Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **4**, July **2015**

Boosting Classifiers for Scene Category Recognition

Fu-Xiang Lu

School of Information Science & Engineering
Lanzhou University
222 Tianshui Road, Lanzhou, 730000, China
lufux@lzu.edu.cn

Jun Huang

Shanghai Advanced Research Institute
Chinese Academy of Sciences
99 Haike Road, Hi-Tech Park, Shanghai, 201210, China
huangj@sari.ac.cn

Kun Zhan

School of Information Science & Engineering
Lanzhou University
222 Tianshui Road, Lanzhou, 730000, China
kzhan@lzu.edu.cn

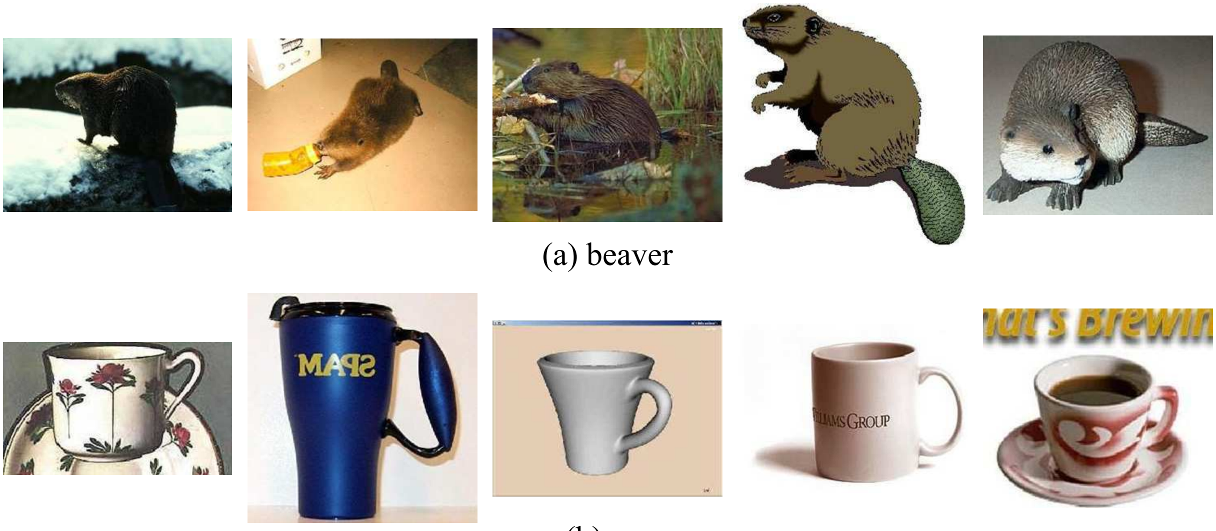
Received August, 2014; revised March, 2015

Abstract. *This paper presents a method for recognizing natural scene categories based not only on approximate global geometric correspondence between the features of the same kind, but also on complementary information cues offered by heterogeneous features. This technique works by dividing the image into increasingly fine sub-cells and computing the bag-of-features found inside each sub-cell. The discriminative power of each resulting spatial pyramid depends largely on its specific choices on interest point detector and local region descriptor involved in computing the bag-of-features. Different choices on interest point detector and local region descriptor lead to a powerful image representation: multiple pyramid histograms of words (mPHOW), which is a simple and computationally efficient extension of pyramid histogram of words (PHOW). In order to recognize an unknown image as correctly as possible, this paper first employs multi-class support vector machine (SVM) classifiers to compute posterior probabilities from the individual PHOWs, and then adopt the boosting algorithm to combine the variants of SVM, each trained on a single PHOW, to obtain the improved estimate of the "final" posterior probabilities. Our proposed method is evaluated on three benchmark scene datasets: OT, FP, and LSP. Results demonstrate that the proposed method outperforms the compared algorithms consistently.*

Keywords: Bag-of-words, Pyramid histogram of words, Support vector machine, Boosting.

1. **Introduction.** With the exponential growth on high quality digital images, the need of semantic scene category recognition is becoming increasingly important to support effective image database indexing and retrieval. However, the recognition of scene category, also called scene categorization, is one of the most challenging problems in computer vision, especially in the presence of intra-class variation, occlusion, clutter, pose and illumination changes.

Figure 1 shows five example images from each of two scene categories: *beaver* and *cup*. It is evident that each category presents high intra-class variation. Taking *cup* images in the second row of Figure 1 as an example, there exists significant variations in appearance, size, shape, color, and etc. This means that we need a method that can generalize across all possible instances of certain categories.



