Object Categorization Utilizing a Codebook Containing Contextual Information of Visual Words

Shuang BAI Yoshinori TAKEUCHI Hiroaki KUDO Noboru OHNISHI

Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
E-mail: {baishuang,kudo,takeuchi,ohnishi}@ ohnishi.m.is.nagoya-u.ac.jp

Abstract In object categorization, bag of visual words is a promising approach. However, in this framework how to obtain discriminative codebook is still an open issue. Since contextual information can be used to reduce ambiguity in object recognition, in this report we propose to build a codebook which takes contextual information of visual words into consideration. Utilizing a codebook in which both the appearance of visual words and their contextual information are contained would help to improve image representation. We first detect interest points in images employing Harris-Laplacian detector, then from each detected point we extract patches of different scales, which are described using SIFT descriptor. After that, based on these extracted patches we build a hierarchical codebook in which visual words in different levels are related, and higher level visual words contain contextual information of lower level visual words. Through this codebook, image representations which are more discriminative and robust could be created. We compared our method with two baseline approaches, and results indicated the effectiveness of our proposed method.

Keyword Object categorization, Bag of Visual Words, Hierarchical codebook, Contextual information

1. Introduction

Object categorization is a challenging problem in the field of computer vision, and is also of great application significance. The objective of generic object categorization is to recognize object classes instead of object instances. Generally, this process involves coping with view, lightening change, object occlusion and clutter as well as intra class variation, all of which are typical for objects in real world. Therefore, a qualified categorization system should be able to capture common features of objects from the same category, at the same time be discriminative with respect to features from different categories [8].

Early works on object categorization mainly focused on global features, such as color and texture, which are extracted to represent image contents [5], [9]. However, because of the large intra class variation of objects, the performance of approaches based on global features is limited. On the other hand, object representation based on discriminative local features has outperformed global features in this field. Promising results to categorize objects using local features were demonstrated in works [1], [4], [6], [7], [8], [12]. Usually, in methods based on local features, a set of image patches are extracted based on salient point detectors [10], [11] or densely [20]. And then, extracted image patches are described by descriptors, like SIFT [10]. Finally, these obtained local patch features are used for representing images.

Recently, an approach called bag-of-visual-words [4] is proposed for object categorization. In bag of visual words framework, first a codebook of visual words would be obtained through applying vector quantization to descriptors of image patches extracted from training images. And then, in image representation stage, an image is represented by assigning extracted local patches from this image to their nearest visual words in the codebook. Consequently, each image is represented as a histogram indicating the frequency of each visual word appearing in the image. This procedure has shown to be robust and characteristic for images represented in object categorization field [6], [7].

However, in the bag of visual words framework, since images are encoded as a collection of quantized visual words, lots of information gets lost in this procedure. Ambiguity may arise in assigning image patches to visual words in the codebook. For example, patches which are similar in appearance but of different semantic interpretation may be assigned to the same visual words. It is quite possible for this process to make the classification performance deteriorate. So in this paper, to alleviate the ambiguity caused by the above mentioned reason, we propose to create a codebook which is augmented with enhanced discriminative capacity by using hierarchical structure with each level containing visual words of different coarseness. In the proposed procedure, we take contextual information of image patches into account in the process of visual words creation and image representation. To achieve this goal, after salient regions in images are detected using Harris-Laplacian [11] detector, at each detected point besides the patch with detected scale, we extract another two sets of patches with larger scales. The coarse scale patches represent image region information coarsely, while fine scale patches and intermediate scale patches can represent image information more precisely. Then, all patches are described by SIFT descriptor. Based on these features, a hierarchical codebook with each hierarchy containing visual words of different coarseness is constructed. We first apply k-means to the

coarse scale patches to create a set of clusters representing region information coarsely. And then we sort fine scale patches and intermediate scale patches into groups based on their corresponding relationship with the coarse scale patches. For each intermediate scale or fine scale patch group, we continue to apply k-means to it to obtain a set of fine scale visual words and intermediate scale visual words, which would be used to represent images. The hierarchical codebook is created by combining intermediate scale visual words and fines scale visual words with their corresponding relationship recorded. And in image representation stage, the pairs of intermediate scale patch and fine scale patch extracted from the same interest points would be assigned to the hierarchical codebook together, with the corresponding relationship among visual words in different hierarchies considered. The novelty of our proposed method lies in that we construct a hierarchical codebook in which visual words are related based on the corresponding relationship among patches extracted from the same point, and in image representation stage, patches are assigned to the codebook in pair with corresponding relationship among visual words considered. Consequently, patches assignment to codebook could be more stable and obtained image representation could be more discriminative.

