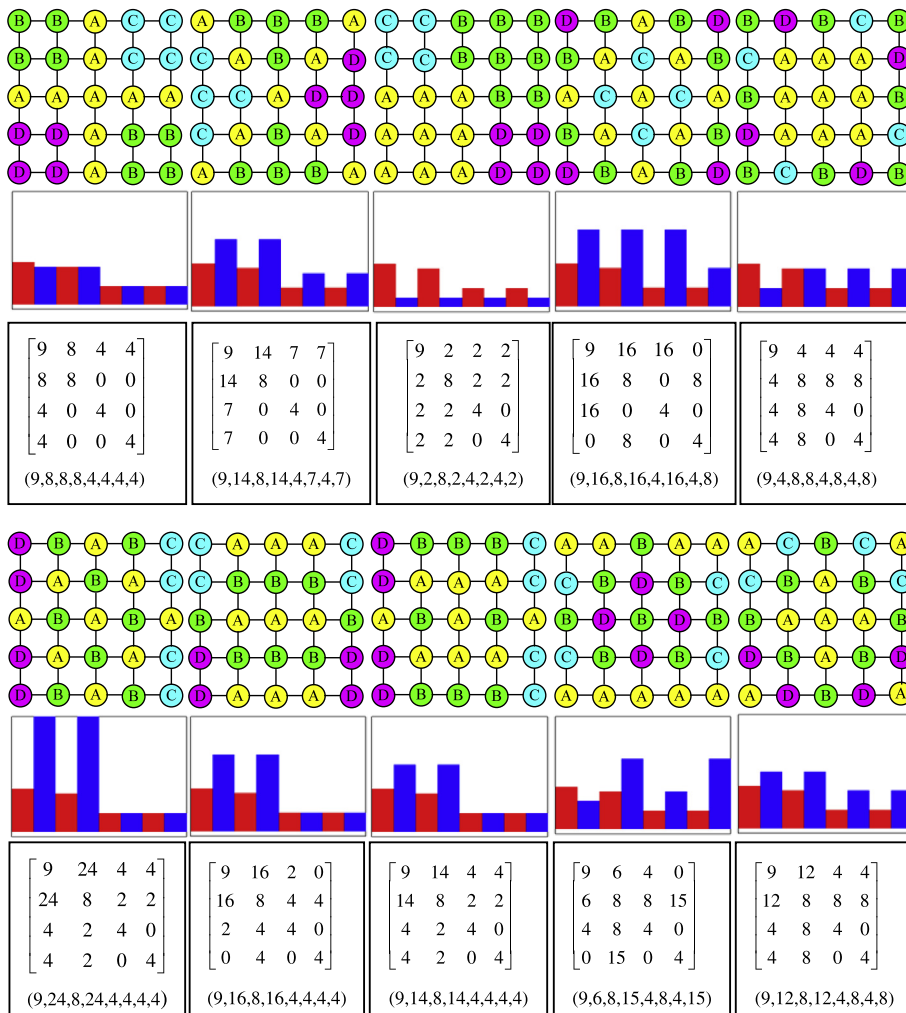
The dimension of the feature histogram in Formula (3) is just twice larger than that of the one in Formula (1). But the discriminative ability is greatly improved.

As illustrated in Fig. 2, we design ten superpixels which are mapped into four keywords A, B, C and D. Under each superpixel, the corresponding histogram, feature matrix and feature vector are shown. The feature matrices are obtained by Formula (2). The histograms and feature vectors are obtained by Formula (3). Using the traditional BOF model, all superpixels can be represented by the same feature histogram (9, 8, 4, 4) as shown by the red bins in each histogram in Fig. 2. This indicates that these superpixels are definitely same. But using the histograms obtained by the SABOF model in Formula (3), these superpixels can be effectively distinguished each other.

### 4. Superpixel adjacent histogram with multiple segmentations

#### 4.1. SAH classification

In our object classification and localization framework, we construct the SABOF classifier operating on superpixels. For each image, the dense SIFT features are extracted at the given orientation and scale using the “VLFeat toolbox” in [45]. The *k*-means method is used to quantize the extracted descriptors to produce the feature dictionary. These descriptors



**Fig. 2.** Comparison of the traditional BOF and SABOF models. In ten groups of experiments, ten superpixels are mapped into four keywords A, B, C and D. Each group of results includes one keyword mapping, one feature histogram by the SABOF, one feature matrix by Formula (2) and one feature vector by Formula (3). In each feature histogram, the red part corresponds to the histogram of the traditional BOF model. Ten feature histograms of the SABOF model are different, but feature histograms of the BOF model are completely same. This indicates that the SABOF model is more discriminative than the traditional BOF model.

are then aggregated into one  $l^1$ -normalized histogram for each superpixel. However, when learning a classifier, the feature histograms are usually sparse because each superpixel includes a small number of pixels. To address this problem, Fulkerson et al. [12] introduced a superpixel adjacent histogram. Firstly, the adjacency graph of superpixels in an image is constructed. Define  $G(S, E)$  as the adjacency graph of superpixels  $s_i \in S$  in an image. Let  $H_i^0$  be the histogram associated with the superpixel  $s_i$ .  $E$  is the edges connecting adjacent superpixels in the image and  $D(s_i, s_j)$  is the length of the shortest path between  $s_i$  and  $s_j$ . We define  $H_i^N = \sum_{s_j | D(s_i, s_j) \leq N} H_j^0$ , as the merged histogram combining the histograms of  $s_i$  and its neighbors whose distances to  $s_i$  are less than  $N$ . We use the same learning framework except that the superpixels are represented by  $h_i^N = H_i^N / \|H_i^N\|$ , where  $\|\cdot\|$  is the  $L_1$  norm. The histograms  $h_i^N$  with the neighborhood information include more abundant features and are less sparse. It also provides spatial consistency in classification and learns some context information.

