LETTER

Extraction of Combined Features from Global/Local Statistics of Visual Words Using Relevant Operations

Tetsu MATSUKAWA[†a)], *Student Member and* Takio KURITA[††], *Member*

SUMMARY This paper presents a combined feature extraction method to improve the performance of bag-of-features image classification. We apply 10 relevant operations to global/local statistics of visual words. Because the pairwise combination of visual words is large, we apply feature selection methods including fisher discriminant criterion and L1-SVM. The effectiveness of the proposed method is confirmed through the experiment.

key words: *combined feature, bag-of-features, feature selection, image classification*

1. Introduction

Generic object recognition technologies have many possible applications such as automatic image search. However, generic object recognition involves some very difficult problems because one has to deal with inherent object/scene variations as well as difficulties in viewpoint, lighting, and occlusion. Thus, although many methods of generic object recognition have been developed so far, the classification performance of these conventional methods is still insufficient, and a method that can achieve high classification accuracy is strongly desired.

The bag-of-features approach is the most popular approach for generic object recognition [1] because of its simplicity and effectiveness. This approach is originally inspired from the text recognition method called "bag-of-words," and this method treats an image as an orderless collection of quantized appearance descriptors extracted from local patches. The main steps of the bag-of-features are (1) detection and description of image patches, (2) assigning patch descriptors to a set of predetermined codebooks with a vector quantization algorithm, (3) constructing a bag of features, which counts the number of patches assigned to each codebook, and (4) applying a classifier by treating the bag of features as the features vector and thus determining the category which an image can be assigned.

It is known that the bag-of-features method is robust with regard to background clutter, pose changes, and intraclass variations and offers good classification accuracy.

In this paper, we propose a combined feature extraction method to improve the performance of the bag-of-features image classification. The proposed method includes 10 operations of pairwise histogram components and feature selection methods. The effectiveness of the proposed feature extraction for bag-of-features is confirmed through experiments using the popular Scene-15 dataset [2].

2. Related Work

We review here only closely related work on our proposed feature extraction. Nakayama et al. proposed the generalized local correlations (GLC) method [3] for scene classification. GLC method use the correlations of the histogram components in local features. However GLC doesn't use the autocorrelations of the visual words, which usually produce high classification accuracy. Cao et al. proposed the second-order HOG features [4] for pedestrian recognition. Second-order HOG feature can extract co-occurrence of local statistics of edge direction among cell, and this feature significantly outperformed HOG features performance. In text classification method, the combination feature are often used to achieve high accuracy [5]. As mentioned above, the effectiveness of such feature combinations is promising. However, there are no reports to introduce such combination feature extraction to bag-of-features. It should be also mentioned that few operators were considered in these literatures; these are product, min, and harmonic mean. Thus, the contributions of this paper is two-fold; 1): we apply the combined feature extraction to bag-of-features image classification and confirmed its effectiveness. 2): we apply 10 operators, that includes new operators that are not used in previous combined feature extraction methods.

3. Spatial Pyramid Matching (Local Statistics)

To alleviate the loss of spatial layout information in bag-of-features image representation, one of the most successful approaches so far is the spatial pyramid matching (SPM) technique proposed by Lazebnik et al. [2]. SPM divides an image into subregions and integrating corresponding results in these regions. Since SPM usually improves image classification accuracy, this method is used in many recent articles [7], [8]. The methods that use overlap grid [7], vertical and horizontal grid [6] are also proposed. This paper use original spatial pyramid layout used in [2]. As shown in Fig. 2, the level 2 split in a spatial pyramid divides the image into $2^2 \times 2^2 = 16$ blocks. Similarly, level 1 and 0 have 4, 1 blocks respectively. Then the histograms in all

Manuscript received April 1, 2010.
Manuscript revised June 22, 2010.

[†]The author is with the School of Systems and Information Engineering, University of Tsukuba, Tsukuba-shi, 305–8577 Japan.
[††]The author is with the Department of Information Engineering, Hiroshima University, Higashihiroshima-shi, 739–8527 Japan.
a) E-mail: t.matsukawa@aist.go.jp

DOI: 10.1587/transinf.E93.D.2870

Table 1 Relevant operations.