2. Related work

The bag of visual words approach for object categorization is motivated from bag of words method for text categorization [14]. In [4], Csurka et al. proposed to identify objects through bag of key points, which is based on a codebook obtained through vector quantizing affine invariant descriptors of image patches. Naïve Bayes and SVM classifiers are used in their work for classification. However, since image representation based on the initial visual words framework make a lot of information get lost. Many works are proposed to improve its performance. Simple geometrical relationships are added to improve classification accuracy [16]. Lazebnik et al. [12] used nearby regions which have a high frequency to appear in the training samples as semi-local parts. These semi-local parts are used to represent object classes. In [18], a new local structural context descriptor is designed for object categorization to capture the relationship between current point and remaining points, so that to some extent, the structural of the image can be represented. The paper [15] presents to represent images in local visual and semantic concept based feature spaces. In their framework through exploiting local neighborhood structure of the codebook, local concept correlation statistics and spatial relationships in individual encoded images, images are represented in correlation and spatial relationship enhanced concept feature spaces. They utilized the intrinsic correlation existing in the codebook constructed via self-organizing map and the spatial relationship among patches in each individual image to improve the performance of image retrieval.

Besides incorporating spatial relationship among patches into bag of visual words framework, researchers tried to utilize contextual information for improving categorization performance. Yang [17] proposed a mechanism to assess roles of context features for different object recognition tasks, by analyzing information entropy and data ambiguity. Based on the evaluation result of the proposed assessing method, they put weights to different context features as well as object appearance features. Finally, combined context features and appearance features are taken as the image representation. Hence, useful features will have more impact on the categorization process. Mehdi et al. [13] used contextual guided bag of visual words model. In their approach, after image patches are extracted from interest points, a contextual space and a feature space are defined separately. Thereafter, a merging process is employed to fuse features based on their proximity in contextual space. In this way a set of visual words with contextual information are created and used for classification. In [19], Qin and Yung proposed contextual visual words approach. They proposed to combine a patch of interest with its coarser scale patch and neighbor patches, so that obtained feature could be more discriminative.

In addition, codebooks which are more powerful are also designed, in the aim of getting discriminative image representation. Methods for generating compact codebooks via informative feature selection has been proposed [21][22]. However, although what has been discarded is less informative features, performance of codebooks obtained this way demonstrated no satisfactory performance gain. An approach which is similar to our work is [23]. In this paper a multi-sample multi-tree approach to computing visual words codebook is proposed. They extracted several complementary patches around the same point as a visual packet and for each type of patch sampling, a specific codebook is created. In image retrieval process, two visual packets need to have full match to be taken as identical. Otherwise, this visual packet would be discarded. Their visual packet representation is equivalent to a fine partition of the joint feature space of patches in the visual packet. Although these visual packets are more discriminative, it is made unstable. And its efficiency in image classification is not proved. On the other hand, we have employed pairs of patches with different scales extracted from the same points. We utilized their corresponding relationship to construct a codebook which contains contextual information of visual words. And when a pair of patches is assigned to the codebook, the ambiguity of patch assignment can be reduced by taking the corresponding relationship of pairs of patches into consideration. Our approach is designed to reduce patch assignment ambiguity and make resulted image representation more robust.

3. Image representation based on a codebook containing contextual information

As mentioned above, in traditional bag of visual words method visual words have ambiguity. They may get patches of similar

appearance but of different semantic interpretation to be assigned in a confusing way. This may play a negative role in the performance of categorization. Intuitively, a simple idea to overcome this shortcoming is to construct a hierarchical codebook with visual words in higher hierarchy containing contextual information of visual words in lower hierarchy. Therefore, in the stage of image representation, both the appearance and the contextual information of a patch can be taken into consideration. Consequently, when the patches are assigned in pair to the codebook, ambiguity could be alleviated.

3.1 Building a hierarchical codebook containing contextual information

To build the hierarchical codebook, we first detect interest points using Harris-Laplacian detector from training images. At each detected interest point, besides patches with detected scale s_d, we extract another two sets of patches, whose scales are larger than the detected scale, $s_c = s_d \times \alpha$, $s_m = s_d \times \beta$, $\alpha > \beta > 1$. We call patch with detected scale s_d fine scale patch, the largest patch with scale s_c coarse scale patch which is used to represent local region information coarsely, patch with scale s_m intermediate scale patch which contains contextual information of fine scale patch.

