Spatial encoding of visual words for image classification

Dong Liu,[a,b] Shengsheng Wang,[a,b] and Fatih Porikli[c,d,*]
[a]Jilin University, College of Computer Science and Technology, Changchun 130012, China
[b]Jilin University, Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Changchun 130012, China
[c]Australian National University, Research School of Engineering, RSISE Building, Canberra 2601, ACT, Australia
[d]Data61, 7 London Circuit, Canberra 2601, ACT, Australia

Abstract. Appearance-based bag-of-visual words (BoVW) models are employed to represent the frequency of a vocabulary of local features in an image. Due to their versatility, they are widely popular, although they ignore the underlying spatial context and relationships among the features. Here, we present a unified representation that enhances BoVWs with explicit local and global structure models. Three aspects of our method should be noted in comparison to the previous approaches. First, we use a local structure feature that encodes the spatial attributes between a pair of points in a discriminative fashion using class-label information. We introduce a bag-of-structural words (BoSW) model for the given image set and describe each image with this model on its coarsely sampled relevant keypoints. We then combine the codebook histograms of BoVW and BoSW to train a classifier. Rigorous experimental evaluations on four benchmark data sets demonstrate that the unified representation outperforms the conventional models and compares favorably to more sophisticated scene classification techniques. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.3.033008]

Keywords: visual descriptors; bag-of-words; spatial feature representations; scene classification.

Paper 15926L received Jan. 7, 2016; accepted for publication May 4, 2016; published online May 31, 2016.

1 Introduction

Image classification is one of the primary tasks for visual understanding, with a wide variety of applications. Most image classification algorithms incorporate bag-of-visual words (BoVW) models since they are simple, robust to certain affine transformations, and capable of providing a uniform feature space irrespective of the number of visual word detections. Nevertheless, BoVW models have also several drawbacks, including the lack of spatial composition of visual words in the model, absence of structural cues, and poor interpretation of context for image classification.

Many techniques have been proposed to model spatial context for BoVW. Relative positions of codewords are taken into account in Ref. 1 for generative models. Lazebnik et al.[2] introduced a spatial pyramid matching model, in which the histograms of local features found inside subregions are concatenated for constructing spatial structures. Correlogram features[3] are proposed to capture spatial co-occurrences of features. Niu et al.[4] developed a discriminative latent Dirichlet allocation model to capture two types of contextual information, global spatial layout, and visual coherence in uniform local regions. Bolovinou et al.[5] presented a spatio-visual descriptor to encode ordered spatial configurations of visual words. Shaohua and Aggarwal[6] used a topic model based on a mixed membership stochastic model, in which the latent topics of adjacent visual words are jointly generated. Xie et al.[7] utilized a fused edge-scale invariant feature transform (SIFT) descriptor at the descriptor extraction level. Although these methods have been shown to provide certain solutions, they may also tend to overfit data, incur high computational complexity, be limited to predefined grids, and pertain to local structure.

Intuitively, contextual information, i.e., the spatial relationships among image features, provides crucial feedback in disambiguating visual words. This subsequently leads to improved recognition performance. To this end, we propose a method to encode spatial information of visual words both globally and locally as shown in Fig. 1.

In comparison to state-of-the-art BoVW models, our method provides several enhancements. First, we analyze two types of BoVW models, where images are sampled from a dense grid and a set of sparse interest points, respectively. For both models, we encode both global and local spatial representations of the visual words as midlevel features after a common codebook is trained and interest point descriptors are quantized onto the visual vocabulary. For the local spatial representation, we determine pairwise interest point cliques by associating the relevant points and compute a structure feature for each clique to describe the spatial relationships between the points of the clique. From these structure features, we learn a bag-of-structural words (BoSW) model that provides structural information across interest points. Finally, we concatenate the dense BoVW model and the BoSW model to train a potent support vector machine (SVM) classifier. To summarize, our contribution is threefold:

- BoSW model using structure features,
- unified representation of local and global information, and
- classification framework built on top of BoVW/BoSW.