Operations	Definition	Type	Explanation
Sum	h$_i$ + h$_j$	OR	marge two components
Sub	h$_i$ - h$_j$	DIFF	degree of big/small
Div	$\frac{h_i}{h_j}$	DIFF	rate of big/small
Prod	h$_i$h$_j$	AND	degree of co-occurrence
Sum(binary)	binary(h$_i$ + h$_j$)	OR	appear/not appear when marge two components
Sub(binary)	binary(h$_i$ - h$_j$)	DIFF	big/small
Prod(binary)	binary(h$_i$ h$_j$)	AND	co-occur/not co-occur
Max	max(h$_i$, h$_j$)	OR	highest component
Min	min(h$_i$, h$_j$)	AND	degree of co-occurrence
Harmonic mean	$\frac{2h_ih_j}{h_i+h_j}$	AND	degree of co-occurrence

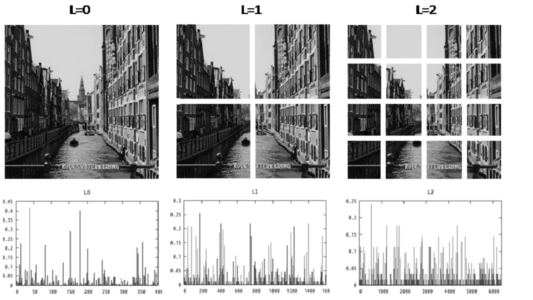

Fig. 1 Spatial pyramid bag of features [2].

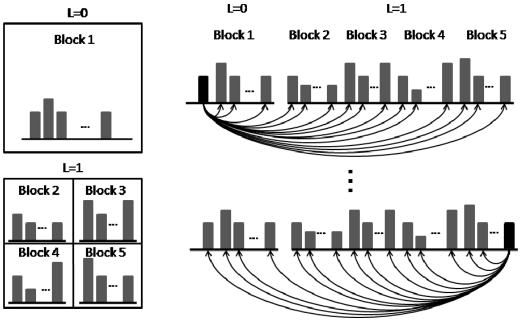

Fig. 2 Pairwise relationship among the histogram components [4]. The histogram components of upto level 1 spatial pyramid are shown.

blocks are concatenated. For example, a level 2 pyramid have 16 + 4 + 1 = 21 blocks and a level 1 pyramid have 4 + 1 = 5 blocks respectively (Fig. 1).

4. Combined Feature Extraction

Let k be the number of visual words and $\mathbf{H} = (h_1, \ldots, h_K)$ be the concatenated histogram of spatial pyramid, where K $\in \{k, 5k, 21k\}$ is the dimension of spatial pyramid bag-of-features for each spatial pyramid level 0, 1 and 2. This paper presents 10 operations for the combinations of the histogram components (h_i, h_j) of **H** (Table 1). The details of these operators are described as follows;

Summation: Summation of two variable indicates weak co-occurence relationship because the summation value is not affected largely if the one value is very low. This operation is also recognized as merging two variances.

Subtraction: Subtraction means the difference of the frequencies of two visual words.

Division: Division represents the rate of the frequencies of two visual words. Because the value $\frac{h_i}{h_j}$ becomes very high if h_j is closely to zero, we set the maximum value of the division operator to 100.

Product: Product of two variable is often used in combination of the features. This operator express strong co-occurrence of two variable.

Summation(binary): In Table 1, binary(*) returns 1 if the value * is higher than 0 and returns 0 in other case. The intension of binarized summation operators is only consider whether the visual words appears or not.

Subtraction(binary): Binarized subtraction operators means the bigger/smaller relationship of frequencies of two visual words.

Product(binary): Binarized product becomes 1 only when the frequency of visual words are bigger than 1 in two histogram bin. So, this operator represents AND relationship.

Max: Different from binarized summation, max represents OR relationship of two histogram components in continuous value.

Min: Min operator also represents strong co-occurrence of two variants. Different from binarized product, the value of min is continuous.

Harmonic mean: We used harmonic means because the effectiveness of this operator is confirmed in [4].

There are K^2 combination for subtraction, division and binarized subtraction operators and $_KC_2$ combination for summation, product, binarized summation, binarized product, max, min and harmonic means operators. This combination is very large. We select M combined features by feature selection method described in following section, and concatenate combined features to spatial pyramid bag-of-features **H**. Then the K+M dimension features is used for classification.