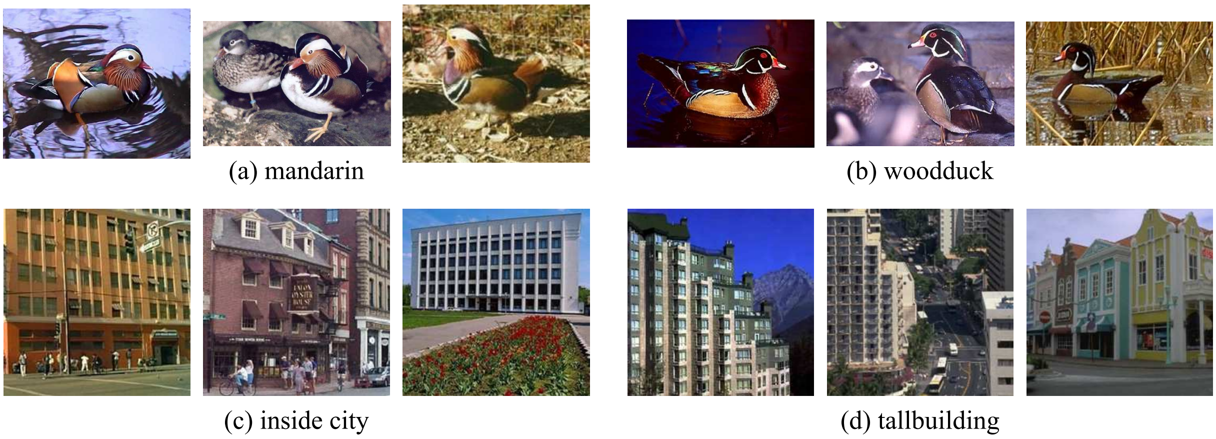
(a) beaver

(b) cup

Figure 1. Intra-class variation problem. Both *beaver* (a) and *cup* (b) present high intra-class variation. Note that in this paper all images in the figures are scaled so as to have same width while the aspect ratio remains unchanged.

Figure 2 shows three example images from each of four categories: *mandarin*, *woodduck*, *inside city*, and *tall building*. Although the *mandarin* and *woodduck* scenes are not labeled as the same category, we can see from Figure 2(a-b) that they would be easily confused each other. Also, *inside city* and *tall building* in Figure 2(c-d) can be easily confused each other because they have a very similar appearance. But for the case of scene category recognition, we do not want to confuse between scenes of different categories that are quite similar.

No one has yet constructed a scene categorization system which approaches the performance level of my two-year-old daughter. However, the progress in the field is quite dramatic, if judged by how much better today's algorithms are compared to those of a decade ago. In summary, most of the scene category recognition algorithms in the literature thus far are based on one of two basic image representations: *parts-and-shape* and *bag-of-features*. In the first category, the models are to represent real-world scenes by combining individual object appearance components with their spatial relations [1, 2]. The disadvantage of these models is that the learning and inference process is still extremely complex when the number of the constituent parts is beyond , say, 15 or 20, which greatly limits their applications. On the contrary, the bag-of-features models retain only the frequencies of the individual visual features and discard all information about their layout. The bag-of-features models have been proven to be effective for scene categorization [3, 4]. The success of these orderless models for scene categorization tasks may be explained with the help of an analogy to *bag-of-words* models for text document analysis, so these two terms, namely bag-of-words and bag-of-features, are used interchangeably throughout this paper. Currently, the bag-of-features models are the most popular methods for scene category recognition.



(a) mandarin

(b) woodduck

(c) inside city

(d) tallbuilding

Figure 2. Inter-class similarity problem. Three *mandarin* images (a) that could be easily confused by the three *woodduck* images (b). Three *inside city* images (c) are very similar with the *tallbuilding* images (d).