#### 4.2. Integrating multiple segmentations

The combination of multiple segmentations and the SAH method is based on two principles. Firstly, because the size of superpixels is related to the scale parameter of the quick shift and the scales of objects in images vary in a large range, multiple segmentations are more helpful than single segmentation to extract enough suitable features to represent the objects with different scales. Also, the variation of regions in multiple segmentations can provide available information in different feature scales. Secondly, multiple classifications trained by the multiple superpixels can provide more information about the characteristics of each pixel. For example, if all classifiers give the highest probability of the “car” class to one pixel, the result that this pixel belongs to the “car” class is more reliable than using a single classifier. In general, we use the average probability of one pixel as the final probability indicating that the pixel belongs to one class.

Let  $I(x)$  be one image pixel, and  $P_i^j(I(x))$  be the probability of  $I(x)$  which belongs to the  $i$ th ( $i = 1, \dots, C$ ) class using the  $j$ th ( $j = 1, \dots, M$ ) classifier, where  $C$  is the number of the object classes and  $M$  is the number of the classifiers. The final class probability of pixel  $I(x)$  can be represented as

$$P_i(I(x)) = \frac{1}{M} \sum_{j=1}^M P_i^j(I(x)) \tag{4}$$

Similar to [12], we can also refine the result with the conditional random fields (CRFs) in order to reduce misclassifications and recover more precise boundaries. Let  $P(c|G; \omega)$  be the conditional probability of the class label  $c$ ,

$$-\log(P(c|G; \omega)) = \sum_{s_p \in S} \Psi(c_i | s_p) + \omega \sum_{(s_p, s_q) \in E} \Phi(c_i, c_j | s_p, s_q) \tag{5}$$

where  $G$  is the adjacency graph and  $\omega$  is the weights. The unary potentials  $\Psi$  can be approximately defined as the probability outputs provided by SVM implemented in LIBSVM [7] for each superpixel as in [12]:  $\Psi(c_i | s_p) = -\log(P(c_i | s_p))$ , and the pairwise potentials  $\Phi$  are defined as  $\Phi(c_i, c_j | s_p, s_q) = \left( \frac{L(s_p, s_q)}{1 + \|s_p - s_q\|} \right) [c_i \neq c_j]$ , where  $[\cdot]$  is the zero-one indicator function and  $\|s_p - s_q\|$  is the  $L_1$  norm of the color difference between superpixels in the LUV color space.  $L(s_p, s_q)$  is the regulation term to avoid small isolated regions [6], defined as the common boundary length between superpixels.

After refining with CRFs, each pixel is given a determinate class label instead of some class probabilities. Therefore, multiple class labels will be produced by multiple classifiers for each pixel. These labels may be different. The final label will be assigned by the majority class voting. If each classifier gives a different class label to a pixel, we will randomly select one as the final class label.

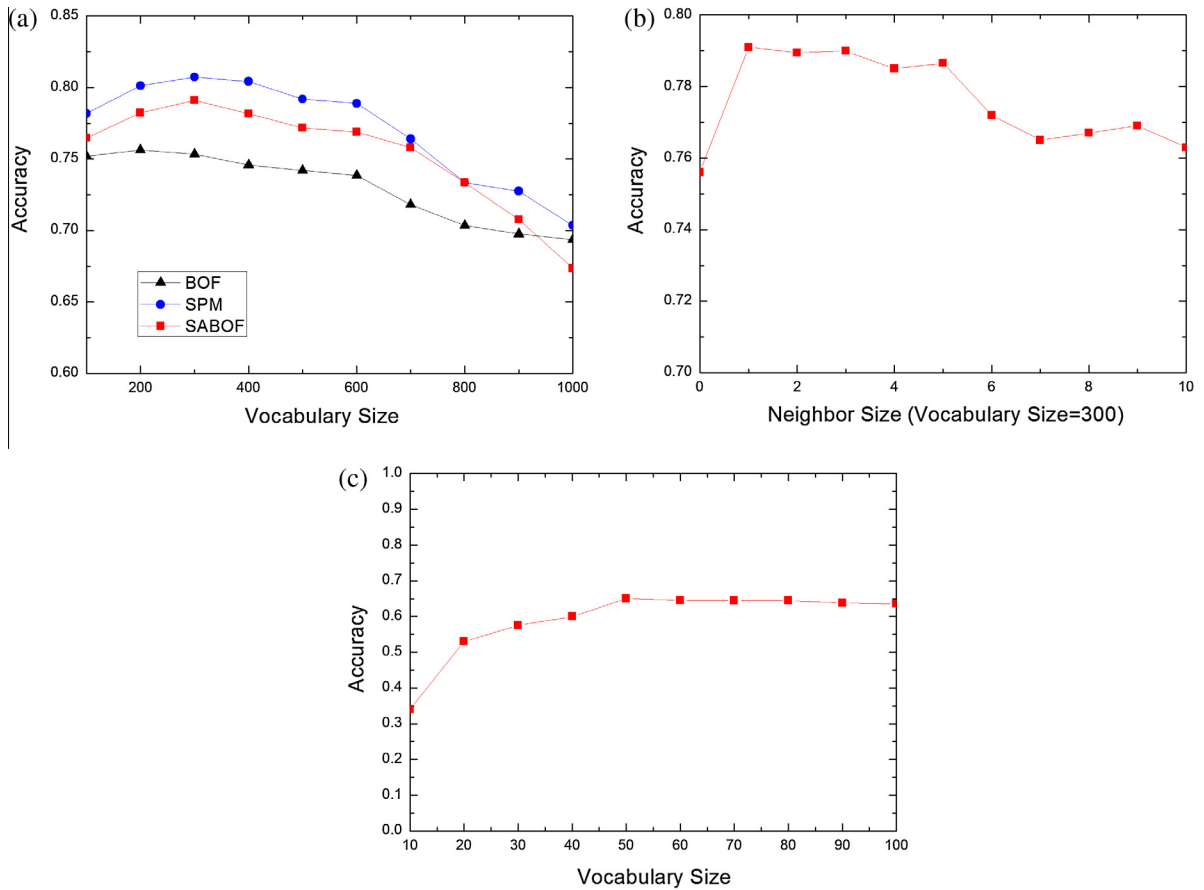
### 5. Experiments

#### 5.1. The SABOF model for scene classification

We first test the performance of our SABOF model on image scene classification. We evaluate the model on three challenging datasets. The first dataset used in [27] contains fifteen scene categories, and the other two are the Graz-02 dataset used in [13,30] and Caltech-101 dataset [15,16]. The dense SIFT descriptors are utilized for better discriminative ability. A visual vocabulary is produced by using the  $k$ -means clustering of a random subset from the training set.