5. Feature Selection

Let x_j, $j = 1, 2, \ldots, P$ be a possible combined feature, where $P \in \{K^2,\ {}_KC_2\}$. Because the combination of the histogram components P is very large, we select a subset from these combinations. The feature selection method used in this paper is as follows;

Algorithm 1. Feature selection using L1-regularized SVM

```
Input: training data, select dimension M, iteration number T(=10), sampling number S(=5000)
current feature set CF = {}
for t=1 to T
    add S combination features to H randomly
    determine parameter of L1-reg.SVM for CF by
     5-fold cross validation
    learn L1-reg.SVM with regards to CF
    remain largest M non zero features and put
      off other features from CF
end for
if |CF| < M then
    add $M - |CF|$ features to CF randomly
end if
Output: M feature combination CF
```

Fisher discriminant criterion: Fisher discriminant score calculated per each possible combined feature is used. Fisher discriminant score J for the feature x_j is defined by $J(x_j) = \sigma_{Bj}/\sigma_{Wj}$. Where σ_{Bj} denotes the between-class variance and σ_{Wj} denotes the within-class variance of feature x_j. We calculate $J(x_j)$ for j = 1,..,P and select the largest M features.

L1 regularized SVM: L2-norm of **w** in Support Vector Machine(SVM) [10] are replaced to L1-norm. L1 regularization generates a sparse solution of **w**. We use implementation in LIBLINEAR [11]. In L1 regularized SVM, the following optimization problem is solved.

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 + C \sum_{i=1}^{l} (max(0, 1 - y_i \mathbf{w}^t \mathbf{x}_i))^2, \tag{1}$$

where $\|\cdot\|_1$ denotes the L1-norm and y_i denotes the class label $\in \{-1, 1\}$ of sample number i. Because SVM is binary classifier, we train SVM by one-against-all and select features per each category. We select features that the absolute values of w are high. The feature selection algorithm using L1-regularized SVM is shown in algorithm 1.

Selection from each operator/ all operators: We select each feature from each operator or all operators. When we are selecting from all operators, the scales of each feature are different. So, we normalize each feature so that the mean of each feature (over all training data) is zero, and the standard deviation is one, i.e. we rescale the feature values x_j to the normalized feature values x'_j, using the relation:

$$x'_j = \frac{x_j - \overline{x_j}}{\sigma_{x_j}}, \tag{2}$$

where $\overline{x_j}$ is the mean feature value, and σ_{x_j} is the standard deviation.

6. Experiment

We performed experiments on Scene-15 dataset [2]. The Scene-15 dataset consists of 4485 images spread over 15 categories. The fifteen categories contain 200 to 400 images each and range from natural scene like mountains and forest to man-made environments like kitchens and office. We selected 100 random images per categories as a training set and the remaining images as the test set. Some examples of dataset images are shown in Fig. 3.

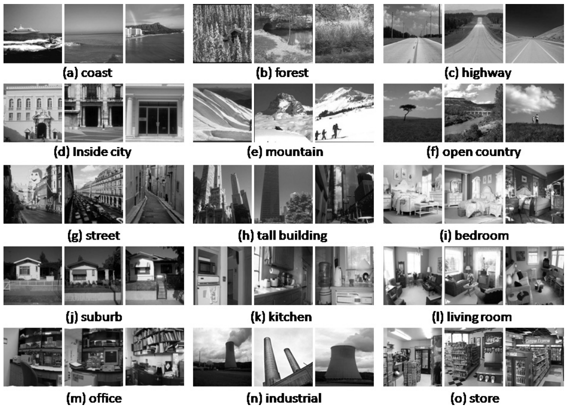

Fig. 3 Example of Scene-15 dataset.

To obtain reliable results, we repeated the experiments 10 times. Ten random subsets were selected from the data to create 10 pairs of training and test data. For each of these pairs a codebook was created by using k-means clustering on training set. Because the cross-validation in feature selection by L1-SVM takes large computation times [†], the result of 1 times is reported with regard to L1-SVM. The combination of spatial pyramid level 1, 2 becomes very large compared to spatial pyramid level 0. Thus, larger iteration number T in Algorithm 1. is required. This means more computation time is needed. Then, we searched feature only for spatial pyramid level 0 with regard to L1-SVM.

