AUGMENTING DESCRIPTORS FOR FINE-GRAINED VISUAL CATEGORIZATION USING POLYNOMIAL EMBEDDING

Hideki Nakayama
Grad. School of Information Science and Technology, The University of Tokyo
{*nakayama@ci.i.u-tokyo.ac.jp*}

ABSTRACT

Fine-grained visual categorization (FGVC), which is a relatively new research area, distinguishes conceptually and visually similar categories such as plant and animal species. While FGVC is expected to lead to many task-specific practical applications, it is known as an extremely difficult problem because interclass variations are often quite subtle.

We believe that the key to FGVC is improving local descriptors to enhance discriminative power at the local patch-level. While the pooling strategy of descriptors has been intensively improved for bag-of-visual-words (BoVW) based image representations, the descriptors themselves are often untouched. In this paper, we propose a descriptor augmentation method that utilizes polynomial embedding and supervised dimensionality reduction. Since our method provides moderate-sized compressed descriptors, it can be naturally integrated with off-the-shelf BoVW techniques. In experiments, we show that our method achieves state-of-the-art performance on standard FGVC datasets, Caltech-Birds, and Oxford-Flowers.

Index Terms— Fine-grained Visual Categorization, Local Descriptors, Polynomial Embedding, Bag-of-Visual-Words, Fisher Vector

1. INTRODUCTION

Recently, generic image classification techniques have been making steady progress. Among them, fine-grained visual categorization (FGVC) [1] is now thought to be a promising new framework. The goal of FGVC is to categorize conceptually (and thus visually) similar classes such as plant and animal species [2, 3, 4, 5]. Such a technique would give rise to many practical applications such as bird-watching assistance and online plant identification [6]. However, it is regarded to be extremely difficult because of its high intra-class and low inter-class variations [2].

To distinguish very similar categories, we need to extract highly informative visual features. We believe that the key to achieving this is to enhance the discriminative power of local descriptors. Hitherto, while the pooling strategy of descriptors has been well studied [7, 8, 9] for bag-of-visual-words (BoVW) [10] based methods, descriptors themselves are often untouched, using standard, expert-provided ones (such as the SIFT [11]) as "given". Although some techniques to augment descriptors have been proposed for image matching and registration [12, 13, 14], this point has been largely overlooked in the context of image categorization problems.

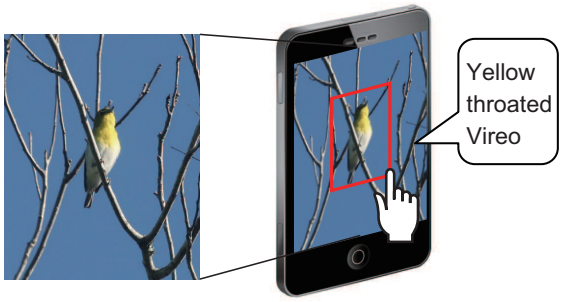

Fig. 1. An illustration of a case of using FGVC techniques (Bird-watching). Users would indicate target objects using tablet interfaces.

In this paper, we show that we can efficiently improve the discriminative performance of arbitrary local descriptors for BoVW-based systems with a simple supervised dimensionality reduction method. Using polynomials of a descriptor and its neighbors, we can efficiently exploit local spatial co-occurrence patterns. Our method is motivated by the recent success of descriptor learning in image matching, wherein descriptors are trained in a supervised learning framework. Although our approach is based on a strong assumption that every patch in an image is somewhat related to its category, this is reasonable for FGVC problems considering its applications. Unlike in a generic image categorization setup, human users do not know what the image content is (e.g., species of birds), and actively want to know it. In such a case, it is reasonable to expect users to provide a bounding box or a mask (Fig. 1).

With this in mind, we perform extensive experiments using the challenging FGVC datasets. Our descriptor augmentation method can dramatically improve the classification performance of BoVW-based image representations. Moreover, when used with the state-of-the-art Fisher vector coding [8], it outperforms the current best performing methods.

2. RELATED WORK

Our method for augmenting descriptors essentially consists of two ideas: exploiting local spatial information, and supervised discriminative dimensionality reduction. Here, we summarize related work respectively.

