2010 International Conference on Pattern Recognition

Bag of Hierarchical Co-occurrence Features for Image Classification

Takumi Kobayashi
Information Technology Research Institute
AIST
1-1-1 Umezono, Tsukuba, Japan
Email: takumi.kobayashi@aist.go.jp

Nobuyuki Otsu
Fellow
AIST
1-1-1 Umezono, Tsukuba, Japan
Email: otsu.n@aist.go.jp

***Abstract*—We propose a bag-of-hierarchical-co-occurrence-features method incorporating hierarchical structures for image classification. Local co-occurrences of visual words effectively characterize the spatial alignment of objects' components. The visual words are hierarchically constructed in the feature space, which helps us to extract higher-level words and to avoid quantization error in assigning the words to descriptors. For extracting descriptors, we employ two types of features hierarchically: narrow (local) descriptors, like SIFT [1], and broad descriptors based on co-occurrence features. The proposed method thus captures the co-occurrences of both small and large components. We conduct an experiment on image classification by applying the method to the Caltech 101 dataset and show the favorable performance of the proposed method.**

***Keywords*-image classification, bag-of-features, cooccurrence, hierarchical visual words**

I. Introduction

Image classification has become an important and attractive research area since the numbers of photo images stored in PCs and across the Internet have been significantly increasing. Over the last decade, the bag-of-features approach [2], [3], which is derived from text mining methods, has shown impressive levels of performance on image classification tasks.

In the standard bag-of-features method, an image is represented by ensembles (*bag*) of local feature vectors (*descriptors*) which are quantized into clusters (*visual words*), and then a histogram of visual words is constructed by counting the occurrences of the words in the image. The bag-of-features method possesses shift invariance since it does not take into account the spatial locations from which the descriptors are extracted. Recently, spatial information has been explicitly incorporated into this framework, for example, by utilizing spatially partitioned images via kernel methods (spatial pyramid match kernel) [3], in order to improve performance, especially for the Caltech 101 dataset [4] in which target objects are roughly aligned in terms of their positions. Most of those methods, however, lead to a loss of the shift invariance property which is desirable for general object recognition.

With respect to spatial characteristics, objects impose constraints on the (relative) spatial positions of their components, e.g., human body parts. In addition, those components should not be regarded equally, as they form hierarchical structures; e.g., a hand belongs to an arm. These relations provide important cues for recognition, and we focus on these characteristics.

In this paper, we propose a method to extract local *co-occurrences* of visual words, utilizing hierarchical structures in images. While global co-occurrences characterize situations, such as context, in the images, local co-occurrences adequately extract relative spatial alignment of components, which enables shift invariant recognition in the proposed method. In addition, we exploit hierarchical structures for both visual words and descriptors. By utilizing the hierarchical structures of the visual words, we can reduce quantization errors in assigning words to descriptors; also we can extract higher-level words, like concepts. Two types of descriptors are constructed in a hierarchical manner: narrow (local) descriptors like SIFT [1] and broad descriptors based on the co-occurrence features of the narrow descriptors. These descriptors are useful for extracting co-occurrences of various kinds of components which form hierarchical structures in the target objects.

II. Bag of co-occurrence features

The proposed method is based on the co-occurrence histogram of visual words, not simple occurrences used in the standard bag-of-features method. By restricting co-occurrences within local neighborhoods, the resultant histogram features are shift invariant with respect to object positions. In the proposed method, the co-occurrences of visual words are effectively calculated by using local auto-correlation functions as in GLAC [5] which extracts image features based on gradient co-occurrences.

In this paper, we assume that descriptors are extracted on (10 pixel spaced) grids in the image, since such dense descriptors result in better performance for the bag-of-features framework [6]. The proposed method also allows sparse interest points, such as by DoG [1] and Harris-Laplace [7], by slightly modifying the displacement vector $\boldsymbol{a}$ defined in the following. We assign to each descriptor not a symbolic label but a word vector $\boldsymbol{w}$ representing a visual word. Suppose we have N visual words (clusters);

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.945

IEEE computer society

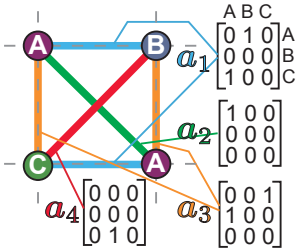

Figure 1. Co-occurrence features of visual words. 'A','B' and 'C' indicate visual words. The co-occurrence histogram is represented by the matrices.