As local features, we used a gradient local autocorrelation (CLAC) descriptor [12] sampled on a regular grid. Because GLAC can extract richer information than SIFT [9] descriptor. GLAC descriptor used in this paper is 256-dimensional co-occurrence histogram of gradient direction that contains 4 types of local autocorrelation patterns. We calculated the feature values from a 16×16 pixel patch sampled every 8 pixels, and histogram of each autocorrelation pattern is L2-Hys normalized. In the codebook creation process, all features sampled every 16 pixel on all training images were used for k-means clustering. The codebook size k is set to 400. Below we refer bag-of-features histogram created by this setup as original bag-of-features. Then we added selected combined features to original bag-of-features. As normalization method, we used L1-norm normalization for both bag-of-features and combined feature vectors respectively and concatenated these vectors.

After creating concatenated vectors by fisher discriminant criterion or L1SVM, we train linear SVMs (these are L2 regularized) by one-against-all for classification. As implementation of SVM, we used LIBLINEAR [11]. Five-fold cross validation was carried out on the training set to tune

[†] The computation time depends on the search range of C parameter of SVM. We searched it from { 2^0, 2^1, 2^2,..,2^{15} }, It takes 3 days when 4-core of Xeon 2.83 GHz is used for feature selection of one operator for 15 categories.

Table 2 Recognition rates of scene-15 by fisher discriminant criterion (plus 2000 features). Bold figure shows the best three operators in each pyramid level.

Pyramid Level	Without	Sum	Sub	Div	Prod	Sum(b)	Sub(b)	Prod(b)	Max	Min	HMean
0	59.59	60.39	60.61	61.36	**63.72**	62.75	63.16	63.71	63.43	**65.66**	**64.60**
1	68.52	69.01	69.43	**69.60**	62.93	68.85	**70.00**	**69.98**	69.41	66.19	66.37
2	71.59	71.71	72.76	**73.05**	69.35	**72.97**	72.74	**73.62**	71.42	70.89	70.84

Table 3 Recognition rates of scene-15 by L1-SVM (plus 2000 features). Bold figure shows the best three operators.

Pyramid Level	Without	Sum	Sub	Div	Prod	Sum(b)	Sub(b)	Prod(b)	Max	Min	HMean
0	59.59	63.68	63.19	65.56	**67.31**	64.09	62.99	64.62	62.09	**69.76**	**68.58**

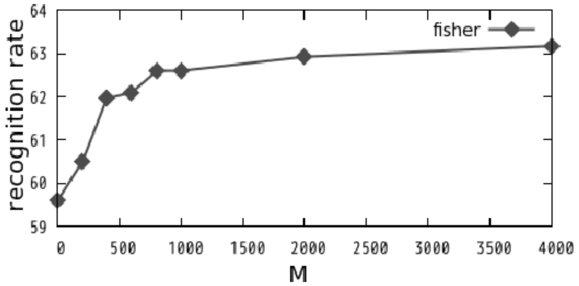

Fig. 4 Average recognition rates of 10 operators in spatial pyramid level 0. Fisher discriminant criterion is used for feature selection. M=0 is original bag-of-features.

the parameters of SVM. The classification rate we report is the average of the per-class recognition rates which in turn are averaged over the 10 random test sets.

6.1 Experimental Results

At first, we check the effect of additional dimension M in spatial pyramid level 0 using feature selection by fisher discriminant analysis. This result is shown in Fig. 4. It is confirmed that recognition rates becomes increase as to increase the number of additional dimension M. So the M can be set by considering trade-off between speed and classification accuracy. Below in this paper, we set M to 2000.

The recognition rates of feature selection by fisher discriminant criterion in 2000 additional dimension is shown in Table 2. It is shown that min operator is the best performance with regard to pyramid level 0. The recognition rates of harmonic mean and product are the next. Binarized operators are also good performances. Summation, subtraction, and division are not effective compared to above operators. The results of pyramid level 1, 2 are slightly different from pyramid level 0. In these cases, division and binarized operators show better performances and operator min, product, and harmonic mean are not effective.

The recognition rates of feature selection by L1-SVM in 2000 additional dimension are shown in Table 3. It is shown that better performances than fisher discriminant criterion are achieved. This is because the fisher discriminant score is calculated per each feature, and L1SVM uses many features combination to train SVM. The non-zero feature numbers of L1SVM were about 200-500. But, we used 2000 features by adding random combination to compare with fisher discriminant criterion in the same dimension.

Table 4 Recognition rates of scene-15 plus 2000 features selected from all operators.