2 Spatial Encoding of Visual Words

2.1 *Bag-of-Visual Words*

Suppose, for a given image data set $\{I_m\}$, we extract a set of keypoints $\{p_n\}$ from all images and compute the corresponding descriptors f_n for the keypoints. The descriptor

*Address all correspondence to: Fatih Porikli, E-mail: fatih.porikli@anu.edu.au

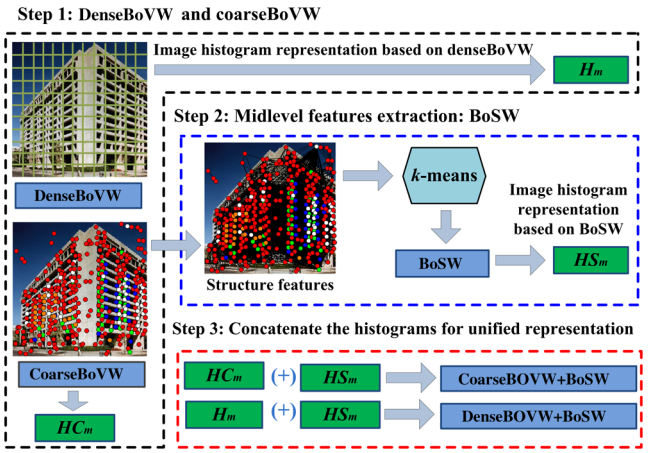

Fig. 1 Overview of the unified BoVW and BoSW representations.

can be, for instance, the histogram of oriented gradients within the local support. By k-means clustering these descriptors into d words with 4-neighbor soft weighting,[8] we construct a BoVW vocabulary $V_V = \{v_1, \ldots, v_d\}$. Then, we use this vocabulary to obtain an encoding F_m of each image

$$F_m = [(p_n, v_n)]_{p_n \in I_m}, \tag{1}$$

where v_n is the visual word assigned to the descriptor f_n of the keypoint p_n in I_m.

Each image I_m is further represented by a d-dimensional histogram $H_m = [h_{m,1}, \ldots, h_{m,d}]$ of the words, where $h_{m,k}$ is the bin corresponding the visual word k. Keypoints $\{p_n\}$ can be sampled on a grid to construct a denseBoVW model, or selected from sparse interest points, e.g., using SIFT keypoints, to build a coarseBoVW model.

2.2 *Structure Features*

We introduce a feature $s_{i,j}$ that encodes the spatial attributes of the edge e_{ij} between a pair of relevant keypoints p_i and p_j, which are selected from the coarse interest points. Two keypoints are considered relevant if they satisfy one of the following conditions:

1. the spatial distance between them is small: $\delta(p_i, p_j) < \xi$, and
2. both v_i and v_j belong to the most frequent k words.

That is, we select the short edges, and also include the edges between the points that have the assigned visual words, which rank within the top K most frequent words for the specific input image (instead of the entire image set). An example is shown in Fig. 2, where the keypoints of short edges are marked in red and the most frequent $K = 4$ visual words are marked in green, white, blue, and orange, respectively.

For each edge e_{ij}, we compute a feature vector of spatial attributes. The first component of this feature characterizes the orientation θ_{ij} of the edge. To provide an efficient representation that allows us computing orientation dissimilarity using vector operations and circumventing the circular ambiguity in orientation computation, we adopt a polar histogram Θ_{ij}. Around the central point of the e_{ij}, we quantize the polar space into B bins. The orientation of the edge votes on the corresponding histogram bins of Θ_{ij} with a Gaussian distribution $N(\theta_{ij}, \sigma)$ centered on the bin of the edge orientation θ_{ij} in a circular fashion, e.g., the last and the first bins are considered consecutive. Typical parameters for this histogram are $\sigma = 5$ and $B = 18$.

The orientation histogram does not contain the information on the word labels v_i and v_j of the edge points p_i and p_j. To this end, we analyze the contribution factor for each visual word to the image classes. Let $\{I_m\}$ contains C classes. An image in the c'th class is represented by the BoVW histogram $H_m^c = [h_{m,1}^c, \ldots, h_{m,d}^c]$. Suppose N_c denotes the number of images in the c'th class. Then, the mean histogram of visual words for the c'th class is denoted as H_μ^c

$$H_\mu^c = \frac{1}{N_c}\sum_{m=1}^{N_c} H_m^c = \frac{1}{N_c}\left[\sum_{m=1}^{N_c} h_{m,1}^c, \ldots, \sum_{m=1}^{N_c} h_{m,d}^c\right] = [\mu_1^c, \ldots, \mu_d^c], \tag{2}$$

