Combined Feature Extraction from Global/Local Statistics of Visual Words using Relevant Operations

Tetsu MATSUKAWA†and Takio KURITA ‡

†:Graduate School of Systems and Information Engineering, University of Tsukuba
‡:National Institute of Advanced Industrial Science and Technology
{t.matsukawa, takio-kurita}@aist.go.jp

Abstract This paper presents a combined feature extraction method to improve the performance of bag-of-features image classification. The bag-of-features approach is the most popular approach for generic object recognition and uses global statistics (histogram) of visual words. To alleviate the loss of spatial layout information, the local statistics of visual words is also used in many researches. However, it is possible to extract richer information from these global or local statistics of visual words. We apply 10 relevant operations to global/local statistics of visual words. Each operation is meaningful to describe the property of the pairwise relationship among histogram components. For example, multiplication indicates strong co-occurrence of different visual words and binarized minus operator indicates bigger/smaller relationship. Because the pairwise combination of visual words is large, we apply feature selection methods including fisher discriminant criterion and L1-SVM. Experimental results using Scene-15 dataset show that min operator achieved the best performance for global statistics, and division and binarized operators achieved better performance for local statistics.

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 are 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. However, several problems exist with regard to its application to image representation. To solve these problems, many methods have been proposed. These methods include spatial pyramid binning that utilizes location information [2, 3], higher level codebook creation based on local co-occurrence of codebooks [4, 5], improvement of codebook creation[6], and incorporate with spatial correlogram[7]. All these methods are based on the global/local statistics (histogram) of local appearance, and further information is not extracted from these histogram representation. In this paper, we call the feature extracted from a combination of predetermined histogram components as combined feature. It is worthful to explore the effectiveness of combined feature extraction in bag-of-features.

In this paper, we propose a combined feature extraction method to improve the performance of the bag-of-features image classification. Proposed method includes 10 operations of pairwise histogram components and feature selection methods. The effectiveness of the proposed feature extrac-

tion for bag-of-features is confirmed through experiments using the popular Scene-15 dataset[2].

The remainder of this paper is organized as follows. Section 2 reviews the related literature on the combined feature extraction. Section 3 and 4 reviews the standard bag-of-features and the spatial pyramid bag-of-features respectively. Sections 5 presents our proposed feature extraction method. Section 6 presents the feature selection method. In Section 7 we presents our experimental results. Finally, we conclude our work in Section 8.

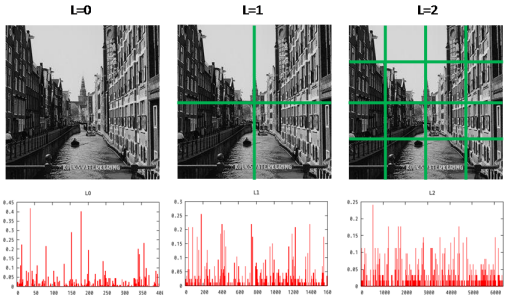

Fig. 1: Spatial pyramid bag of features[2].

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[8] 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[9] 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[10]. 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 Bag-of-Features (Global Statistics)

In this section, we briefly review the standard bag-of-features method [1]. The bag of features method is a classification method using orderless collection of quantized local features. The main step of bag-of-features are :

1. Detection and description of image patches; Dense[2] or random sampling[6] is better than keypoints[1] for scene classification. SIFT descriptor [11] is widely used for its good performance in classification task.

2. Assigning patch descriptors to a set of predetermined clusters (a vocabulary) with a vector quantization algorithm; typically k-means clustering is used.

3. Constructing a bag of features, which counts the number of patches assigned to each cluster

4. Applying a classifier by treating the bag of features as the features vector, and thus determine which category to assign to the image

It is known that the bag-of-features method is robust for background clutter and produces good classification accuracy. One drawback of the standard bag-of-feature method is the limited descriptive ability of spatial information because orderless feature collection does not contain spatial layout of the features.