2.1. Local Spatial Information

Local spatial information, i.e., the local arrangement of neighboring descriptors, has been shown to have rich discriminative power. The standard approach for exploiting such information is descriptor coupling [15, 16]. These methods pair descriptors in the visual word space, where each descriptor is assigned one visual word as in the usual BoVW approach, and the pairwise histogram is used as an image signature. However, the number of histogram bins can be quite large when a large number of visual words are used. On the other hand, Morioka *et al.* proposed a pairing method in the descriptor space by concatenating spatially neighboring descriptors into one long descriptor [17, 18]. However, since the dimension of concatenated descriptors becomes large as the number of included neighbors increases, the computational complexity for coding image signatures also becomes high.

Harada *et al.* included local spatial information by simply using the correlation of elements of neighboring local descriptors [19, 20]. Despite its simplicity, their method showed good performance for image classification problems. However, this method is not designed for BoVW-based feature representations. Motivated by their idea, in this study, we exploit polynomials of neighboring descriptors to encode local spatial information in BoVW-based approaches.

2.2. Discriminative Dimensionality Reduction

Reducing the dimensionality of descriptors has been an attractive topic within the image matching and retrieval community. Recent studies have aimed at learning compact binary descriptors that are computationally quite efficient in terms of both calculation and storage use [12, 21, 22, 23]. Although binarization (hashing) is beyond the scope of this paper, we also apply dimensionality reduction methods to raw descriptors.

Many methods have been proposed for this task, both in unsupervised and supervised approaches. We focus on supervised methods, which we anticipated to be the key to improving system performance. The method proposed by Brown *et al.* [14] appears to be the closest to ours, wherein linear discriminant analysis is used for compressing local descriptors. The objective of their work was descriptor-level image matching. Therefore, each descriptor in the training dataset is assigned labels (true match or not). LDAHash [13] and [12] extends this idea to learn compact binary codes. In this paper, we follow the same strategy in the context of generic image categorization. We apply a simple linear dimensionality reduction method to descriptors using image-level labels.

3. OUR APPROACH

3.1. Augmented Descriptor

We densely extract local features $\boldsymbol{v} \in R^d$ from images. Each patch at position (x, y) is described by $\boldsymbol{v}_{(x,y)}$. We augment this by explicitly including the polynomials [1] of its elements. Let $\boldsymbol{p}^c_{(x,y)}$ denote the augmented descriptor, where c is the number of neighbors considered. When no neighbor is considered,

$$\boldsymbol{p}^0_{(x,y)} = \begin{pmatrix} \boldsymbol{v}_{(x,y)} \\ upperVec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x,y)}\right) \end{pmatrix}, \quad (1)$$

where $upperVec()$ is the flattened vector of the components in the upper triangular part of a symmetric matrix.

Moreover, we can efficiently exploit local spatial information by taking the polynomials between neighboring descriptors. When considering two neighbors as shown in Fig. 2 (a),

$$\boldsymbol{p}^2_{(x,y)} = \begin{pmatrix} \boldsymbol{v}_{(x,y)} \\ upperVec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x,y)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x-\delta,y)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x+\delta,y)}\right) \end{pmatrix}, \quad (2)$$

where $Vec()$ is the flattened vector of the components of a matrix, and δ is an offset parameter for defining neighbors. Similarly, when considering four-neighbors (Fig. 2 (b)),

$$\boldsymbol{p}^4_{(x,y)} = \begin{pmatrix} \boldsymbol{v}_{(x,y)} \\ upperVec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x,y)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x,y-\delta)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x-\delta,y)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x+\delta,y)}\right) \\ Vec\left(\boldsymbol{v}_{(x,y)}\boldsymbol{v}^T_{(x,y+\delta)}\right) \end{pmatrix}. \quad (3)$$

3.2. Label Vector

Our label vector is quite simple; if the image has the label w_i, the i-th element of the labels feature is one; otherwise, it is zero. Therefore, for categorization problems, the dimension of the label vector is the number of categories. Only one element that corresponds to the image's category is one; all other elements are zero [2].

[1] We use at most the second-order polynomials in this paper considering the computational cost, although our framework supports higher-order ones.

[2] This label vector could be naturally used for multi-label problems, although this is beyond the scope of this paper.