Multi-class classification is done by a one-vs-rest SVM trained with an RBF- $X^2$  kernel. We use the SVM implemented in LIBSVM [7]. Each classifier is learned to partition each class from the rest. One test image will be given the label of the classifier with the highest response. In Figs. 3–5, using the dataset with fifteen scene categories, Graz-02 and Caltech-101 datasets, we compare the performance of the proposed SABOF model with those of the BOF and SPM models [27]. We also test the effect of the vocabulary and neighbor sizes in several cases.

In Figs. 3(a), 4(a) and 5(a), we compare the performance of the BOF, SABOF and SPM models as the vocabulary size changes. For the dataset with fifteen scene categories, the best results of the BOF, SABOF and SPM are 75.6% with the vocabulary size of 200, 79.1% with the vocabulary size of 300, and 80.9% with the vocabulary size of 300, respectively. For the Graz-02 dataset, the best results of the BOF, SABOF and SPM are 80.6% with the vocabulary size of 300, 84.7% with the vocabulary size of 400, and 85.2% with the vocabulary size of 200, respectively. For the Caltech-101 dataset, the best



**Fig. 3.** Experiments using the fifteen scenes dataset. (a) The classification accuracy comparison of the BOF, SABOF and SPM models [27] versus the vocabulary size. (b) The vertical ordinate denotes the classification accuracy and the horizontal ordinate denotes how many largest neighboring pairs are used in the feature histogram, for example, the proposed SABOF just uses one largest neighboring pair. (c) All the neighbor pairs are considered to construct feature histogram in the extreme case.

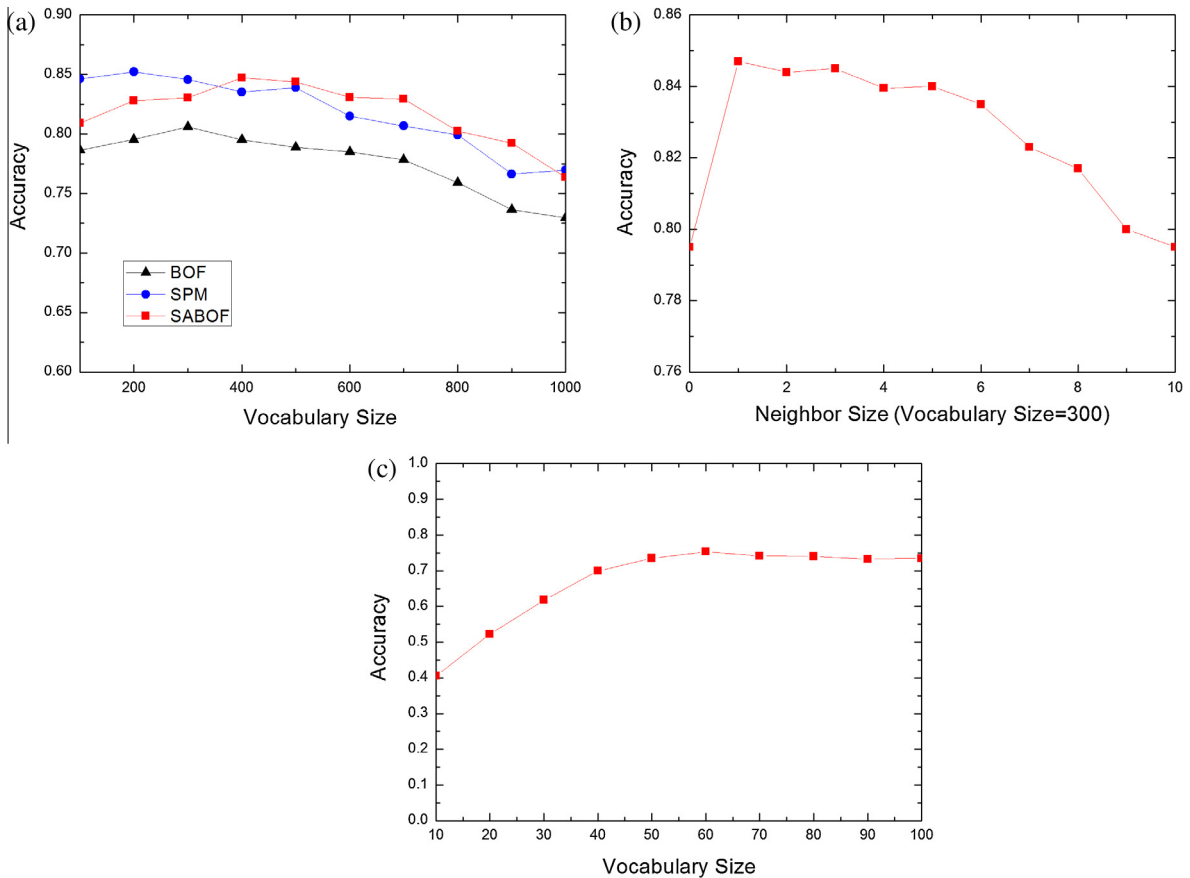
results of the BOF, SABOF and SPM are 55.2% with the vocabulary size of 200, 65.1% with the vocabulary size of 300, and 65.2% with the vocabulary size of 200, respectively.