Inspired by pyramid histogram of words (PHOW) [5,6], this paper represents an image as multiple pyramid histograms of words (mPHOW), where the discriminative power of each PHOW depends on its specific choices on interest point detector and local region descriptor used in the framework of the bag-of-words. Since different interest point detectors and local region descriptors place their respective emphases on different aspects of a given image such as corner, texture, shape, and etc., such representation is expected to reflect the content of the image more comprehensively. Then, the multi-class support vector machine (SVM) classifiers, with kernels based on the histogram intersection distance (HID) for comparing visual word distributions, are learned from the individual PHOWs. Finally, in order to take advantage of the complementary information cues that seem to reside in the individual PHOWs, we adopt the boosting algorithm to obtain the improved estimate of the "final" *a posteriori* probabilities for an unknown image. Our proposed method is evaluated on three benchmark scene datasets: OT [7], FP [8], and LSP [5]. Experimental results demonstrate that the proposed method outperforms the compared algorithms on all benchmark datasets consistently.

The rest of this paper is organized as follows. Section 2 describe our method in detail. Experimental results on standard datasets are shown in Section 3. Section 4 makes a conclusion.

2. **Our method.** Figure 3 shows the basic stages involved in the design of our scene categorization system: mPHOW based image representation, recognition with the individual PHOWs, and decision-level classifier combination by the boosting. System evaluation stage is also included here for completeness, but it is not our focus in this work. In the following, Section 2.1 extends the PHOW to mPHOW based on different interest detectors and local region descriptors, Section 2.2 presents the design of multi-class SVM classifiers where each one is trained on a single PHOW, and at last Section 2.3 shows a decision-level classifier combination by the boosting algorithm.

2.1. **mPHOW.** Let $\mathbf{x}^m$ denote the PHOW of type m for an image I and $\mathbf{x}^m_{lc}$ denote the bag-of-words feature of the cell c at level l of the spatial pyramid (SP) constructed from the image I. Note that the first level of the SP consists of only one cell, i.e., the whole image itself. In each subsequent level each cell is split into four non-overlapping subcells. The process is repeated up to level L [5]. Computing the bag-of-words feature of the cell c at level l, $\mathbf{x}^m$, involves three main steps:

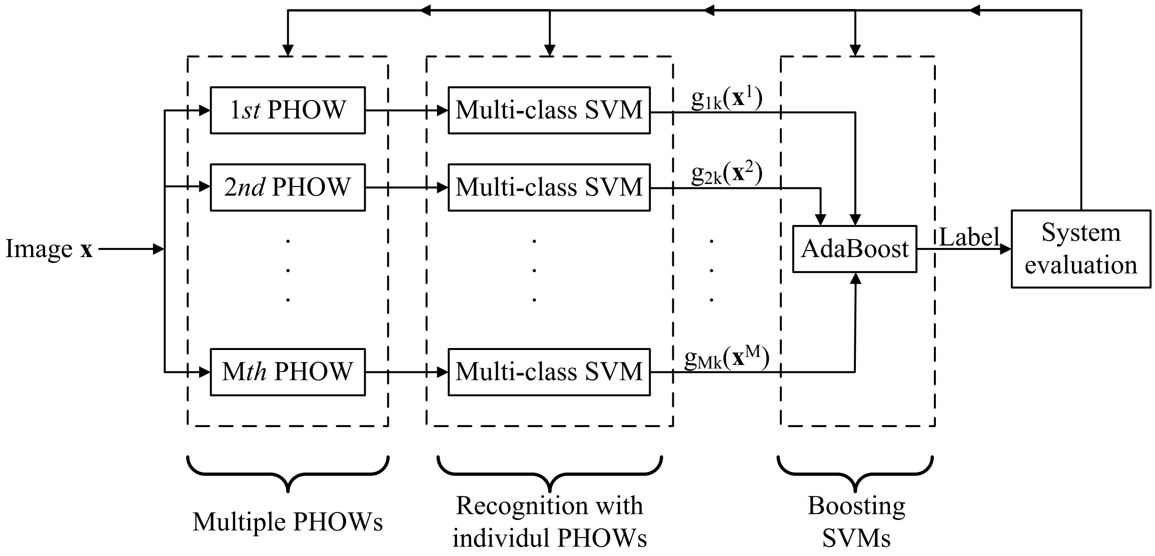


Figure 3. The basic stages involved in the design of our scene categorization system.

- Extract local region descriptors (either at interest points, or densely sampled).
- Generate a code book by K-means and vector quantize (VQ) descriptors to code words according to nearest neighbor (NN) rule.
- Encode the cell as a histogram of visual code words.