After we have extracted patches with different coarseness, we first cluster the coarse scale patches to get a number of clusters which can represent region information coarsely. After that, we sort intermediate scale patches and fine scale patches into groups based on their corresponding relationship with coarse scale patches. For each cluster of coarse scale patches, we get a group of intermediate scale patches and a group of fine scale patches. Based on the sorted patch groups, from each intermediate scale patch group and each fine scale patch group, we create a set of intermediate scale visual words and a set of fine scale visual words. And then, we combine all the obtained intermediate scale visual words and fine scale visual words in a hierarchical codebook, with intermediate scale visual words in the higher level and fine scale visual words in the lower level. At the same time, the set of intermediate scale visual words constructed from a coarse scale patch cluster would be related to the set of fine scale visual words created from the same coarse scale patch cluster. Since we sorted the intermediate scale patches and fine scale patches based on the clustering result of the coarse scale patches, outliers may exist in each sorted group. Before the visual words creation step, we apply K-nearest neighbor method over each group of intermediate scale patches and fine scale patches to reduce outliers. The whole procedure is described in Fig.1. Through utilizing the above codebook construction process, different levels of visual words may have different coarseness, and higher level visual words may contain contextual information of lower level visual words. In this framework, both the appearance information and contextual information of a patch would be taken into consideration. Images would be represented by assigning patches to the related visual words of the codebook in pair.

3.2 Image encoding utilizing a codebook containing contextual information

In image representation stage, we use intermediate scale patches and fine scale patches to create image histogram features. Now we propose a method to assign patches extracted from the same interest point to the hierarchical codebook in pair, so that the assignment of patches is based on both the fine scale patch and the intermediate scale patch which provides contextual information for the fine scale patch. In the image representation stage, for each fine scale patch and its related intermediate scale patch, first we find their nearest visual words in the corresponding codebook hierarchy respectively. If the pair of patches are assigned to intermediate scale visual words and fine scale visual words corresponding to the same coarse scale patch cluster, both patches would be assigned to their corresponding visual words with a high confidence with the value of 1. However, when the pair of patches is assigned to intermediate scale visual word and fine scale visual word corresponding to different coarse scale patch clusters, the two patches would not be assigned to their nearest visual words directly, since there is ambiguity for the assignment of the pair of related patches. Instead, we find a sub-nearest visual word for each patch in the pair in their corresponding level of the codebook. The sub-nearest visual word for a patch would be searched from the set of visual words created from the coarse scale patch cluster corresponding to its related patch in the pair. Fig 2 demonstrated the process for assigning pairs of patches to the hierarchical codebook, with their corresponding relationship taken into consideration. In Fig. 2 solid arrow denotes the nearest visual word for a patch, while the dashed arrow denotes sub-nearest visual word for a patch. In this case, to relieve ambiguity in patch assignment, while we assign each patch to its corresponding nearest visual word, we also assign it to its sub-nearest word. And the assignment of the nearest visual word and sub-nearest visual word is weighted by feature to word distance. Assignment weights are calculated as follows: For a patch i with an intermediate scale or a fine scale, and its related patch j with a fine scale or an intermediate scale, we first calculate its nearest visual word weight factor denoted as NWFactor as follows,

$$\mathrm{NWFactor}_i = \frac{1}{d_{js}} \exp(1/d_{in}) \quad (1)$$

where d_{in} is the distance between patch i and i's nearest visual word and d_{js} denotes the distance between patch j and j's sub-nearest visual word. And then we calculate weight factor for patch i's sub-nearest visual words denoted as SNWFactor,

$$\mathrm{SNWFactor}_i = \frac{1}{d_{jn}} \exp(1/d_{is}) \quad (2)$$

where d_{jn} is the distance of patch j and j's nearest visual words, and d_{is} is the distance between patch i and i's sub-nearest visual words. And then, the weight of patch i's nearest visual word and sub-nearest