From Fig. 3(a), we can see that, for the dataset with fifteen scene categories, the best result of the SABOF method is much better than that of the traditional BOF model but slightly lower than that of the SPM method. For the Graz-02 and Caltech-101 datasets, our results (Figs. 4(a) and 5(a)) are extremely close to those of the SPM method [27] and much better than that of the BOF model. However, the proposed SABOF method can be easily extended to the pixel-level segmentation to locate objects in the input image while the SPM method [27] does not. We will give the experimental results of the object segmentation and classification in the following section.

In the proposed SABOF model, we not only count the number of occurrences of each keyword in the superpixel, but also count the number  $h(i, k)$  that keyword  $k$  appears next to keyword  $i$ . Moreover, we simplify the SABOF model by considering only the largest neighbor pair. In (b) and (c) of Figs. 3–5, we test the effect of the SABOF model when considering more neighbor pairs. In Figs. 3(b), 4(b) and 5(b), we demonstrate the performance of SABOF when the size of the largest neighbor pairs increases from 0 to 10. From the results, we can find that the best result appears when the size of the largest neighbor pairs is one. If the size further increases, the accuracy does not increase (see Fig. 3(b)) or even greatly decreases (see Fig. 4(b)). In Figs. 3(c), 4(c) and 5(c), we test the extreme case where all neighbor pairs are used to construct the feature histogram. For  $K$  keywords, the dimension of the feature histogram will be  $K(K-1)/2$ . We can see that the accuracy increases as the number of keywords increases when the number of the keywords is small. While the number of keywords increases to about 60, the accuracy reaches its maximum (68.2% for the dataset with fifteen scene categories, 76.1% for the Graz-02 dataset and 51.4% for the Caltech-101 dataset). The accuracy does not change too much when the number of keywords is over 60.

## 5.2. The SABOF model integrating multiple segmentations in SAH framework for object classification

We further evaluate our object segmentation and classification algorithm based on the SABOF model on the PASCAL VOC 2007 segmentation competition dataset [10]. The PASCALVOC 2007 dataset includes 21 object categories. Each object class is

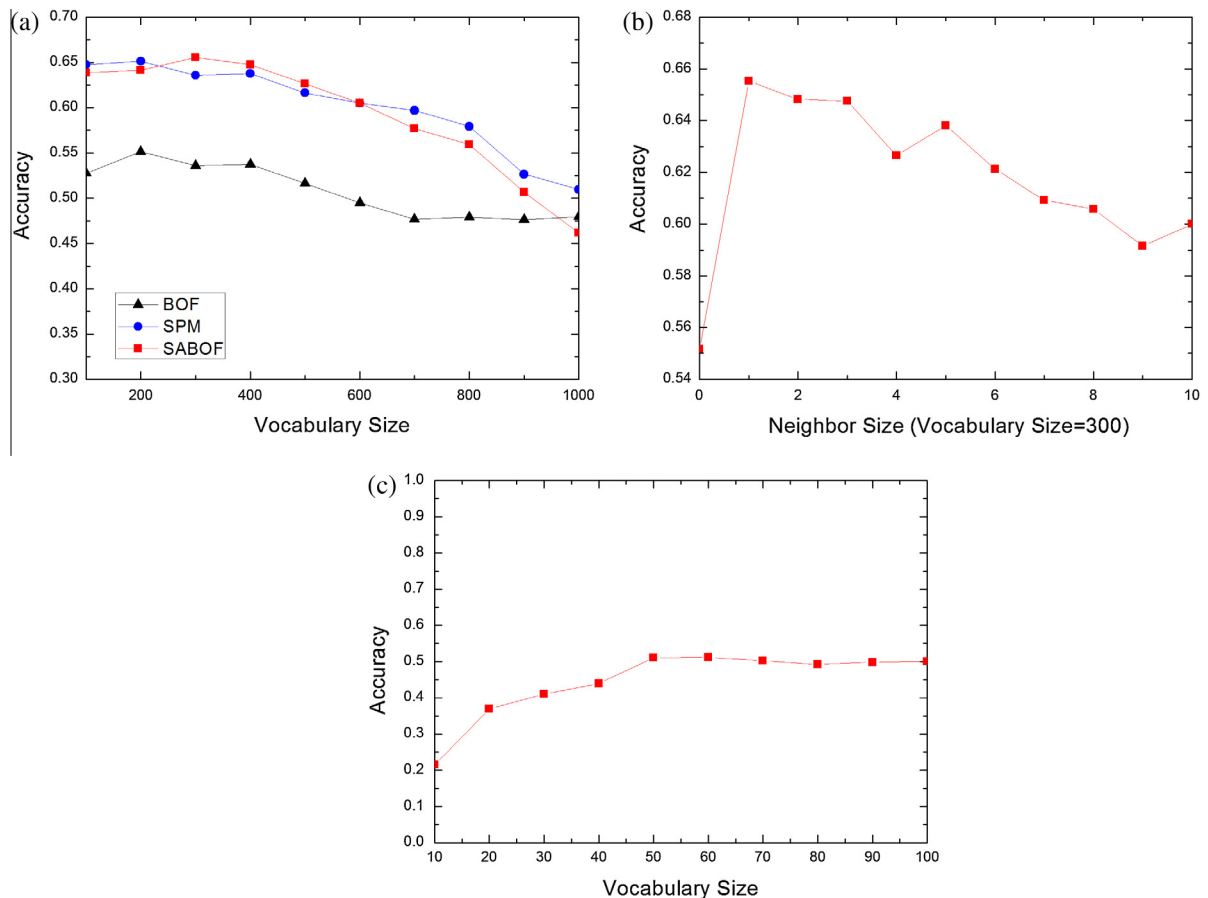


**Fig. 4.** Experiments using the Graz dataset. (a) The classification accuracy comparison of the BOF, SABOF and SPM models [27] versus the vocabulary size. (b) The vertical ordinate denotes the classification accuracy and the horizontal ordinate denotes how many largest neighboring pairs are used in the feature histogram, for example, the proposed SABOF just uses one largest neighboring pair. (c) All the neighbor pairs are considered to construct feature histogram in the extreme case.

with extreme variations in deformation, scale, illumination, pose, and occlusion. The ground truth segmentation is used for training. We use the average pixel accuracy to measure the performance of the algorithm. For each object class, the average pixel accuracy is defined as the ratio between the number of correctly classified pixels and the number of the ground truth pixels plus incorrectly classified pixels. The total percentage of pixels which are correctly classified is also reported.