In this way, $\mathbf{x}^m$ is of dimensionality $D = V\sum_{l=1}^{L} 4^{l-1}$ for the spatial pyramid of L levels and the dictionary of V words. In order to capture multiple information cues resided in an image, we can apply different interest point detectors and local region descriptors to the framework of bag-of-words. Let us take for example local descriptors. To date there exist a number of descriptors in the literature, such as scale invariant feature transform (SIFT) [9], census transform (CT) [10] and self-similarity (SSIM) [11]. Therefore, selecting different interest point detectors and local region descriptors leads to multiple pyramid histograms of words (mPHOW), $\{\mathbf{x}^m\}$.

2.2. **Kernel-Based Recognition With the individual PHOWs.** Having computed the PHOW, $\mathbf{x}^m$, which is based on certain choices (indicated by the superscript, m, in $\mathbf{x}^m$) on interest point detector and local region descriptor, we use SVM for classification. In a two-class case, the decision function of SVM for a test image with the PHOW, $\mathbf{x}^m$, is of the following form:

$$g(\mathbf{x}^m) = \sum_{n=1}^{N} \alpha_n^m y_n \mathbf{K}(\mathbf{x}^m, \mathbf{x}_n^m) + b^m, \tag{1}$$

where $\mathbf{K}(\mathbf{x}^m, \mathbf{x}_n^m)$ is the value of a kernel function for the test image and the nth training image, y_n the class label of $\mathbf{x}_n^m$ (+1 or −1), α_n^m the learned weight of the nth training image, and b^m the learned threshold parameter.

In Eq. (1), we use extended Gaussian kernels:

$$\mathbf{K}(\mathbf{x}^m, \mathbf{x}_n^m) = \exp\left(-\frac{1}{\mu} d_{hi}(\mathbf{x}^m, \mathbf{x}_n^m)\right), \tag{2}$$

where

$$d_{hi}(\mathbf{x}^m, \mathbf{x}_n^m) = \sum_{i=1}^{D} \min(\mathbf{x}^m(i), \mathbf{x}_n^m(i)) \tag{3}$$

is the histogram intersection distance (HID) between $\mathbf{x}^m$ and $\mathbf{x}_n^m$, and μ is set to the average HID between all the training images.

LIBSVM [12], an integrated software for SVM, is used to train the two-class classifiers. Multi-class classification in LIBSVM is conducted following the *one-versus-one* strategy and LIBSVM is able to provide posterior probabilities at the outputs of the respective classifiers, i.e.,

$$g_k(\mathbf{x}^m) = P(\omega_k|\mathbf{x}^m), k = 1, 2, \dots, K, \tag{4}$$

which satisfies $g_k(\mathbf{x}^m) \geq 0$ and $\sum_{k=1}^{K} g_k(\mathbf{x}^m) = 1$, where K is the total number of classes. However, in this work, we conduct multi-class classification following the *one-versus-the-rest* strategy. The reason is that we have found the recognition accuracy rates of employing the one-versus-the-rest strategy are higher than those of employing the one-versus-one strategy according to our preliminary experiments. The steps involved in the *one-versus-the-rest* multi-class classification is summarized in Algorithm 1, where the superscript, m, is omitted in order to simplify the presentation.

Algorithm 1 *One-versus-the-rest* multi-class classifier

```
Require: Training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ with $y_n \in \{1, 2, \dots, K\}$, $n = 1, \dots, N$.
for $j = 1$ to $K$ do
    $\mathcal{D}_j^+ = \{(\mathbf{x}_n, y_n)|(\mathbf{x}_n, y_n) \in \mathcal{D}_{\text{train}}, \forall y_n = j\}$
    $\mathcal{D}_j^- = \{(\mathbf{x}_n, y_n)|(\mathbf{x}_n, y_n) \in \mathcal{D}_{\text{train}}, \forall y_n \neq j\}$
    Learn a two-class SVM and obtain a discrimination function
        $g_j(\mathbf{x}) = \sum_{n=1}^{N} \alpha_{jn} y_n \mathbf{K}(\mathbf{x}, \mathbf{x}_n) + b_j.$
end for
Assign $\mathbf{x}$ in $\omega_i$ if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ if $\forall j \neq i$.
```