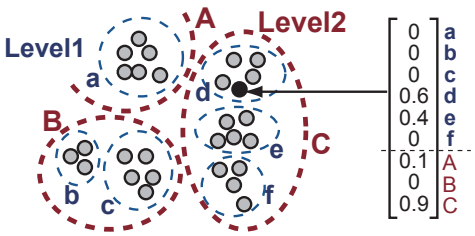

Figure 2. Hierarchical structures in the feature space. Dots represent descriptors extracted from the images.

the word vector then has dimension N ($\boldsymbol{w} \in \boldsymbol{R}^N$), such that there are only a few non-zero elements associated with the visual words that the descriptor belongs to. In this paper, since we employ soft assignment [8] and hierarchical words as described in the next section, the word vector has a few (more than one) non-zero elements.

The first order auto-correlation function for the visual words is then defined as follows:

$$\boldsymbol{H}(\boldsymbol{a}) = \sum_{\boldsymbol{r} \in D} \boldsymbol{w}(\boldsymbol{r}) \otimes \boldsymbol{w}(\boldsymbol{r} + \boldsymbol{a}), \tag{1}$$

where $\otimes$ denotes the outer-product of vectors, $\boldsymbol{r}$ is a position vector $\boldsymbol{r} = (x, y)$ in the whole image grid $D = X \times Y$, and $\boldsymbol{a}$ is a displacement vector, $\boldsymbol{a} \in \{(\Delta r, 0), (\Delta r, \Delta r), (0, \Delta r), (-\Delta r, \Delta r)\}$ in which shift-equivalent patterns are excluded. The parameter Δr indicates the interval for local co-occurrences. Note that $\boldsymbol{r}, \boldsymbol{a}$ and Δr are defined in the 2D grid coordinates, not pixel coordinates. The co-occurrence histogram $\boldsymbol{H}(\boldsymbol{a})$ counts all visual word pairs that co-occur along the displacement vector $\boldsymbol{a}$, as shown in Fig. 1, and it is actually unfolded to a vector $\boldsymbol{h}(\boldsymbol{a})$. The image feature $\boldsymbol{h}$ is finally constructed by concatenating the co-occurrence histogram vectors $\boldsymbol{h}(\boldsymbol{a})$ over all displacements. This is an extended formulation of the standard bag-of-features method which can be represented by the zeroth order auto-correlation function:

$$\boldsymbol{H}^{(0)} = \sum_{\boldsymbol{r} \in D} \boldsymbol{w}(\boldsymbol{r}). \tag{2}$$

These zeroth order features are not employed in this study.

The method in [9] also utilizes similar co-occurrences of visual words. However, displacement vectors are not introduced and soft representations for words are not dealt with. In this paper, we define the local auto-correlation functions in Eq.(1) and provide a more general formulation. It should be noted that the computational cost of Eq.(1) is low since the word vector $\boldsymbol{w}$ is sparse, and it is of the same order $O(n)$ as that of the bag-of-features, where n is the number of pixels.

III. Hierarchical Structures

We employ hierarchical representations in the bag-of-co-occurrence-features method described in the previous section. We focus on hierarchical structures for both visual words and descriptors.

A. Hierarchical visual words

In general, visual words are constructed by clustering descriptor vectors in the feature space. This procedure corresponds to quantization of the feature space and a quantization error is inevitable, depending on the assumed number of visual words (clusters). Therefore, we exploit the hierarchical structure of the feature space as in [10] (Fig. 2). For extracting hierarchical structures, we repeatedly apply k-means clustering, decreasing the number k and obtaining the clusters (visual words) at different levels. A vocabulary tree [10] could also be applied, but this causes quantization errors by definitely assigning words to the descriptor at each level in a sequential manner. At each level, visual words are assigned to the descriptor by soft assignment and then the word vectors at all levels are concatenated into the whole word vector $\boldsymbol{w}$, the dimensionality of which is $N = \sum_i N_i$ (N_i is the number of words at the i-th level), as shown in Fig. 2. Since the concatenated word vector $\boldsymbol{w}$ contains visual words at every level, the co-occurrences between visual words at different levels can be naturally extracted by the auto-correlations in Eq.(1). This hierarchical representation with soft assignment possibly reduces quantization error and robustly extracts higher-level words like concepts.