4 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 [12, 13, 14]. The methods that use overlap grid[13], vertical and horizontal grid[3] 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 blocks are concatenated. For example, a level 2 pyramid have 16+4+1=21 blocks.

5 Combined Feature Extraction

Let k be the number of visual words and $\mathbf{H} = (h_1, ..., h_{21k})$ be the concatenated histogram of spatial pyramid. This paper presents 10 operations for the combinations of the histogram components (h_i, h_j) of $\mathbf{H}$ (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 wheter 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 [9].

Table 1: Relevant operations

operations	definition	explanation
Sum	$h_i + h_j$	OR
Sub	h_i - h_j	difference
Div	$\frac{h_i}{h_j}$	difference rates
Prod	$h_i h_j$	AND
Sum(binary)	binary($h_i + h_j$)	OR
Sub(binary)	binary(h_i - h_j)	big/small
Mul (binary)	binary(h_i h_j)	AND
Max	max(h_i, h_j)	OR
Min	min(h_i, h_j)	AND
Harmonic mean	$\frac{2h_i h_j}{h_i + h_j}$	co-occurrence

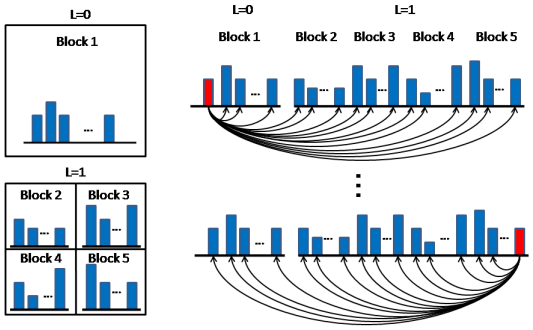

Fig. 2: Pairwise relationship among the histogram components[9].The histogram components of upto level 1 spatial pyramid are shown.

6 Feature Selection

Because the combinations of the histogram components is very large, we select a subset from these combinations. The feature selection method used in this paper is as follows;

Fisher discriminant criterion: Fisher discriminant score for each feature is used. Fisher discriminant score J for the feature i is defined by J(i) $= \sigma_{Bi}/\sigma_{Wi}$. Where σ_{Bi} denotes the between-class variance and σ_{Wi} denotes the within-class variance of feature i. We select largest the M features.

L1 regularized SVM: L2-norm of $\mathbf{w}$ in Support Vector Machine(SVM)[15] are replaced to L1-norm. L1 regularization generates a sparse solution of $\mathbf{w}$. We use implementation in LIBLINEAR[16]. In L1 regularized SVM, the following optimization problem is solved.

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

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:

$$x'_j = \frac{x_j - \overline{x_j}}{\sigma_{x_j}}, \qquad (2)$$

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 H = {}
for t=1 to T
  add S combination features to H randomly
  determine parameter of L1-reg.SVM for H by 5-fold cross validation
  learn L1-reg.SVM with regards to H
  remain largest M non zero features and put off other features from H
end for
if $|\mathbf{H}| < \mathbf{M}$ **then**
  add $M - |H|$ features to H randomly
end if
Output: M feature combination H

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

7 Experiment

7.1 Experimental Setup

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.

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. For classification, a linear SVM was used by one-against-all. As implementation of SVM, we used LIBLINEAR[16]. Five-fold cross validation was carried out on the training set to tune 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.

As local features, we used a gradient local autocorrelation(CLAC) descriptor [17] sampled on a regular grid. Because GLAC can extract richer information than SIFT[11] descriptor. GLAC descriptor used in this paper is 256-dimensional cooccurrence 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. 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.