Experiments should have many parameter settings. We first extract the dense SIFT features with a patch size of 12 pixels at a given orientation. These SIFT feature descriptors are quantized into the learned vocabulary dictionary from the training data. The number of vocabularies is set to 200 in our experiments. Thus, the number of keywords is 200 for the traditional BOF method and 400 for the SABOF model. The quick shift algorithm has three parameters  $\lambda$ ,  $\sigma$  and  $\tau$  to control the extraction of superpixels. More details can refer to [44]. We choose  $\lambda = 0.5$  and  $\tau = 8$ . The multiple segmentation results are obtained by adjusting the scale parameter  $\sigma$ . In this paper, we set  $\sigma = 2, 4, 6$  and  $8$  to produce four different segmentation results. Correspondingly, the histogram parameter  $N$  in the superpixel adjacent histogram (SAH) is set as 4, 3, 2, and 1, respectively. We randomly choose an equal number of feature histograms from each category to train SVM. Thus, we can obtain four different classifiers which correspond to a group of superpixel results and a SAH parameter  $N$ , respectively. For testing and comparison with other methods, we convert the superpixel labels into a pixel-labeled map and make the evaluation at the pixel level. The CRF model is used to refine the results of the SVM classifiers. For the PASCAL VOC 2007 dataset, we randomly select 250 training feature histograms from each class to train the SVM.

For simplicity, we call the SAH method based on the traditional BOF in [12] as SAH-BOF. If the CRF model is integrated, then call it as SAH-BOF-CRF. If the SABOF model is used instead of the BOF model, then call them as SAH-SABOF and SAH-SABOF-CRF, respectively. We compute the classification results of four SAH-BOF and four SAH-SABOF methods corresponding to four segmentation results, respectively. Then four SAH-BOF-CRF and SAH-SABOF-CRF classification results are obtained by refining the results using the CRF model. For four SAH-BOF and SAH-SABOF classification results, we can obtain the average classification results by Formula (4). For four SAH-BOF-CRF and SAH-SABOF-CRF classification results, we use the majority class voting method to assign the final pixel label.



**Fig. 5.** Experiments using the Caltech-101 dataset. (a) The classification accuracy comparison of the BOF, SABOF and SPM models[27] versus the vocabulary size. (b) The vertical ordinate denotes the classification accuracy and the horizontal ordinate denotes how many largest neighboring pairs are used in the feature histogram, for example, the proposed SABOF just uses one largest neighboring pair. (c) All the neighbor pairs are considered to construct feature histogram in the extreme case.

We apply SAH-BOF, SAH-BOF-CRF, SAH-SABOF and SAH-SABOF-CRF on four superpixel segmentation results, which are indicated as SAH1, 2, 3 and 4, SAHC1, 2, 3 and 4, SA1, 2, 3 and 4, and SAC1, 2, 3 and 4 in Table 1, respectively. The SAHA and SAA indicate the fusion results of four classifications produced by four segmentations based on BOF and SABOF, respectively. The SAHCV and SACV indicate the voting results of four classifications after refining with CRF.

The first five rows in Table 1 show four results by SAH-BOF and one fusion result. The results show that the fusion of multiple classifications can greatly improve the performance. The best and the worst results of the single classification are 52% and 47% respectively while the fusion result reaches to 55%. The second five rows show four results by SAH-BOF-CRF and one fusion result. The best and the worst results of the single classification are 57% and 53% respectively while the fusion result is 59%. The third five rows show four results by SAH-SABOF and one fusion result. The best and the worst results of the single classification are 58% and 53% respectively while the fusion result is 60%. The fourth five rows show four results by SAH-SABOF-CRF and one fusion result. The best and the worst results of the single classification are 61% and 58% respectively while the fusion result is 63%. All experimental results show that the fusion of multiple superpixel classifications is extremely helpful for improving the classification performance. Comparing the results by SAH-BOF (in the first five rows) with SAH-SABOF (in the third five rows), for all four classifications and the fusion result, the classification accuracy of the proposed SABOF model is higher than that of the traditional BOF model. The comparison also shows that CRF refining can remarkably increase the performance of object classification and localization no matter BOF or SABOF is being used. The final fusion result by the proposed SABOF model integrating multiple segmentations in the SAH framework is 63% shown in the last row. This result improves by about 6% compared with the best result 57% reported in [12].