Soft assignment is achieved by determining the weights for the m nearest visual words as follows:

$$\omega_i = \frac{d_1/d_i}{\sum_{j=1}^{m} d_1/d_j} = \frac{\prod_{l \neq i} d_l}{\sum_{j=1}^{m} \prod_{l \neq j} d_l}, \tag{3}$$

where d_i is the distance to the i-th nearest cluster center (word). This is derived from the distance ratio on the basis of the nearest ($i = 1$) distance.

B. Hierarchical descriptors

Objects have different localities for their components. For example, the human body has hierarchical structures of fingers, hands, and arms, which are ordered according to their locality. We describe such components by different types of features considering their locality. For narrow (local) descriptors, we employ features such as SIFT [1] and GLAC [5], and for broad descriptors, we directly apply

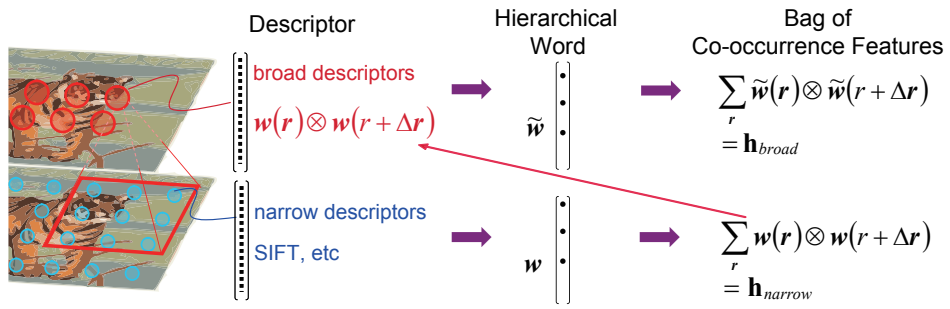

Figure 3. Hierarchical structure using narrow and broad descriptors.

co-occurrence features (Sec.II) using narrow descriptors, as follows (Fig. 3).

The proposed co-occurrence features can be applied to broad descriptors by using visual words of the narrow descriptors and by replacing the whole image region D with a local region D_s (7×7 grid points) in Eq.(1). The feature is based on combinations (co-occurrences) of the narrow descriptors, and effectively characterizes bigger components, such as an arm composed of an elbow and a hand in the human body.

For these two types of descriptor, hierarchical visual words are constructed (Sec.III-A) and then the co-occurrence features for the visual words are extracted (Sec.II), respectively, as shown in Fig. 3. The extracted features are finally concatenated to form the complete image feature vector with a weighting parameter α ($0 \leq \alpha \leq 1$):

$$\boldsymbol{h} = \left[\sqrt{\alpha}\,\boldsymbol{h}_{narrow}^T, \sqrt{1-\alpha}\,\boldsymbol{h}_{broad}^T\right]^T, \quad (4)$$

where $\boldsymbol{h}_{narrow}, \boldsymbol{h}_{broad}$ are the co-occurrence feature vectors based on the narrow and broad descriptors, respectively. In this study, we assume the feature vectors are normalized ($||\boldsymbol{h}_{narrow}|| = ||\boldsymbol{h}_{broad}|| = 1$); the above weighting is to keep constant the norm of the total feature vector ($||\boldsymbol{h}|| = 1$).

IV. Experiment

We applied the proposed method to image classification using the Caltech 101 dataset [4]. The dataset contains images in 101 categories with large intra-class variability, as shown in Fig. 4.

A. Experimental setting

For narrow (local) descriptors, we employed SIFT [1] and GLAC [5] features, and their performances are compared. We used three hierarchical levels for visual words: 250 words for level 1, 50 words for level 2, and 10 words for level 3. The number of words for soft assignment was 3 ($m = 3$ in Eq.(3)). The spatial interval Δr was set to 3 grid points for the co-occurrence features of both narrow and broad descriptors. The weight α in Eq.(4) for those two types of co-occurrence features was determined based on

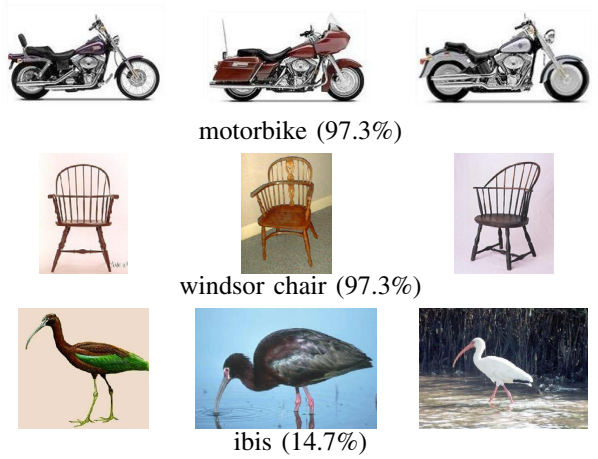

Figure 4. Example images in Caltech101. The classification rates of the proposed method are also shown in parentheses.

cross validations. For classification, we applied multi-class linear SVM [11].