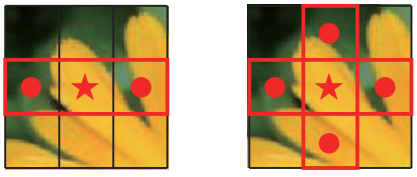

Fig. 2. Models for spatial embedding. Star mark and circles represent the target local descriptor and its neighbors, respectively.

In this paper, we use this image-level label vector for descriptor compression. That is, all $\boldsymbol{p}$ within an image are coupled with the same label vector for supervised dimensionality reduction. Obviously, this is a rather rough approach, since not all local features within an image are actually related to the image-level labels. Nevertheless, we note that this assumption is justified somewhat for FGVC problems, as discussed in the introduction.

3.3. Supervised Dimensionality Reduction

We apply canonical correlation analysis (CCA) [24] to the pairs of the augmented descriptor $\boldsymbol{p}$ and label vector $\boldsymbol{l}$. CCA finds the linear projections $\boldsymbol{s} = A^T\boldsymbol{p}$ and $\boldsymbol{t} = B^T\boldsymbol{l}$ that maximize the correlation between the projected vectors $\boldsymbol{s}$ and $\boldsymbol{t}$. We randomly sample $\{\boldsymbol{p}^{(x,y)}, \boldsymbol{l}^{(x,y)}\}$ pairs from the entire training dataset, and let $C = \begin{pmatrix} C_{pp} & C_{pl} \\ C_{lp} & C_{ll} \end{pmatrix}$ denote their covariance matrices. Namely,

$$C_{pp} = \frac{1}{N}\sum(\boldsymbol{p}-\bar{\boldsymbol{p}})(\boldsymbol{p}-\bar{\boldsymbol{p}})^T, \quad (4)$$

$$C_{ll} = \frac{1}{N}\sum(\boldsymbol{l}-\bar{\boldsymbol{l}})(\boldsymbol{l}-\bar{\boldsymbol{l}})^T, \quad (5)$$

$$C_{pl} = \frac{1}{N}\sum(\boldsymbol{p}-\bar{\boldsymbol{p}})(\boldsymbol{l}-\bar{\boldsymbol{l}})^T, \quad (6)$$

$$C_{lp} = C_{pl}^T, \quad (7)$$

where N is the number of sampled pairs, and $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{l}}$ are their means. The solution of CCA can be obtained by solving the following eigenvalue problem.

$$C_{pl}C_{ll}^{-1}C_{lp}A = C_{pp}A\Lambda^2 \quad (A^TC_{pp}A = I_m), \quad (8)$$

where Λ is the diagonal matrix of the first m canonical correlations, and m is the dimension of the canonical elements. The parameter m corresponds to the dimension of the embedded descriptor, and needs to be tuned manually. One problem is that m can be at most the dimension of the label vector because of the rank problem. If we need more features, we can project $\boldsymbol{p}$ into the orthogonal subspace and iteratively apply CCA to further extract discriminative components.

Using the projections obtained by CCA, we get a compact vector $\boldsymbol{s}$ that embeds a high-dimensional augmented vector, which we call the latent descriptor.

$$\boldsymbol{s} = A^T\boldsymbol{p}. \quad (9)$$

Once the latent descriptor is computed, it can be used in the exact same manner as widely-used raw descriptors such as SIFT.

4. EXPERIMENT

4.1. Image Features and Classifiers

We use several standard local descriptors to test our method, such as SIFT [11], C-SIFT [25], opponent-SIFT [26], and the self-similarity descriptor [27]. The dimension of the self-similarity descriptor is 40 in our experiments (4 radial bins and 10 angle bins). All these local features are extracted in a dense sampling approach. We extract local features from 24x24 patches on regular grids spacing five pixels. These descriptors are compressed into 64 dimensions via PCA, except for the self-similarity descriptor [3]. Finally, we apply our polynomial embedding (PE) method with CCA and obtain 64-dimensional latent descriptor ($m = 64$). We fix the offset parameter $\delta = 20$ for defining neighbors.