where μ_k^c denotes the mean occurrences of the k'th visual word in the c'th class, which represents the contribution of that visual word to the corresponding class. Then, the contribution vector of the k'th visual word to all classes is obtained as

$$C(k) = [\mu_k^1, \ldots, \mu_k^C], \quad k = 1, \ldots, d. \tag{3}$$

For an edge e_{ij} of a pair of relevant points p_i and p_j with the visual word labels v_i and v_j, the final structure feature s_{ij} is composed by combining the orientation histogram, the contribution vectors $C(v_i)$ and $C(v_j)$

$$s_{ij} = [\Theta_{ij}; C(v_i) + C(v_j)], \tag{4}$$

which is symmetric in the order of points, i.e., $s_{ij} = s_{ji}$. For a data set with 10 classes, the dimension of s_{ij} is $18 + 10$.

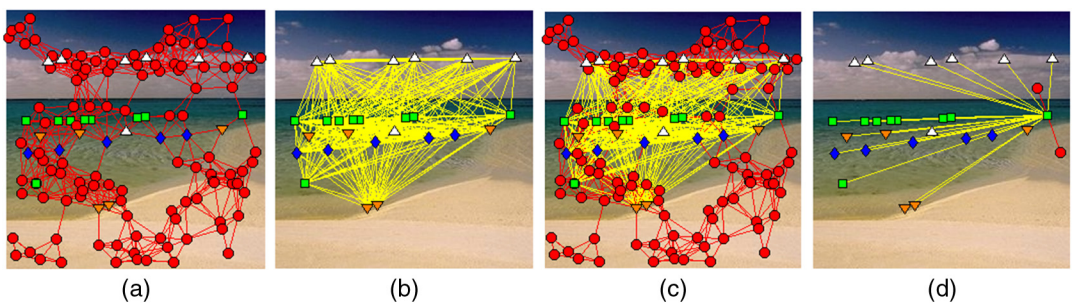

Fig. 2 (a) Relevant points obtained by distance cutoff, (b) relevant points obtained from the most frequent visual words, (c) all relevant points, and (d) relevant points of a point (the rightmost green point).

Similar to the appearance counterpart, we cluster the structure features s_{ij} of all images in the data set by k-means clustering to construct the BoSW vocabulary $V_S = \{s_1, \ldots, s_t\}$. Then, we use the BoSW vocabulary to obtain an encoding S_m of each image

$$S_m = [(s_{ij}, e_{ij})]_{e_{ij} \in \text{relevant keypoints}}. \quad (5)$$

Each image I_m is then represented by a t-dimensional structure histogram HS_m.

2.3 *Image Representation*

As shown in Fig. 1, a given image I_m is sampled densely (on a grid) and coarsely (on keypoints), which results in denseBoVW and coarseBoVW models, respectively. We then concatenate the encoding of BoSW with the encodings of denseBoVW and coarseBoVW models, respectively. For instance, the image representation of the denseBoVW+BoSW version concatenates the denseBoVW encoding H_m of the image I_m with the BoSW encoding HS_m to construct the final descriptor. Note that feature representations of the BoSW in "BoSW," "coarseBoVW+BoSW," and "denseBoVW+BoSW" are equivalent, yet the later two models employ both appearance and structure features, while "BoSW" uses only the structure features.

3 Experimental Results

We evaluated the performance of our methods on four large-scale data sets, i.e., MSRC-9[9] (9 categories), LabelMe[10] (8 categories), UIUC-Sports[11] (8 categories), and the LULC data set[12] (21 categories), depicting natural scenes, aerial orthoimagery images, and complex events. For the sake of consistency with reference works, we employ the published protocols on these data sets. For the LULC and MSRC-9 data sets, we report the mean classification accuracy upon a five-fold cross-validation setting for all categories following Refs. 13 and 14. Similar to Refs. 4 and 6, we randomly select 100 images per category for training and 100 for testing on the LabelMe data set. For the UIUC-Sports data set, we randomly select 70 and 60 images per class for training and testing, respectively.

3.1 *Bag-of-Structural Words Variants*

We first analyzed the effect of the kernel on the classification performance. As the base classifier, we tested SVM with linear, radial bases function (RBF), sigmoid, and histogram intersection (HI) kernels on the MSRC-9 data set. The results are shown in Table 1. For consistency, we trained the BoSW variants using the same codebook size of coarseBoVW; e.g., the base BoVS codebook size is $d = 300$ for both variants. Our results indicated that the HI kernel provides the best classification performance for both variants. Based on this observation, we used the HI for the following evaluations.