2.3. **Combination By the Boosting.** Following the procedure given in Section 2.2, we have obtained $M \times K$ binary SVMs $g_{mk}(\mathbf{x}^m)$, $m = 1, \dots, M$ and $k = 1, \dots, K$, where M is the total number of PHOWs in mPHOW and K is the total number of scene classes. Now let $\mathbf{x}_n^m$ be the mth PHOW computed from the nth training image. We can obtain a $M \times K$ matrix, $\mathbf{G}_n = (g_{mk}(\mathbf{x}_n^m))$, where the (m, k) element is the response of $g_{mk}(\mathbf{x})$ for $\mathbf{x}_n^m$. Then, the new feature $\mathbf{t}_n$ for the nth training image is generated by lexicographic ordering of the elements of $\mathbf{G}_n$. With the reformulated training set $\{(\mathbf{t}_1, y_1), (\mathbf{t}_2, y_2), \dots, (\mathbf{t}_N, y_N)\}$, and once again, we first concentrate on the two-class classification. The goal of AdaBoost [13] (adaptive boosting, the most popular algorithm of the boosting family) is to construct an optimally designed classifier of the form

$$f(\mathbf{t}) = \text{sign}(F(\mathbf{t})) \tag{5}$$

where

$$F(\mathbf{t}) = \sum_{t=1}^{T} \alpha_t \phi(\mathbf{t}, k_t, p_t, \theta_t). \tag{6}$$

In Eq. (6), the base classifier $\phi(\mathbf{t}, k, p, \theta)$ is defined as

$$\phi(\mathbf{t}, k, p, \theta) = \begin{cases} +1 & \text{if } p[\mathbf{t}]_k < p\theta, \\ -1 & \text{otherwise} \end{cases} \tag{7}$$

Algorithm 2 The Adaboost approach to combine classifiers

```
Initialize $w_n^{(0)} = \frac{1}{N}, n = 1, 2, \ldots, N$
$t = 1$
repeat
    Compute
        $$(k_t, p_t, \theta_t) = \arg\min_{k,p,\theta} \left\{ \sum_{n=1}^{N} w_n^{(t)} I(1 - y_n \phi(\mathbf{t}_n, k, p, \theta)) \right\}$$
    $P_t = \sum w_n^{(t)}$ for $y_n \phi(\mathbf{t}_n, k_t, p_t, \theta_t) < 0$
    $\alpha_t = \frac{1}{2} \ln \frac{1-P_t}{P_t}$
    $Z_t = 0$
    for $n = 1$ to $N$ do
        $w_n^{(t+1)} = w_n^{(t)} \exp\left(-y_n \alpha_t \phi(\mathbf{t}_n, k_t, p_t, \theta_t)\right)$
        $Z_t = Z_t + w_n^{(t+1)}$
    end for
    for $n = 1$ to $N$ do
        $w_n^{(t+1)} = w_n^{(t+1)} / Z_t$
    end for
    $T = t$
    $t = t + 1$
until A termination criterion is met.
return $f(\mathbf{t}) = \text{sign} \sum_{t=1}^{T} \alpha_t \phi(\mathbf{t}, k_t, p_t, \theta_t)$.
```

where $[\mathbf{t}]_k$ is the kth element of the vector $\mathbf{t}$, p is a polarity indicating the direction of the inequality, and θ is a threshold. The weak classifier of Eq. (7) is usually called *decision stump* in the machine learning field. Algorithm 2 shows the pseudocode of the AdaBoost algorithm used in this paper. Note that in Algorithm 2, Function $I(\cdot)$ is either 0 or 1, depending on its argument, whether it is zero or positive, respectively. To obtain a classifier response, we use the raw outputs of the AdaBoost, given by Eq. (6). Multi-class classification is then done with the *one-versus-the-rest* rule. Specifically, an AdaBoost classifier is learned to separate each class from the rest, and a testing image is assigned the label of the classifier with the highest response.

3. **Experimental Results.**

3.1. **Datasets and Protocols.** We evaluated our scene categorization algorithm over three benchmark datasets:

1. Oliva and Torralba [7],
2. Fei-Fei and Perona [8], and
3. Lazebnik et al. [5]

We will refer to these datasets as OT, FP, and LSP, respectively. Figure 4 shows some example images from these three datasets. Here, we give an brief overview of these datasets.

- **OT**. It includes 2688 images classified as eight categories: 360 *coast*, 328 *forest*, 374 *mountain*, 410 *open country*, 260 *highway*, 308 *inside of city*, 356 *tall building*, and 292 *street.* The average size of each image is 250 × 250 pixels.
- **FP**. It is composed of thirteen scene categories and is only available in gray scale. It consists 2688 images of the OT dataset plus: 241 *suburb residence*, 174 *bedroom*,

151 *kitchen*, 289 *living room*, and 216 *office*. The average size of each image is approximately 300 × 250 pixels.