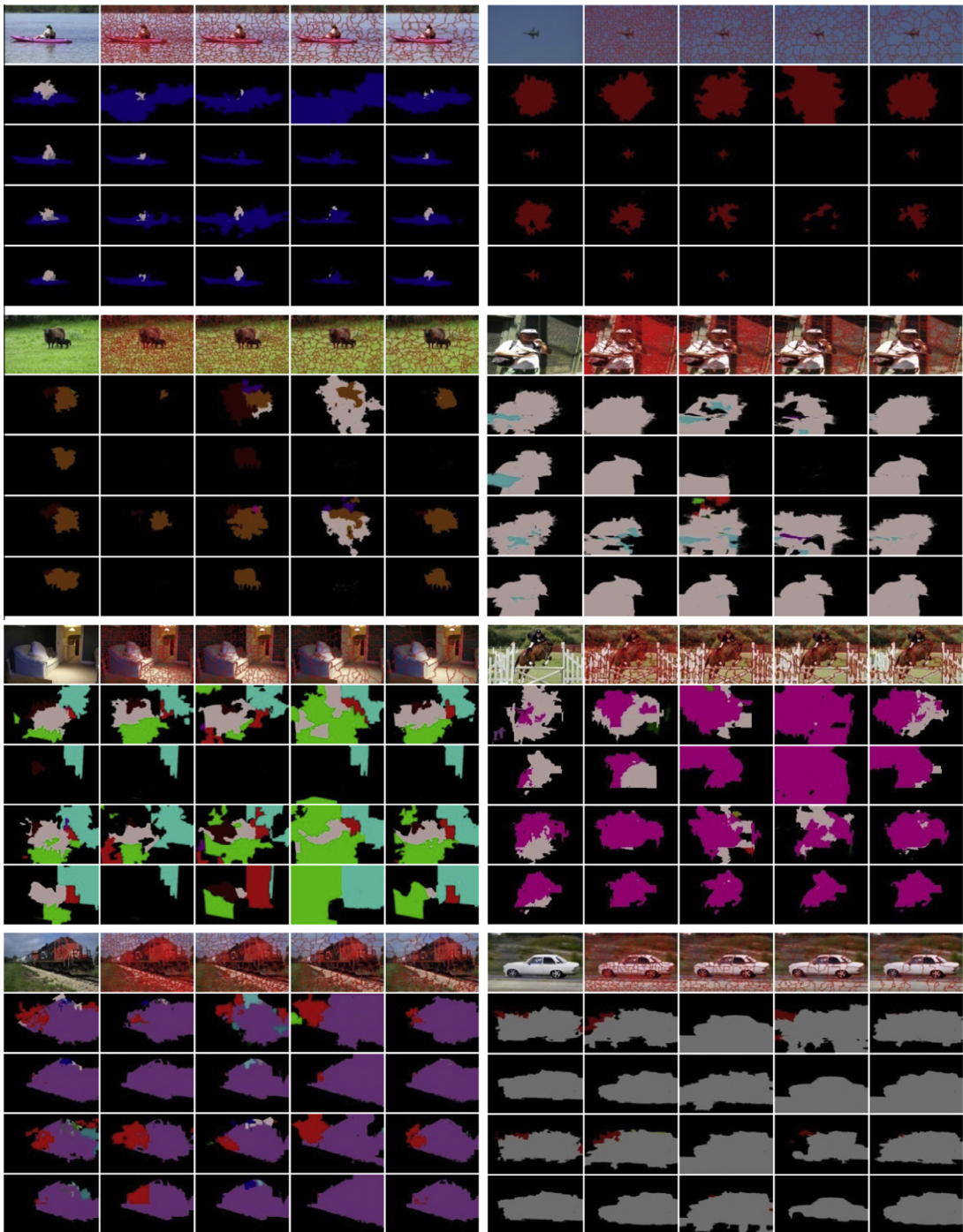
Selected localization examples are shown in Fig. 6. In each group of experiments, the first row shows the original images and their four segmentations, the second row shows four classifications corresponding to four segmentations and the final fusion results of four classifications by SAH-BOF, the third row shows four CRF refining results of the classifications in row 2 and the final voting results, the fourth row shows four classifications corresponding to four segmentations and the final fusion results of four classifications by SAH-SABOF, the fifth row shows four CRF refining results of the classifications in












**Table 1**

SAH-BOF, SAH-BOF-CRF, SAH-SABOF and SAH-SABOF-CRF are applied to four superpixel segmentation results, which are indicated as SAH1-SAH4, SAHC1-SAHC4, SA1-SA4, and SAC1-SAC4, respectively. SAHA and SAA indicate the fused results of four classifications produced by four segmentations based on BOF and SABOF, respectively. SAHCV and SACV indicate the voting results of four classifications based on BOF and SABOF after refining with CRF, respectively.

	Background	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tv monitor	Pixel (%)
SAH1	57	16	18	14	6	1	20	38	44	16	4	7	25	21	46	58	17	13	15	41	31	50
SAH2	60	13	18	10	6	1	15	28	42	15	2	8	28	19	53	63	15	9	13	44	36	52
SAH3	54	13	14	10	8	1	20	34	46	14	4	11	31	25	47	58	20	14	13	24	38	49
SAH4	48	6	17	13	8	1	21	31	39	23	6	8	29	16	36	56	20	14	16	47	42	47
SAHA	65	14	20	14	9	1	22	35	47	22	5	11	29	23	53	66	21	14	15	50	42	55
SAHC1	69	20	18	17	8	1	17	37	48	12	3	9	28	20	48	61	16	11	18	43	32	56
SAHC2	74	15	19	14	10	1	21	37	50	13	7	10	23	10	55	67	23	9	15	49	46	57
SAHC3	62	16	24	18	9	1	18	36	46	14	8	10	30	22	55	56	22	14	12	47	42	539
SAHC4	66	5	21	12	6	1	24	31	49	15	6	7	25	13	40	57	24	11	18	45	43	54
SAHCV	70	18	24	18	10	2	24	39	50	16	9	11	29	24	554	65	25	13	15	47	45	59
SA1	65	12	16	17	5	1	20	36	37	19	6	8	26	25	55	60	13	14	11	33	42	56
SA2	67	12	19	15	12	1	25	42	36	17	5	10	27	29	49	62	26	14	22	39	41	58
SA3	59	15	17	13	7	2	16	39	36	22	9	13	29	30	46	57	18	15	15	42	41	55
SA4	50	11	16	16	7	2	26	17	34	14	8	7	30	21	45	61	16	13	13	43	42	53
SAA	66	14	19	17	11	2	25	39	45	21	9	12	31	28	51	66	24	16	21	46	45	60
SAC1	73	16	25	17	6	1	22	36	46	19	5	6	27	25	49	60	22	18	19	36	38	60
SAC2	76	17	22	13	9	1	25	29	44	25	8	14	28	22	54	65	26	10	17	45	18	61
SAC3	69	12	21	15	11	1	24	32	49	24	7	6	21	27	55	65	25	11	23	45	29	60
SAC4	69	14	22	15	7	1	22	33	44	23	6	12	26	31	56	59	24	13	19	46	43	58
SACV	75	17	26	19	10	1	26	37	48	24	8	13	28	28	54	66	28	15	24	48	44	63