Using the latent descriptors, we encode an image-level feature vector by two approaches. The first is the standard BoVW histogram with spatial pyramid matching [28]. The second is the Fisher vector [8], which is a recently proposed powerful representation. We use 64 Gaussians for estimating a Gaussian mixture model and concatenate feature vectors from an entire image and three horizontal regions. For classification, we use LIBSVM [29] and LIBLINEAR [30] packages for the BoVW and Fisher vector, respectively.

4.2. Fine-grained Visual Categorization

4.2.1. Dataset

We experiment with two publicly available datasets: the Oxford-Flower102 dataset [4] and the Caltech-Bird200-2010 dataset [3] (Fig. 4).

The flower dataset consists of 102 flower categories. For each class, 20 images are specified as training samples by the dataset authors. The remaining samples are used for testing. Images are roughly cropped, as shown in Fig. 4. The bird dataset contains 200 bird species. We use 15 samples per class for training and the remaining samples for testing, as specified by the authors. In the same manner as the previous study, we crop images using the provided bounding box and rescale them so that the shorter axis has 150 pixels [31].

[3] We use the raw self-similarity descriptor for polynomial embedding without applying PCA.

Table 1. Comparison of classification rates on the Oxford-Flower102 dataset (%).

Descriptor	SIFT (128dim)		C-SIFT (384dim)		Opp.-SIFT (384dim)		Self Sim. (40dim)	
	BoVW	Fisher	BoVW	Fisher	BoVW	Fisher	BoVW	Fisher
Raw	45.2	-	54.0	-	53.4	-	45.2	60.3
PCA64	45.5	60.8	53.5	76.1	53.4	73.7	-	-
CCA64	45.7	57.4	55.6	74.5	56.0	72.5	-	-
CCA64 (2 neighbors)	43.4	57.7	56.7	74.1	57.7	71.9	47.8	65.8
CCA64 (4 neighbors)	43.4	60.7	57.8	73.3	59.0	73.3	48.1	66.4
PCA64-PE0-CCA64	52.9	64.0	61.8	79.6	62.7	78.0	52.2	62.3
PCA64-PE2-CCA64	54.9	67.5	65.2	80.1	65.2	**80.9**	56.5	69.9
PCA64-PE4-CCA64	**56.8**	**68.9**	**67.6**	**80.5**	**67.9**	80.8	**57.3**	**71.0**

Table 2. Comparison of classification rates on the Caltech-Bird200-2010 dataset (%).

Descriptor	SIFT (128dim)		C-SIFT (384dim)		Opp.-SIFT (384dim)		Self Sim. (40dim)	
	BoVW	Fisher	BoVW	Fisher	BoVW	Fisher	BoVW	Fisher
Raw	7.5	-	10.1	-	10.7	-	8.6	10.6
PCA64	7.7	11.0	10.2	18.3	10.3	19.7	-	-
CCA64	6.7	10.0	10.0	17.9	10.2	17.6	-	-
CCA64 (2 neighbors)	7.4	11.0	9.5	17.9	10.2	17.7	8.0	12.2
CCA64 (4 neighbors)	7.9	12.3	11.8	18.7	11.0	18.2	8.4	**13.2**
PCA64-PE0-CCA64	9.1	12.5	10.6	17.7	11.8	19.7	8.8	11.4
PCA64-PE2-CCA64	9.6	12.9	11.8	20.0	13.6	21.3	**9.3**	12.4
PCA64-PE4-CCA64	**10.2**	**14.3**	**12.0**	**20.8**	**14.7**	**22.9**	**9.3**	11.7

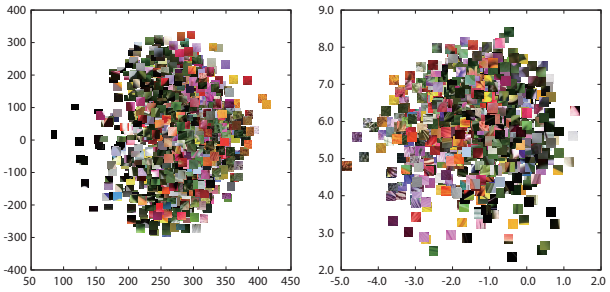

Fig. 3. First two components of compressed SIFT descriptors in the flower dataset. Left: PCA64. Right: PCA64-PE4-CCA64 (ours).

Fig. 4. Images from FGVC benchmark datasets. Top: Oxford-Flower102 [4]. Bottom: Caltech-Bird200-2010 [3].