- **LSP**. It is composed of fifteen categories and, as with FP, is only available in gray scale. It consists of thirteen scene categories of the FP dataset plus: 315 *store* and 311 *industrial*. The average size of each image is approximately 300 × 250 pixels.

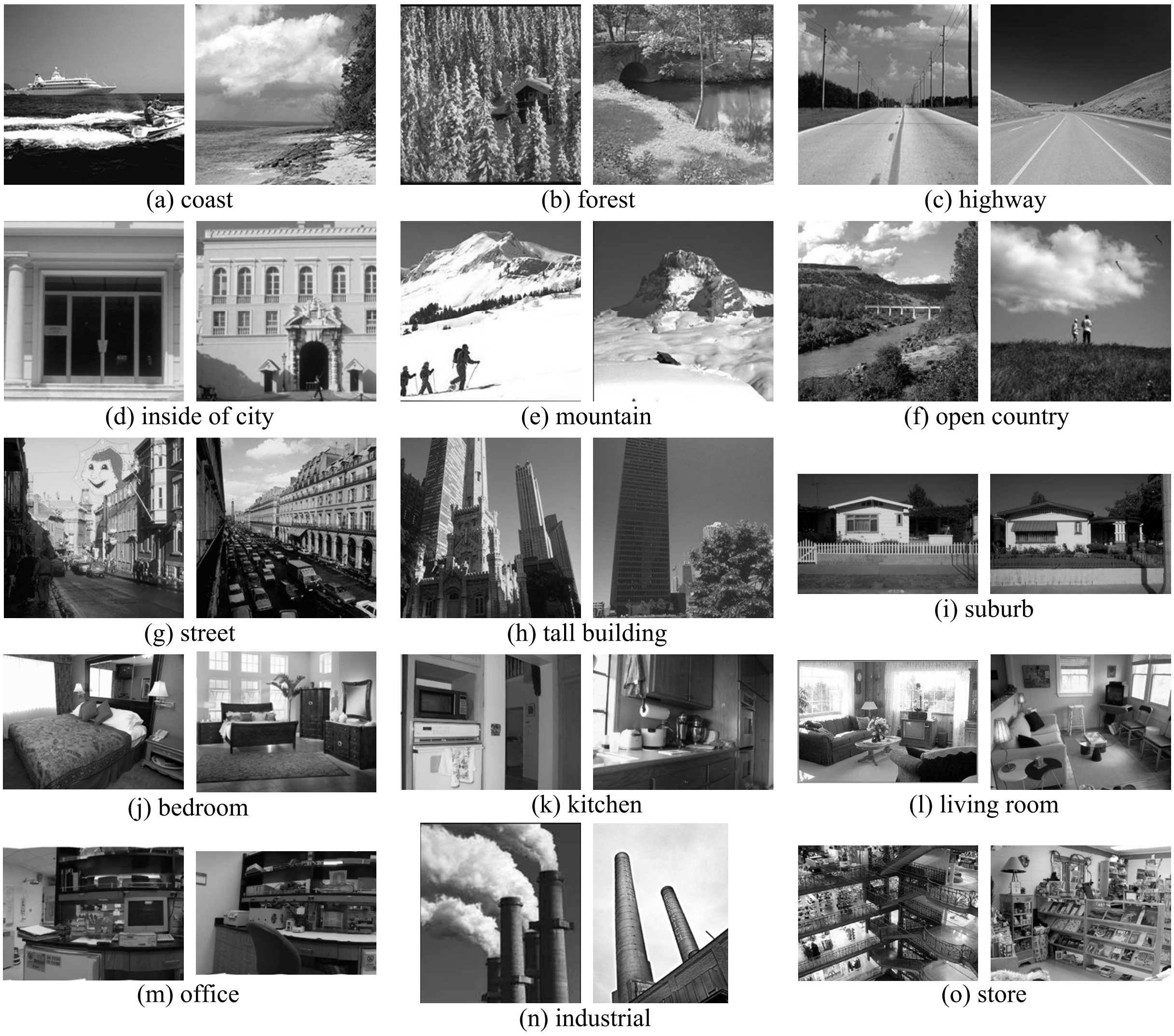

(a) coast (b) forest (c) highway (d) inside of city (e) mountain (f) open country (g) street (h) tall building (i) suburb (j) bedroom (k) kitchen (l) living room (m) office (n) industrial (o) store

Figure 4. Example images from the OT (a)–(h), FP (a)–(m) and LSP (a)–(o) datasets.