Table 1 Classification accuracy for different SVM kernels on MSRC-9 data set. Base codebook size is $d = 300$.

Methods	Linear (%)	RBF (%)	Sigmoid (%)	HI (%)
BoSW	67.04	70.07	72.96	**73.70**
DenseBoVW+BoSW	82.22	87.41	84.81	**90.00**

Note: Bold values indicate the best results.

Next, we examined the classification accuracy against the different base codebook sizes. The results of different variants of our method for different base codebook sizes on four data sets are summarized in Table 2. We observe

1. The accuracy of the conventional coarseBoVW is inferior to other choices. In addition, increasing the base codebook size causes performance degradation. A striking outcome is that learning the representation on a dense grid, i.e., denseBoVW, achieves better classification accuracy than the prevailing bag-of-word reconstruction on keypoints, i.e., coarseBoVW. However, the improvement quickly saturates for larger base codebook sizes.

Table 2 Classification accuracy of baseline and our methods with different base codebook sizes on various data sets.

Data sets	Methods	100 (%)	200 (%)	400 (%)	1024 (%)
MSRC-9	CoarseBoVW	57.41	63.70	68.51	67.78
	DenseBoVW	79.62	83.33	85.50	85.91
	BoSW	59.62	71.85	77.78	95.56
	CoarseBoVW.BoSW	60.00	73.33	82.22	97.03
	DenseBoVW+BoSW	**81.85**	**87.41**	**92.96**	**97.78**
LabelMe	CoarseBoVW	59.22	62.83	66.96	68.19
	DenseBoVW	82.63	82.91	85.30	86.18
	BoSW	55.02	62.02	69.75	75.29
	CoarseBoVW+BoSW	62.79	67.15	70.83	76.73
	DenseBoVW+BoSW	**86.35**	**88.03**	**88.67**	**90.30**
UIUC	CoarseBoVW	45.47	48.70	51.67	53.21
	DenseBoVW	71.52	73.68	75.12	77.62
	BoSW	46.61	54.78	60.22	65.32
	CoarseBoVW+BoSW	48.38	55.15	61.53	66.58
	DenseBoVW+BoSW	**75.61**	**79.92**	**83.28**	**87.52**
LULC	CoarseBoVW	68.04	71.33	75.04	75.38
	DenseBoVW	77.85	79.85	81.12	81.71
	BoSW	58.81	66.52	72.19	74.85
	CoarseBoVW+BoSW	67.91	72.33	77.67	78.42
	DenseBoVW+BoSW	**81.57**	**83.57**	**85.61**	**88.33**

Note: Bold values indicate the best results.

Table 3 Running times (s) for baseline BoVW and BoSW on MSRC-9.

	CoarseBoVW	DenseBoVW	BoSW
Feature extraction	0.21	0.93	0.82
Clustering	68.23	209.67	173.47

2. The proposed BoSW has potential to attain further improvement with increasing base codebook size. Employing only discriminative structure features outperforms appearance-based baselines, in particular when the base codebook size is large.
3. Albeit handicapped, the coarseBoVW's performance increases significantly when it is combined with the BoSW.
4. The proposed denseBoVW+BoSW is superior to all other baselines in all base codebook sizes. This demonstrates the effectiveness of our method. Variants using the structure features consistently outperform the baseline BoVW models independent of the sampling mode.

Lastly, we analyze the computational load of the proposed method implemented in MATLAB® R2014a running on an Intel i5-2400 processor at 3.1 GHz with 4 GB RAM. The computational times for feature extraction and assignment (per image) and clustering (into 200 visual words) on the MSRC-9 data set are shown in Table 3.

3.2 *Comparisons with the State of the Art*

We compared against 15 recent state-of-the-art image classification methods. Table 4 presents the image classification performance. The second column shows the base codebook sizes for the BoVW-based methods.