**Fig. 6.** Eight selected group results for PASCAL VOC 2007. In each group, the first row shows the original images and their four segmentations; the second row shows four classifications corresponding to four segmentations and the final fused results of four classifications by the SAH-BOF method; the third row shows four CRF refining results of the classifications in the second row and the final voting results; the fourth row shows four classifications corresponding to four segmentations and the final fused results of four classifications by the SAH-SABOF method; the fifth row shows four CRF refining results of the classifications in the fourth row and the final voting results. The color index table for all objects is as followed: 1. Plane, ; 2. Bike, ; 3. Bird, ; 4. Boat, ; 5. Bottle, ; 6. Bus, ; 7. Car, ; 8. Cat, ; 9. Chair, ; 10. Cow, ; 11. Table, ; 12. Dog, ; 13. Horse, ; 14. Motorbike, ; 15. Person, ; 16. Pottedplant, ; 17. Sheep, ; 18. Sofa, ; 19. Train, ; 20. Tv monitor, . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**

Comparison of the proposed SACV with several state-of-the-art methods.

	Background	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tv monitor	Average	Pixel (%)
[7]	59	27	1	8	2	1	32	14	14	4	8	32	9	24	15	81	11	26	1	28	17	20	
[19]	33	46	5	14	11	14	34	8	6	3	10	39	40	28	23	32	19	19	8	24	9	20	
[6]	65	20	30	22	2	2	39	25	57	10	3	7	36	23	66	62	15	17	8	46	11	27	57
SACV	75	17	26	19	10	1	26	37	48	24	8	13	28	28	54	66	28	15	24	48	44	30	63

row 4 and the final voting results. The results show that the SAH-SABOF model integrating multiple superpixels has the best performance.

In Table 2, we compare the results by the proposed method with those by the Shotton's method [40], Pantofaru's method [33], and Fulkerson's method [12]. The results of [33,40] and [12] come from Table 3 in [12]. Compared to [40] and [33], the average performance of the proposed method is improved by about 10%, and compared to the [12], the average performance of the proposed method is improved by about 3% and the percentage of pixels which were correctly classified is improved by about 7%. All the results demonstrate the performance of the proposed method.

## 6. Conclusion

This paper has developed an effective algorithm for object localization and classification based on the SABOF model, SAH framework and multiple segmentation cues. The SABOF model integrates the spatial information into the traditional BOF model for more powerful object discriminating capability. The fusion of multiple segmentations into the SAH framework makes it possible to handle the objects with the large size variation more effectively. The scene recognition and object classification experiments on the dataset with fifteen scene categories [27], Graz dataset [13] and PASCAL VOC 2007 segmentation competition dataset [10] have demonstrated the performance of the proposed model and algorithm.

## Acknowledgements

The research has been supported by the National Natural Science Foundation of China (Grants 61371140 and 61305044), and in part by Fundamental Research Funds for the Central Universities of China (HUST: 2014TS101), by Grant 20130144120004, by the Macau Science and Technology Development Fund under Grant 017/2012/A1, and by the Research Committee at University of Macau under Grants MYRG2014-00003-FST, MYRG113(Y1-L3)-FST12-ZYC and MRG001/ZYC/2013/FST.