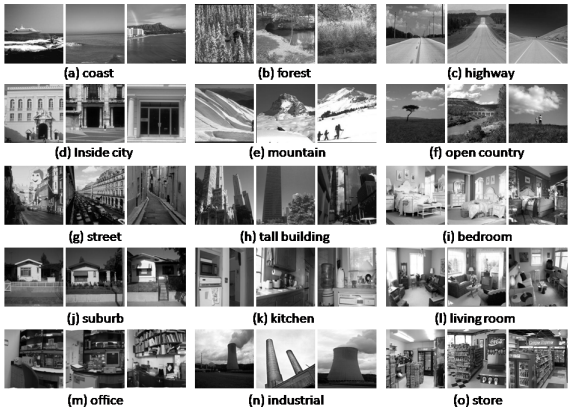

Fig. 3: Example of Scene-15 dataset. Eight of these (a-h) were originally collected by Oliva and Torralba [18], five (i-m) by Fei-Fei and Perona [19], and two (n-o) by Lazebnik et al. [2].

7.2 Experimental Results

The recognition rates of feature selection by fisher discriminant criterion is shown in Table 2. It is shown that all combined feature extraction methods improve the accuracy as to increase the number of the combined features.

The recognition rates of feature selection by fisher discriminant criterion in 2000 additional dimension is shown in Table 3. 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 operatore 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 Ta-

ble 4. 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.

The classification results by selected from all operators are shown in Table 5. By selecting from all operators, the classification rates becomes slightly lower than min operators only in the case of pyramid level 0. In table 6, the selected rates per feature selection methods is shown. It is confirmed that selected rates product, min, and harmonic mean are high in L1-SVM, these operators were good performances in the result of single operator. But in fisher discriminate criterion, these operators are not selected well. This shows that L1-SVM can select better combinations.

8 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 fifteen scene 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.

Our feature work includes extraction of more complex relevant features to extract more richer information from bag-of-features and propose efficient feature selection method.

References

[1] G.Csurca, C.Dance, L.Fan and J.Willamowsli and C.Bray, Visual categorization with bags of keypoints, In 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, In CVPR, pp.2169-2178, 2006.

[3] S.Battiato, G.Farinella, G.Gallo, D.Ravi, Scene categorization using bag of textons on spatial hierarchy, in: IEEE International Conference on Image Processing, 2008, pp.2536-2539.

[4] Y.-T. Zheng, M.Zhao, S.-Y.Neo, T.-S.Chua, Q.Tian, Visual synset: towards a higher-level visual representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp.1–8.

[5] J.Yuan, Y.Wu, M.Yang, Discovery of collocation patterns: from visual words to visual phrases, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp.1–8.

[6] E.Nowak, F.Jurie and B.Triggs, Sampling Strategies for Bag-of-Features Image Classification, In ECCV, pp.490-503, 2006

[7] Y.Zheng, H.Lu, C.Jin, and X.Xue, Incorporating spatial correlogram into bag-of-features model for scene categorization, Asian Conference on Computer Vision, 2009.

[8] H.Nakayama, T.Harada, and Y.Kuniyoshi: Scene Classification Using Generalized Local Correlations. In: the Eleventh IAPR Conference on Machine Vision Applications,pp.195-198, (2009).

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

[10] D.Okanohara, J.Tsuji, Learning Combination Features with L_1 Regularization, in: NAAACL HLT, 2009, pp.97–100.

[11] D.G.Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[12] B.Yao, J.C.Niebles, and L.Fei-Fei: Mining Discriminant Adjectives and Prepositions for Natural Scene Recognition, In: CVPR2009 International Workshop on Visual Scene Undarstanding, 2009.

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

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

[15] V.Vapnik, Statistical Learning Theory, John Wiley & Sones, 1998.

[16] 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 9(2008), 1871-1874.

[17] T.Kobayashi and N.Otsu: Image Feature Extraction Using Gradient Local Auto-Correlations. European conference on computer vision, 2008, Part I, LNCS 5302, pp.346-358 (2008).

[18] A.Oliva, A.Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope , International Journal of Computer Vision 42(3) (2001) 145–175.

[19] L.FeiFei, P.Perona, A bayesian hierachical model for learning natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp.524–531.