As is visible, our method consistently generates the most accurate results among the BoVW-based approaches. The proposed denseBoVW+BoSW achieves the best performance in the MSRC-9 and UIUC-Sports, and LULC data sets, and ranks second best in the LabelMe data set.[16]

Several of the methods we evaluated[4,11,17,18,20,22] employ dense models and learn sophisticated higher level representations; thus, their final feature dimensions for image classification in many cases are much greater than their base codebook sizes. Our method using even 200 codebook size outperforms many other BoVW-based approaches. This high classification accuracy underlines the advantage of our joint use of low-level and structure features. As expected, insensitivity to spatial information is a major weakness of other BoVW-based methods. In our experiments, Niu et al.[4] achieves the second best among BoVW-based methods in LabelMe and UIUC datasets as a result of its explicit modelling of the contextual information.

Also notice that our goal is not low-level feature learning. For example, the work in Ref. 21 attains 86.64% on the LULC data set; however, it requires multiple features [SIFT, local binary pattern, and color], while our method uses only SIFT. The work in Ref. 16 is not a BoVW-based method, uses a hybrid generative score-space scheme on top of local features and 40 latent topics, and imposes the variational free energy of a generative model as a primary source for classification. For the same reason, we did not compare against Ref. 23, which learns spatial pooling regions jointly with discriminative part appearances in a unified optimization framework. In Ref. 19, a latent SVM model regularized by group sparsity to learn class-specific part detectors is proposed, which is not based on BoVW either.

There is also very recent work[13,24] on image classification using convolutional neural networks (CNNs), which can be considered data-driven feature learning. They are not bag-of-words models. The proposed denseBoVW+BoSW outperforms[13] these, as shown in Table 4. Furthermore, our method has the potential to incorporate features learned by a CNN framework to provide additional structure via the receptive field of the fully connected layer neurons.

Table 4 Classification accuracy of state-of-the-art (ordered according to publication dates).

		LabelMe (%)	MSRC9 (%)	UIUC (%)	LULC (%)
Lazebnik et al.[2]	200	81.20	84.20	72.00	81.30
Li and Fei-Fei[11]	300	86.00	—	73.40	—
Wang et al.[15]	240	81.87	—	65.00	—
Perina et al.[16,a]	—	**92.00**	80.40	—	—
Yang and Newsam[12]	300	—	—	—	81.19
Wu and Rehg[17]	200	86.20	—	78.30	—
Niu et al.[18]	800	80.00	—	68.00	—
Niu et al.[4]	500	87.00	—	78.00	—
Sun and Ponce[19,b]	—	—	—	*86.40*	—
Zheng et al.[20]	240	83.43	—	77.29	—
Shaohua and Aggarwal [6,c]	—	81.20	—	77.60	—
Zhao et al.[21]	300	—	—	—	86.64
Zang et al.[22]	240	80.00	—	71.00	—
Romero et al.[13,d]	—	—	—	—	84.53
Chen and Tian[14]	1000	—	—	—	*87.60*
Ours	200	88.03	**87.41**	79.92	83.57
Ours	1024	*90.30*	**97.78**	**87.52**	**88.33**

Note: Best values are denoted in bold and second-best values are denoted in italics.
[a]Uses a probabilistic index map generative model followed by a free-energy–based optimization.
[b]Employs part-based models.
[c]Imposes local spatial homogeneity of latent topics.
[d]A CNN-based method.

3.3 *Parameters*

For the dense sampling strategy, we extracted SIFT features from 16×16 patches over a grid with spacing of 8 pixels for

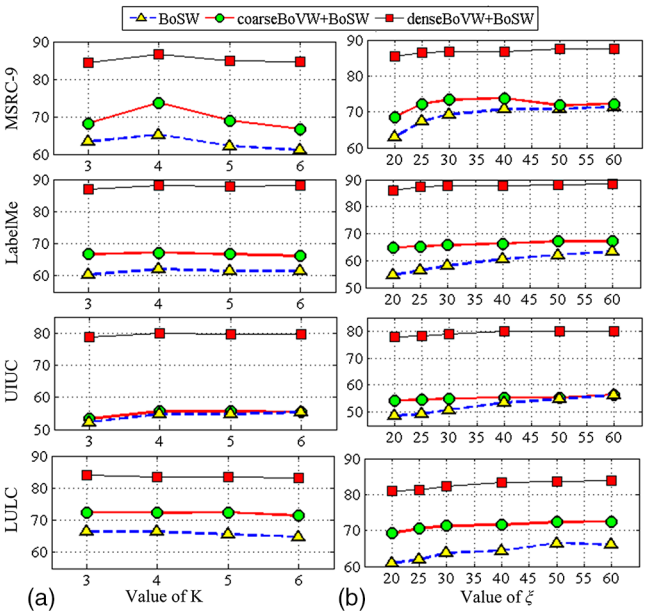

Fig. 3 Classification accuracy for different parameter settings ($d = 200$). (a) The prevalence limit K and (b) the distance cutoff ξ. As shown, the proposed method has relatively low sensitivity to parameters.