## Reference

- [1] X. Ao, P. Luo, X. Ma, F. Zhuang, Q. He, Z. Shi, Z. Shen, Combining supervised and unsupervised models via unconstrained probabilistic embedding, *Inf. Sci.* 257 (2) (2014) 101–114.
- [2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (11) (2012) 2274–2281.
- [3] D. Aldavert, A. Ramisa, RL de Mantaras, Fast and robust object segmentation with the integral linear classifier, in: Proc. CVPR, 2010.
- [4] M. Blaschko, C. Lampert, Learning to localize objects with structured output regression, in: Proc. ECCV, 2008.
- [5] M.G. Baydogan, G. Runger, E. Tuv, A bag-of-features framework to classify time series, *IEEE Trans Pattern Anal Mach Intell* 35 (11) (2013) 2796–2802.
- [6] F. Chen, H. Yu, R. Hu, Shape sparse representation for joint object classification and segmentation, *IEEE Trans. Image Process.* 22 (3) (2013) 992–1004.
- [7] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. <<http://www.csie.ntu.edu.tw/~cjlin/libsvm>>.
- [8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. CVPR, 2005.
- [9] K. Engel, K.D. Toennies, Hierarchical vibrations for part-based recognition of complex objects, *Pattern Recogn.* 43 (8) (2010) 2681–2691.
- [10] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.
- [11] N.M. Elfiky, F.S. Khan, J. van de Weijer, J. González, Discriminative compact pyramids for object and scene recognition, *Pattern Recogn.* 45 (4) (2012) 1627–1636.
- [12] B. Fulkerson, A. Vedaldi, S. Soatto, Class Segmentation and Object Localization with Superpixel Neighborhoods, in: Proc. ICCV, 2009.
- [13] B. Fulkerson, A. Vedaldi, S. Soatto, Localizing objects with smart dictionaries, in: Proc. ECCV, 2008.
- [14] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: Proc. CVPR, 2003.
- [15] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proc. CVPR, 2005.
- [16] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, in: IEEE CVPR Workshop on Generative-Model Based Vision, 2004. <<http://www.vision.caltech.edu/ImageDatasets/Caltech101>>.
- [17] B. Glocker, O. Pauly, E. Konukoglu, A. Criminisi, Joint classification-regression forests for spatially structured multi-object segmentation, in: Proc. ECCV, 2012.
- [18] Y. Gao, B. Zheng, G. Chen, Qing Li, Chun Chen, Gang Chen, Efficient mutual nearest neighbor query processing for moving object trajectories, *Inf. Sci.* 180 (11) (2010) 2176–2195.
- [19] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, *Int. J. Comput. Vision* (2008).
- [20] N. Herve, N. Boujemaa, Visual word pairs for automatic image annotation, in: Proc. ICME, 2009.
- [21] C. Galleguillos, B. Babenko, A. Rabinovich, S. Belongie, Weakly supervised object localization with stable segmentations, in: Proc. ECCV, 2008.
- [22] K. Grauman, T. Darrell, Pyramid match kernels: discriminative classification with sets of image features, in: Proc. ICCV, 2005.
- [23] J. Guo, H. Zhou, C. Zhu, Cascaded classification of high resolution remote sensing images using multiple contexts, *Inf. Sci.* 221 (2) (2013) 84–97.
- [24] A. Ion, J. Carreira, C. Sminchisescu, Probabilistic joint image segmentation and labeling, in: Proc. NIPS, 2011.
- [25] Y.-G. Jiang, J. Yang, Chong-Wah Ngo, A.G. Hauptmann, Representations of keypoint-based semantic concept detection: a comprehensive study, *IEEE Trans. Multimedia* 12 (1) (2010) 42–53.
- [26] C. Lampert, M. Blaschko, T. Hofmann, Beyond sliding windows: object localization by efficient subwindow search, in: Proc. CVPR, 2008.
- [27] S. Lazebnik, C. Schmid, J. Ponce, Beyond bag of features: spatial pyramid matching for recognizing natural scene categories, in: Proc. CVPR, 2006.
- [28] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision* 2 (60) (2004) 91–110.
- [29] T. Li, T. Mei, I.S. Kweon, X.S. Hua, Contextual bag-of-words for visual categorization, *IEEE Trans. Circ. Syst. Video Technol.* 21 (4) (2011) 381–392.
- [30] M. Marszałek, C. Schmid, Accurate object localization with shape masks, in: Proc. CVPR, 2007.
- [31] A. Moore, S. Prince, J. Warrell, U. Mohammed, G. Jones, Superpixel lattices, in: Proc. CVPR, 2008.
- [32] O. Nempont, J. Atif, I. Bloch, A constraint propagation approach to structural model based image segmentation and recognition, *Inf. Sci.* 246 (10) (2013) 1–27.
- [33] C. Pantofaru, C. Schmid, M. Hebert, Object recognition by integrating multiple image segmentations, in: Proc. ECCV, 2008.
- [34] S. Park, S. Kim, M. Park, S.-K. Park, Vision-based global localization for mobile robots with hybrid maps of objects and spatial layouts, *Inf. Sci.* 179 (24) (2009) 4174–4198.

- [35] S.-H. Peng, D.-H. Kim, S.-L. Lee, C.-W. Chung, A visual shape descriptor using sectors and shape context of contour lines, *Inf. Sci.* 180 (16) (2010) 2925–2939.
- [36] B.C. Russell, A.A. Efros, J. Sivic, W.T. Freeman, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: *Proc. CVPR*, 2006.
- [37] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: *Proc. ICCV*, 2007.
- [38] J. Shotton, J. Winn, C. Rother, A. Criminisi, Texton-boost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: *Proc. ECCV*, 2006.
- [39] H.-H. Su, T.-W. Chen, Chieh-Chi Kao, W.H. Hsu, Shao-Yi Chien, Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features, *IEEE Trans. Multimedia* 14 (3) (2012) 833–843.
- [40] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: *Proc. CVPR*, 2008.
- [41] P. Tirilly, V. Claveau, P. Gros, Language modeling for bag-of-visual words image categorization, in: *Proc. CIVR*, 2008.
- [42] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, Multiple kernels for object detection, in: *Proc. ICCV*, 2009.
- [43] J. Verbeek, B. Triggs, Region classification with markov field aspect models, in: *Proc. CVPR*, 2007.
- [44] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: *Proc. ECCV*, 2008.
- [45] A. Vedaldi, B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms. <<http://www.vlfeat.org/index.html>>.
- [46] J. Wang, Y. Gong, Discovering image semantics in codebook derivative space, *IEEE Trans. Multimedia* 14 (4) (2012) 986–994.
- [47] Y. Wei, L. Tao, Efficient Histogram-Based Sliding Window, in: *Proc. CVPR*, 2010.
- [48] J. Winn, N. Jojic, LOCUS: Learning object classes with unsupervised segmentation, in: *Proc. ICCV*, 2005.
- [49] Y. Zhao, Y. Lu, Y. Tian, L. Li, Q. Ren, X. Chai, Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision, *Inf. Sci.* 180 (16) (2010) 2915–2924.
- [50] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrases for image applications. *ACM Int. Conf. on Multimedia*, 2009.
- [51] Y. Zheng, M. Zhao, S. Neo, T. Chua, Q. Tian, Visual synset: towards a higher-level visual representation, in: *Proc. CVPR*, 2008.
- [52] Y. Zhang, Z. Jia, Tsuhan Chen, Image retrieval with geometry-preserving visual phrases, in: *Proc. CVPR*, 2011.
- [53] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, C. Chen, Probabilistic graphlet cut: exploring spatial structure cue for weakly supervised image segmentation, in: *Proc. CVPR*, 2013.
- [54] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, C. Chen, Probabilistic graphlet transfer for photo cropping, *IEEE Trans. Image Process.* 21 (5) (2013) 2887–2897.
- [55] L. Zhang, Y. Han, Y. Yang, M.i. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, *IEEE Trans. Image Process.* 22 (12) (2013) 5071–5084.