We followed the standard evaluation protocol: The dataset was split randomly into 15 training images per category and 15 images for testing. We calculated the classification rate for each category and then averaged them across all 101 categories. The trial was repeated five times and the average performance is reported.

B. Experimental result

Firstly, we show the effectiveness of the proposed method by varying the settings in the method. The baseline results for the standard bag-of-features method are 35.8% for SIFT descriptors and 41.8% for GLAC descriptors (Fig. 5(a)). In this case, the number of visual words was 250, which corresponds to level 1 in hierarchical words. The method of bag-of-co-occurrence-features (Sec.II) with 250 visual words only at level 1 produces 49.17% for SIFT and 52.2% for GLAC (Fig. 5(b)). By employing local co-occurrences of visual words, performance was improved by more than 10% for both types of descriptor. This result shows that local co-occurrences of visual words are effective for classification by capturing spatial alignment of the words. Then, we

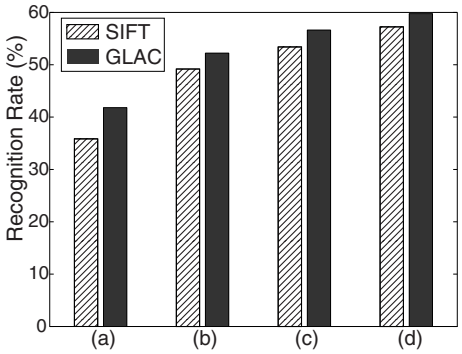

Figure 5. Performance results. Details are in the text.

Table I
PERFORMANCE COMPARISON.

Method	[14]	[3]	[13]	[15]	[16]	Ours
Acc.(%)	49.5	56.4	59.1	52.0	51	59.8

additionally applied hierarchical visual words (Sec.III-A). Performance was further improved by about 4%: 53.4% for SIFT and 56.6% for GLAC (Fig. 5(c)). Finally, the proposed method using a bag-of-co-occurrence-features with hierarchical words and hierarchical descriptors was applied. The results were *57.2%* for SIFT and *59.8%* for GLAC (Fig. 5(d)). The proposed method significantly improved performance for both types of descriptor by about 20% compared to the baseline results.

Next, we compare the result of the proposed method using GLAC descriptors to the other methods, as shown in Table I. For fair comparison, we applied only methods based on a single type of feature, not using multiple kernel learning [12]. State-of-the-art results were obtained by using exhaustive classifier [13] and spatial pyramid match kernel [3] which spatially partitions images and thus is position-specific. The proposed method, which is shift invariant and uses simple linear classification with a low computational cost, produces better results compared to the previous methods.

V. CONCLUSION

We have proposed a bag-of-hierarchical-co-occurrence-features method incorporating hierarchical structures for image classification. The spatial alignments of objects' components are effectively characterized by the local auto-correlation function of visual words for local co-occurrences, and the method is shift invariant. Both narrow and broad descriptors are extracted and the visual words are hierarchically assigned to the descriptors to capture various levels of characteristics. In the experiment for image classification using the Caltech 101 dataset, the proposed method exhibited favorable performance compared to state-of-the-art methods.

REFERENCES

[1] D. Lowe, "Distinctive image features from scale invariant features," *International Journal of Compuater Vision*, vol. 60, pp. 91–110, 2004.

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

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[5] T. Kobayashi and N. Otsu, "Image feature extraction using gradient local auto-correlations," in *ECCV*, 2008.

[6] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *ICCV*, 2007.

[7] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Compuater Vision*, vol. 65, pp. 43–72, 2005.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[9] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *ICCV*, 2007.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006, pp. 2161–2168.

[11] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[12] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV*, 2007.

[13] H. Zhang, A. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *CVPR*, 2006.

[14] K. Grauman and T. Darrell, "Discriminative classification with sets of image features," in *ICCV*, 2005.

[15] R. Zhang, C. Wang, and B. Xiao, "A strategy of classificaion via sparse dictionary learned by non-negative k-svd," in *ICCV*, 2009.

[16] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," in *CVPR*, 2006.