Pyramid Level	Without	ALL-fisher	ALL-L1SVM
0	59.59	63.26	**69.31**
1	68.52	70.14	**72.69**
2	71.59	72.22	**74.18**

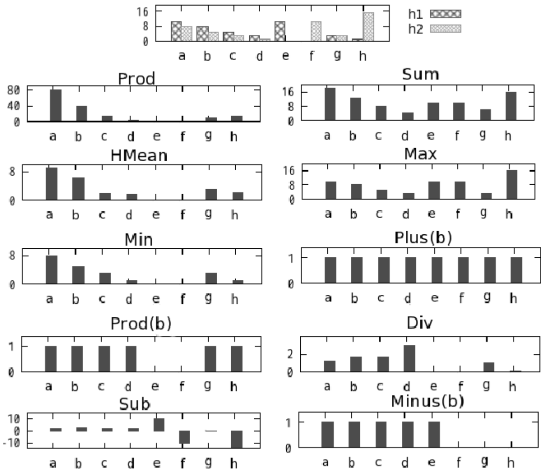

Fig. 5 Toy examples of combined feature values for histogram components h1 and h2.

The classification results by selected from all operators are shown in Table 4. In the case of L1SVM, the features were selected from pyramid level 0 for all pyramid level. By selecting from all operators, the classification rates becomes slightly lower than min operators only in the case of pyramid level 0.

6.2 Discussion

We observed that the type of AND operators (prod, min, harmonic means) show better performance in spatial pyramid level 0. To discuss about each operators' properties, we prepared toy examples of feature values obtained by each operators in Fig. 5. It can be said that the AND operators can produce more sparse features than OR operators. This means more separable features can be obtained by AND operators. The low performance of AND type in spatial pyramid level 1,2 may be caused by the sparseness of the original bag-of-features. In average, 50.57% of the histogram components of the original bag-of-features in spatial pyramid level 0 were zero. On the other hand, 72.75% and 87.74% of the histogram components were zero in spatial pyramid level 1 and 2, respectively. Therefore, the additional sparse features were not required to spatial pyramid level 1,2.

The superiorities of AND operators over DIFF operators can not be theoretical explained. It can be only said that AND type operators were better at least spatial pyramid level 0 by the experiment.

In AND type operators, the min operator showed the best performance. This can be explained by the desirable properties of the min operator. The prod, harmonic means and prod produces similar features (Fig. 5). However, prod is too sensitive to only combination of high histogram components. The prod emphasis combination of large values (Fig. 5.a) and the product of combination of middle/small values becomes relative low (Fig. 5.d,g). On the other hands, the harmonic means and min can more emphasis co-occurrence of middle/low values (Fig. 5.b, c, d, g) than prod. Furthermore, min doesn't sensitive to a high value of only one component (Fig. 5.h) compared to prod and harmonic mean. Thus, the min is the most adequate operator to express co-occurrence relations.

7. Conclusion

In this paper, we proposed a combined feature extraction method for global/local statistics of visual words using 10 relevant operations. Experimental results using Scene-15 dataset show all operations are effective for combined features extraction. Especially, product, min, and harmonic mean operators exhibited high improvements of accuracy for global statistics. Division and binarized operators exhibited high improvement for local statistics.

References

[1] G. Csurca, C. Dance, L. Fan, J. Willamowsli, and C. Bray, "Visual categorization with bags of keypoints," ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[2] S. Lazebnik, C. Schmid, and J. Ponece, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, CVPR, pp.2169–2178, 2006.

[3] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Scene Classification Using Generalized Local Correlations," Eleventh IAPR Conference on Machine Vision Applications, pp.195–198, 2009.

[4] H. Cao, K. Yamaguchi, T. Naito, and Y. Nimomiya, "Pedestrian Recognition using Second-Order HOG Feature," Asian Conference on Computer Vision, 2009.

[5] D. Okanohara and J. Tsuji, "Learning Combination Features with L_1 Regularization," NAAACL HLT, pp.97–100, 2009.

[6] S. Battiato, G. Farinella, G. Gallo, and D. Ravi, "Scene categorization using bag of textons on spatial hierarchy," IEEE International Conference on Image Processing, pp.2536–2539, 2008.

[7] J. Wu and J.M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[8] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification," IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis., vol.60, pp.91–110, 2004.

[10] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

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

[12] T. Kobayashi and N. Otsu, "Image Feature Extraction Using Gradient Local Auto-Correlations," European Conference on Computer Vision, Part I, LNCS 5302, pp.346–358, 2008.