4.2.2. Experimental Results

For each descriptor, we apply various embedding and dimensionality reduction methods. To illustrate the effectiveness of PE, we also test applying CCA to the raw descriptor. Tables 1 and 2 show the results for the flower dataset and bird dataset, respectively. For example, "PCA64-PE2-CCA64" means three steps: (1) compress the original raw descriptor into $d = 64$ dimensions using PCA, (2) compute the augmented descriptor using PE with two neighbors (Eq. 2), (3) compress the augmented descriptor into a 64-dimensional latent descriptor using CCA. Also, "CCA64 (2 neighbors)" means simply applying CCA to concatenated raw descriptors including two neighbors.

For both BoVW and Fisher vectors, PE substantially improves the performance of the original descriptors. Moreover, the performance is often in the following order: PE4>PE2>PE0. This means that including more neighbors is important for improving the discriminative power. The results also show that just applying CCA to raw descriptors does not improve the performance well, even when neighboring descriptors are concatenated. This fact corresponds to the result in [14]; it indicates the importance of PE for exploiting local spatial information. Figure 3 illustrates the patch distribution of compressed SIFT descriptors in the flower dataset. We see that our latent descriptor separates color better than

Table 3. Classification performance using multiple descriptors (%). Fisher vectors with 64 Gaussians are extracted for each descriptor and integrated at the classifier level.

	Flowers	Birds
4 desc. (PCA64)	81.6	23.9
4 desc. (PCA64-PE4-CCA64)	**87.2**	28.1
8 desc. (PCA64 + PCA64-PE4-CCA64)	85.7	**28.8**
Previous Work	85.6 [32]	28.2 [33]
	80.0 [34]	26.7 [32]
	76.3 [35]	26.4 [36]
	73.3 [37]	22.4 [37]
		19.2 [31]
		19.0 [38]
		18.0 [7]

PCA, although both of them are based on gray-SIFT. Considering that color is strongly related to categories in this dataset, we can expect that our descriptor is more discriminative.

Next, we combine Fisher vectors from each descriptor (PCA64 and PCA64-PE4-CCA64 descriptors of SIFT, C-SIFT, opponent-SIFT, and self-similarity) by a late-fusion approach. We take the average log-likelihood of posterior probability for each classifier weighted by its individual confidence in validation. Table 3 shows the result. Our best method outperforms all previously published results [4].

4.3. Object and Scene Classification

Although the primary target of our method is FGVC, we try other problems for further considerations. Here, we perform object and scene categorization problems using Caltech-101 dataset [39] and MIT indoor scene dataset [40].

The Caltech-101 dataset contains 101 objects and a background class. Each class has between 31 to 800 images. In all, we perform 102 classes classification task. We use 30 samples for training and 50 for testing per class. The MIT indoor scene dataset consists of 67 indoor scene categories. For this dataset, training and testing samples are specified by the authors. We use 80 training samples and 20 testing samples per class.

Table 4 shows the result. Not surprisingly, the PE method seems less powerful than in FGVC problems, because strong supervision assumption does not hold in these datasets. However, we note that combining PCA descriptors and our descriptors in late fusion still improves the performance. This fact suggests that our method can point out additional discriminative information in addition to standard methods.

[4]For the bird dataset, [32] uses the bounding box only for training images, therefore the result is not directly comparable to ours. Nevertheless, we note that we outperform them on the flower dataset, wherein both our method and theirs use raw images for training and testing.

5. CONCLUSIONS

In this paper, we presented a simple but powerful method for augmenting arbitrary local descriptors in the context of fine-grained visual categorization. We found that polynomials of descriptors can efficiently capture local spatial information, thus leading to high performance. Although the dimensionality of polynomials can be quite large, we can easily obtain a small-dimensional latent vector by simply using image-level labels for supervised dimensionality reduction. Our latent descriptors can be used naturally with off-the-shelf BoVW techniques. Using our method with the sophisticated Fisher representation, we outperformed state-of-the-art methods on the standard FGVC datasets.

6. REFERENCES

[1] I. Biederman, S. Subramaniam, and M. Bar, "Subordinate-level object classification reexamined," *Psychological Research*, vol. 62, pp. 131–153, 1999.