all images. For the BoSW, we used $t = 2d$. We chose the distance threshold ξ and the prevalence limit K on the validation data. The results for different parameter settings on four data sets are shown in Fig. 3. The limit K determines how many most frequent words should contribute to the structure features. Its lower values improve the robustness but also cause loss of discriminative potential of the BoSW model. The distance cutoff threshold ξ sets the maximum possible distance between the relevant keypoints, thus determining the size of the local support. As shown, our method is robust to parameterization and demonstrates relatively consistent performance. The performance changes only slightly with respect to K. For the parameter ξ, the performance saturates for values of ξ larger than 50 on all data sets. To achieve a trade-off between discriminative power and robustness, we set $K = 4$ and $\xi = 50$ in all experiments.

4 Conclusion

We present an efficient spatial encoding of visual words that incorporates both local and global structures for image classification using a second layer of spatial encoding with features. We show that by constructing features using appearance- and structure-based bag-of-words models, it is possible to achieve more accurate and robust representations.

Acknowledgments

This work was supported under the Australian Research Council's Discovery Projects funding scheme (Project No. DP150104645) and the National Natural Science Foundation of China (No. 61472161).

References

1. E. Sudderth et al., "Learning hierarchical models of scenes, objects, and parts," in *IEEE Int. Conf. on Computer Vision* (2005).
2. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2006).
3. S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlations," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2006).
4. Z. Niu et al., "Context aware topic model for scene recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2012).
5. A. Bolovinou, I. Pratikakis, and S. Perantonis, "Bag of spatiovisual words for context inference in scene classification," *Pattern Recognit.* **46**(3), 1039–1053 (2013).
6. W. Shaohua and J. K. Aggarwal, "Scene recognition by jointly modeling latent topics," in *IEEE Winter Conf. Applications of Computer Vision* (2014).
7. L. Xie et al., "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.* **23**(5), 1994–2008 (2014).
8. Y. G. Jiang, C. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM Int. Conf. on Image and Video Retrieval* (2007).
9. J. Shotton et al., "Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision* **81**(1), 2–23 (2009).
10. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vision* **42**(3), 145–175 (2001).
11. L. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *IEEE Int. Conf. on Computer Vision* (2007).
12. Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems* (2010).
13. A. Romero, C. Gatta, and C. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.* **99**, 1–14 (2015).
14. S. Chen and Y. L. Tian, "Pyramid of spatial relations for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.* **53**(4), 1947–1957 (2015).
15. C. Wang, D. Blei, and F. F. Li, "Simultaneous image classification and annotation," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2009).
16. A. Perina et al., "A hybrid generative/discriminative classification framework based on free-energy terms," in *IEEE 12th Int. Conf. Computer Vision* (2009).
17. J. Wu and J. Rehg, "Centrist: a visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(8), 1489–1501 (2011).
18. Z. Niu et al., "Spatial-DiscLDA for visual recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2011).
19. J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *IEEE Int. Conf. on Computer Vision* (2013).
20. Y. Zheng, Y. J. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2014).
21. L. J. Zhao, P. Tang, and L. Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(12), 4620–4631 (2014).
22. M. Zang et al., "A novel topic feature for image scene classification," *Neurocomputing* **148**, 467–476 (2015).
23. L. Di et al., "Learning important spatial pooling regions for scene classification," in *IEEE Conf. Computer Vision Pattern Recognition* (2014).
24. M. Koskela and J. Laaksonen, "Convolutional network features for scene recognition," in *ACM Int. Conf. on Multimedia* (2014).

Fatih Porikli is an IEEE fellow and a professor in the Research School of Engineering, Australian National University. He is also managing the Computer Vision Research Group at Data61. He has received his PhD from New York University in 2002. He is the recipient of the R&D 100 Scientist of the Year Award in 2006. He won four best paper awards at premier IEEE conferences and received five professional prizes.

Biographies for the other authors are not available.