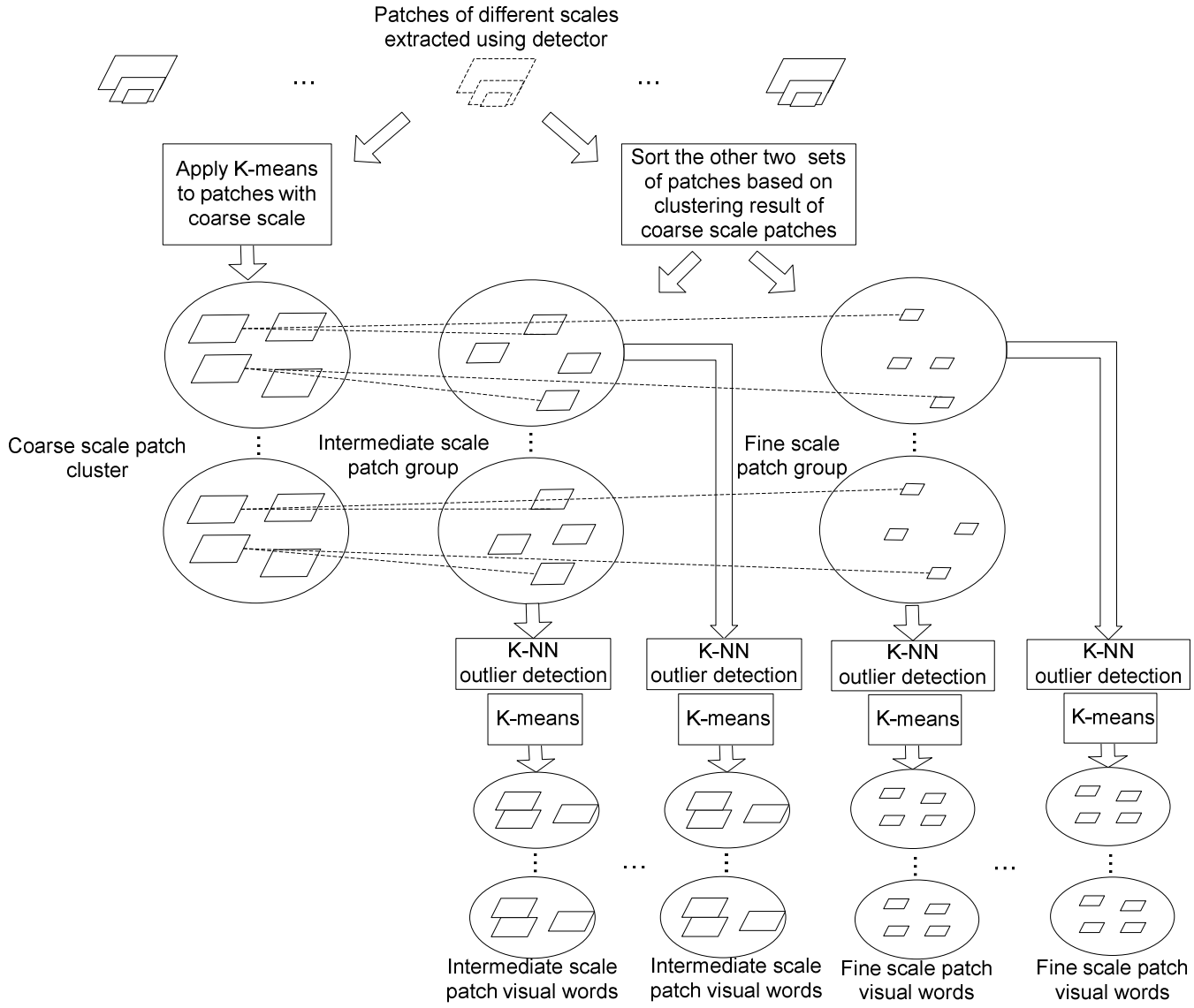

Fig. 1 Hierarchical codebook creation based on images patches of different coarseness