We perform all processing in grayscale, even when color images are available. In each dataset, the available data are randomly split into a training set and a testing set based on published protocols on these datasets. The random splitting is repeated five times, and the average recognition rate is reported for each run. The final results are reported as the mean and standard deviation of the results from the individual runs.

3.2. **Implemental Details.** SIFT, SSIM, CENTRIST and original normalized pixel patch are computed at points on a regular grid with spacing 8 pixels. At each grid point all descriptors are computed over a 16×16 pixel patch. This leads to four PHOWs. The size of all code books is empirically set to 200 based on Lazebniks' work [5]. Spatial pyramids of depth $L = 3$ are used for all experiments.

3.3. **Results.** Table 1 lists the results of the proposed method. For comparison, the results obtained by the individual PHOWs are also shown.

Table 1. The results of our proposed method and obtained by the individual PHOWs.

Datasets	OT	FP	LSP
Patch	82.6 ± 0.3	79.3 ± 0.5	75.9 ± 0.6
SIFT	87.2 ± 0.3	84.6 ± 0.4	80.2 ± 0.3
SSIM	87.3 ± 0.5	81.2 ± 0.7	77.8 ± 0.4
CENTRIST	84.3 ± 0.6	82.3 ± 0.4	79.7 ± 0.5
Boosting	$\mathbf{88.4} \pm 0.3$	$\mathbf{87.6} \pm 0.4$	$\mathbf{85.5} \pm 0.4$

Over all three scene datasets, it is seen from Table 1 that the proposed algorithm consistently outperforms the recognition methods of only using single PHOW. For example, our algorithm achieves a higher recognition rate 85.5% compared to that obtained by the best single feature, i.e. the PHOW with SIFT being local descriptor. This does not come as a surprise since our method exploits multiple information cues that may be complementary for discriminating the scene categories.

Figure 5 shows the resulting confusion matrix on LSP from the second run of using our proposed method. A closer look at the confusion matrix reveals that the most heavy confusion occurs among the four indoor categories: *bedroom*, *living room*, *kitchen* and *office*. This can be explained by the fact that there exists similar components (e.g., windows, tables, and so on) and similar configuration.

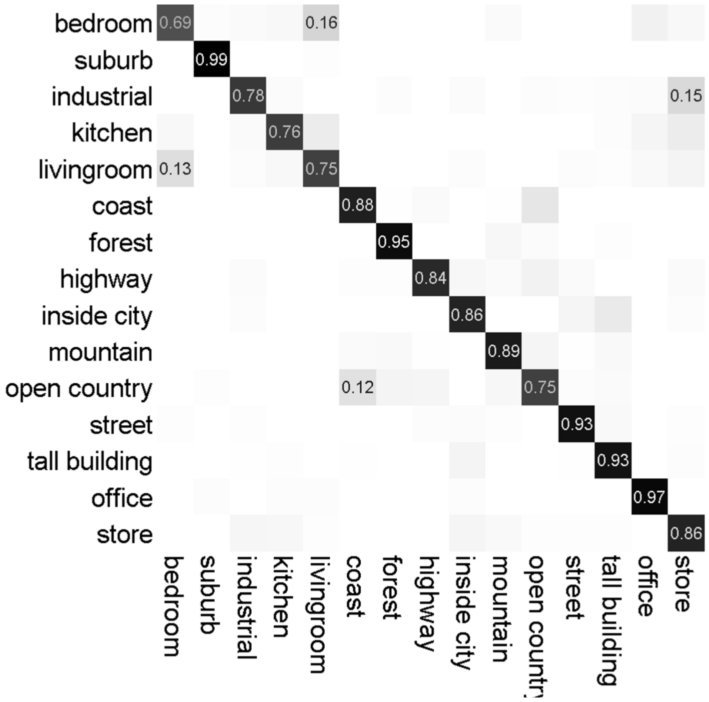


Figure 5. Confusion matrix resulted from our algorithm on LSP. The recognition rate is 85.5%. Only accuracies higher than 10% are shown.

Table 2 compares the proposed method with state-of-the-art ones over the benchmark datasets. For the OT dataset, the proposed method obtains an accuracy rate of 88.4%, which is much higher than the result of 83.7%, achieved with GIST [7]. For the FP dataset, the variant of latent Dirichlet allocation (LDA) obtained an accuracy of 65.2% [8].