[2] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?," in *Proc. ECCV*, 2010, pp. 71–84.

[3] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[4] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conference on Computer Vision, Graphics & Image Processing*, 2008, pp. 722–729.

[5] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization: Stanford Dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011.

[6] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. Lopez, and V. B. Soares, "Leafsnap : A computer vision system for automatic plant species identification," in *Proc. ECCV*, 2012.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR*, 2010, pp. 3360–3367.

[8] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. ECCV*, 2010.

[9] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE CVPR*, 2010, pp. 3304–3311.

[10] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE ICCV*, 1999, pp. 1150–1157.

[12] T. Trzcinski and V. Lepetit, "Efficient discriminative projections for compact binary descriptors," in *Proc. ECCV*, 2012.

Table 4. Classification rate on Caltech-101 and MIT-Indoor datasets (%). Fisher vectors with 64 Gaussians are extracted for each descriptor. Late-fusion is used to combine PCA64 and PCA64-PE4-CCA64 features.

	Caltech-101			MIT-Indoor		
	SIFT	C-SIFT	Opp.-SIFT	SIFT	C-SIFT	Opp.-SIFT
PCA64	65.2	57.9	62.5	51.3	49.6	52.6
PCA64-PE4-CCA64	68.1	60.1	65.8	51.5	50.6	54.8
PCA64 + PCA64-PE4-CCA64	**69.7**	**62.3**	**68.2**	**54.7**	**53.6**	**58.0**

[13] C. Strecha, A. Bronstein, M. Bronstein, and P. Fua, "LDA-Hash: Improved matching with smaller descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2011.

[14] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors.," *IEEE Transactions on Pattern Analysis and Machine Intelligencettern*, vol. 33, no. 1, pp. 43–57, 2011.

[15] X. Lan, C. L. Zitnick, and R. Szeliski, "Local bi-gram model for object recognition," Tech. Rep. MSR-TR-2007-54, Microsoft Research, 2007.

[16] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. IEEE CVPR*, 2008.

[17] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *Proc. ECCV*, 2010, pp. 1–14.

[18] N. Morioka and S. Satoh, "Learning directional local pairwise bases with sparse coding," in *Proc. BMVC*. 2010, pp. 32.1–32.11, British Machine Vision Association.

[19] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Improving local descriptors by embedding global and local spatial information," in *Proc. ECCV*, 2010.

[20] T. Harada and Y. Kuniyoshi, "Graphical Gaussian vector for image categorization," in *Proc. NIPS*, 2012.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE ICCV*, 2011, pp. 2564–2571.

[22] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE ICCV*, 2011, pp. 2548–2555.

[23] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.

[24] H. Hotelling, "Relations between two sets of variants," *Biometrika*, vol. 28, pp. 321–377, 1936.

[25] G. J. Burghouts and J.-M. Geusebroek, "Performance evaluation of local colour invariants," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 48–62, 2009.

[26] K. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–96, 2010.

[27] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE CVPR*, 2007.

[28] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE CVPR*, 2006, vol. 2, pp. 2169–2178.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[30] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[31] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Proc. IEEE CVPR*, 2011.

[32] Y. Chai, E. Rahtu, and V. Lempitsky, "TriCoS: A tri-level class-discriminative co-segmentation method for image classification," in *Proc. ECCV*, 2012, pp. 794–807.

[33] S Yang, L. Bo, J. Wang, and L. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. NIPS*, 2012.

[34] V. Lempitsky and A. Zisserman, "BiCoS: A bi-level co-segmentation method for image classification," in *Proc. IEEE ICCV*, 2011, pp. 2579–2586.

[35] M.-E. Nilsback, *An automatic visual flora: segmentation and classification of flower images*, Ph.D. thesis, University of Oxford, 2009.

[36] L. Bo and D. Fox, "Kernel descriptors for visual recognition," in *Proc. NIPS*, 2010.

[37] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Vanrell, "Portmanteau vocabularies for multi-cue image representation," in *Proc. NIPS*, 2011.

[38] S. Branson, C. Wah, and F. Schroff, "Visual recognition with humans in the loop," in *Proc. ECCV*, 2010.

[39] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Journal of Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[40] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE CVPR*, 2009.