Table 2: Recognition rates of scene-15 per feature dimension (fisher discriminant criterion). Bold figure shows the best recognitions rates in each operator.

additional dimension	0	200	400	600	800	1000	2000
Sum	59.51	58.02	60.22	**61.24**	61.05	60.75	60.39
Sub	59.51	59.89	60.18	60.24	60.30	60.43	**60.61**
Div	59.51	60.45	61.01	61.52	**61.90**	61.84	61.36
Prod	59.41	56.74	60.67	60.69	62.24	62.69	**63.72**
Sum(binary)	59.51	64.42	63.92	63.12	**63.97**	63.04	62.75
Sub(binary)	59.51	62.34	63.65	63.27	**63.68**	63.26	63.16
Prod (binary)	59.51	63.78	**64.40**	63.41	63.61	63.09	63.71
Max	59.51	61.48	62.57	61.57	63.18	63.17	**63.43**
Min	59.51	58.77	61.24	63.01	63.27	64.09	**65.66**
Harmonic mean	59.51	59.03	61.77	62.94	62.94	63.80	**64.60**

Table 3: 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(1trial)	71.59	71.71	72.76	**73.05**	69.35	**72.97**	72.74	**73.62**	71.42	70.89	70.84

Table 4: 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**

Table 5: Recognition rates of scene-15 plus 2000 features selected from all operators. Bold figure shows the best recognition rates in each pyramid level.

Pyramid Level	without	ALL-fisher	ALL-L1SVM
0	59.59(± 0.43)	63.26(± 0.72)	**69.31**
1	68.52(± 0.37)	70.14(± 0.60)	**72.69**
2	71.59(± 0.32)	72.22(± 0.37)	**74.18**

Table 6: Selected rates of each operators (pyramid level 0). Bold figure shows operators those selected rate is more than 20%.

	Sum	Sub	Div	Prod	Sum(b)	Sub(b)	Prod(b)	Max	Min	HMean
fisher	6.5	**22.7**	**31.1**	0.0	**30.7**	0.3	0.0	6.5	0.6	1.2
L1SVM-suburb	1.6	3.2	1.0	**30.0**	0	6.0	10.9	0	**24.5**	**22.4**
L1SVM-coast	3.8	4.6	3.3	**42.7**	15.2	2.1	3.3	4.2	12.2	8.0
L1SVM-forest	1.9	1.2	4.5	**29.0**	17.4	10.9	5.1	4.5	12.9	12.2
L1SVM-highway	1.6	1.9	0.6	**36.2**	11.1	7.5	9.4	2.9	18.3	10.1
L1SVM-inside city	0.2	3.2	2.7	**28.4**	8.6	11.1	11.1	2.4	16.8	14.0
L1SVM-mountain	2.8	2.3	3.1	**35.5**	10.4	5.2	6.3	3.1	15.3	15.6
L1SVM-open country	1.2	4.1	1.9	**34.8**	19.3	2.2	7.0	1.9	11.9	15.1
L1SVM-street	0.7	1.4	0	**29.0**	6.0	9.6	8.9	1.4	**24.0**	18.6
L1SVM-tall building	2.0	3.3	2.0	**32.7**	14.7	4.0	11.0	4.0	13.7	12.3
L1SVM-office	0.4	1.4	2.3	**21.3**	7.1	5.6	14.2	0	**23.6**	**23.6**
L1SVM-bedroom	0.2	1.1	0.5	**29.9**	9.6	10.8	9.9	2.7	**22.1**	12.7
L1SVM-industrial	2.3	3.4	1.7	**29.2**	9.0	10.0	14.2	2.6	17.7	9.4
L1SVM-kitchen	0.9	1.2	5.1	**27.2**	6.3	9.0	13.3	2.1	15.6	18.7
L1SVM-livingroom	0.2	0.5	3.5	**22.8**	6.1	12.3	14.1	0.7	**22.5**	16.9
L1SVM-store	0.6	0.3	0.9	**30.3**	7.9	10.4	11.0	2.7	**20.2**	15.3