Classification accuracy for the bag-of-visterms representation was 66.5% [3]. In our previous work, we proposed a powerful image representation: pyramid histogram of topics (PHOTO), which introduced approximate global geometric correspondence between the topics learned by probabilistic latent semantic analysis (pLSA). PHOTO gave 75.0% accuracy. The proposed method in this paper achieves the highest accuracy rate, 87.6%. Finally, for the LSP dataset, the standard pLSA achieves a classification rate of 65.9% [5] and spatial pyramid matching yields an accuracy of 81.1% [5]. Both Wu [10] and Liu [15] reported an 83.8% accuracy, which is also below the result of 85.5% achieved by the propose method in this paper.

Table 2. Comparison with state-of-the-art methods on the OT, FP and LSP datasets.

OT		FP		LSP	
Ours	88.8	Ours	87.6	Ours	85.5
[7]	83.7	[8]	65.2	[5]	81.1
		[3]	66.5	[10]	83.3
		[14]	75.0	[15]	83.3

4. Conclusion.

Based on various choices of interest point detectors and local region descriptors, this paper generalizes pyramid histogram of words (PHOW) to multiple pyramid histograms of words (mPHOW), which are shown to succeed in catching different information cues for scene category recognition. Given an unknown image, in order to recognize its class label as correctly as possible, this paper firstly employs support vector machine classifiers to compute posterior probabilities from the individual PHOWs, and then adopt the boosting algorithm to combine the variants of SVM, each trained on a single PHOW. Experimental results on several benchmark scene datasets have shown that the proposed algorithm leads to an overall better performance than could be achieved by using each of the variants of the same classifier separately.

Obviously, the boosting is not the only classifier combination algorithm available. For example, recently many experts in the field of machine learning developed the multiple kernel learning to combine different types of kernels in the framework of support vector machine, and obtained the satisfactory results on some datasets. However, from the practical point of view, the multiple kernel learning is computationally prohibitive in practical applications, and therefore, is not suitable for large scale dataset scenarios. Our future work is to conduct experimental comparative studies, and furthermore, to make what is a strategy that one has to adopt for combining the individual outputs in order to reach the final conclusion.

Acknowledgment. This work was partially supported by the Fundamental Research Funds for the Central Universities of China (No. lzujbky-2013-41 and lzujbky-2013-38) and the Project NSFC 61201446. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. Amores, N. Sebe and P. Radeva, Context-based object-class recognition and retrieval by generalized correlograms, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1818–1833, 2007.

[2] P. F. Felzenszwalb, R. B. Girshichk, D. McAllester and D. Ramanan, Object detection with discriminatively trained part based models, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] P. Quelhas, F. Monay, J-M. Odobez, D. G-Perez and T. Tyytelaars, A thousand words in a scene, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[4] J. Zhang, M. Marszalek, S. Lazebnik and C. Schmid, Local feature and kernels for classification of texture and object categories: A comprehensive study, *International Journel of Computer Vision*, vol. 73, no. 2, pp.213–238, 2007.

[5] S. Lazebnik, C. Schmid and J. Ponce, Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2169–2178, 2006.

[6] F. Lu, X. Yang, W. Lin, R. Zhang and S. Yu, Image classification with multiple feature channels, *Optical Engineering*, vol. 50, no. 5, pp. 057210, 2011.

[7] A. Oliva and A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[8] L. Fei-Fei, R. Fergus and P. Perona, "A bayesian hierarchical model for learning natural scene categories, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.524–531, 2005.

[9] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol. 60, no. 2, pp.91–110, 2004.

[10] J. Wu and J. M. Rehg, CENTRIST: A visual descriptor for scene categorization, *IEEE Transactons on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1502, 2011.

[11] E. Shechtman and M. Irani, Matching local self-similarities across images and videos, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2007.

[12] Chih-Chuang Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, *Software available at* `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

[13] P. Viola and M. J. Jones, Robust real-time face detection, *International Journel of Computer Vision*, vol. 57, no. 2, pp.137–154, 2004.

[14] F. Lu, X. Yang, R. Zhang and S. Yu, Image classification based on pyramid histogram of topics, *Proc. of IEEE Conference on Multimedia and Expo*, pp. 398-401, 2009.

[15] J. Liu and M. Shah, Scene modeling using co-clustering, *Proc. of IEEE Conference on Computer Vision*, pp. 1-7, 2007.