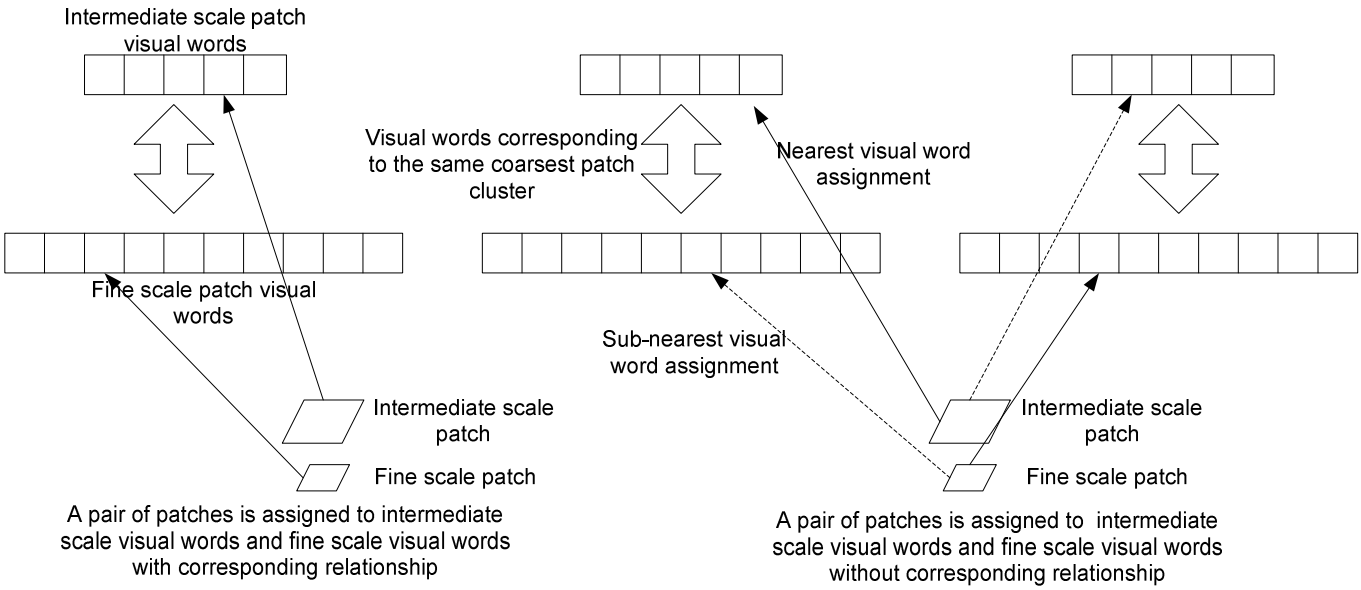

Fig. 2 Patch assignment to hierarchical codebook

visual word are calculated below:

$$\mathrm{NW_i weight} = \frac{\mathrm{NWFactor_i}}{\mathrm{NWFactor_i + SNWFactor_i}} \quad (3)$$

$$\mathrm{SNW_i weight} = \frac{\mathrm{SNWFactor_i}}{\mathrm{NWFactor_i + SNWFactor_i}} \quad (4)$$

where NW weight and SNW weight are values added to image feature histogram for nearest visual word and sub-nearest visual word of a patch. The patch assignment process is summarized in algorithm 1, $\mathrm{bin}(W_i)$ denotes the W_ith bin value in the image feature histogram.

Algorithm 1. Assigning patches to codebook in pair

1 Initialize image histogram feature
2 **for** all pairs of patches (p_i, q_i) exacted from an image
3 find nearest word NW_{p_i} for intermediate scale patch p_i
4 find nearest word NW_{q_i} for fine scale patch q_i
5 find coarse scale cluster index C_{p_i} for word NW_{p_i}
6 find coarse scale cluster index C_{q_i} for word NW_{q_i}
7 **if** $C_{p_i} == C_{q_i}$
8 $\mathrm{bin}(NW_{p_i}) = \mathrm{bin}(NW_{p_i}) + 1$
9 $\mathrm{bin}(NW_{q_i}) = \mathrm{bin}(NW_{q_i}) + 1$
10 **else**
11 $\{W_{m1} \dots W_{mn}\}$ = intermediate scale words set (C_{q_i})
12 $\{W_{f1} \dots W_{fk}\}$ = fine scale words set (C_{p_i})
13 find sub-nearest word SNW_{p_i} from $\{W_{m1} \dots W_{mn}\}$ for p_i
14 find sub-nearest word SNW_{q_i} from $\{W_{f1} \dots W_{fk}\}$ for q_i
15 calculate words weight for NW_{p_i}, NW_{q_i}, SNW_{p_i}, SNW_{q_i} based on feature to word distance
16 add word weights to image histogram feature
17 **end if**
18 **end for**
19 **return** image histogram feature

4. Experiments and results

In the following part, we conduct experiments on two sets of images which are selected from dataset Caltech101 [3] and Caltech256 [25] respectively. To evaluate the performance of the proposed method, we compare it with traditional bag of visual words approach and the recently proposed contextual words approach [19]. LIBSVM [2] is used for classifying images. Results of each method are evaluated through 5-cross validation. For the proposed hierarchical codebook, we used 100 coarse scale patch clusters, and from each coarse scale patch cluster we create 5 intermediate scale patch visual words and 10 fine scale patch visual words, which are combined as the final codebook. For the traditional bag of visual words method, it is based on a codebook of 1000 visual words obtained by applying k-means to image patch descriptors extracted from training images using Harris-Laplacian detector. Another baseline is the contextual words approach. In the contextual words approach, features obtained by combining the patch of interest and its contextual information from coarser scale patch and neighbor patches are used to create a codebook and to represent images. For comparison, we only adopt their procedure to incorporate contextual information from coarser scale patch of the patch of interest. In this method images are divided into regular grids in 5 levels of scales, from scale level 5 to scale level 1, 2000, 1000, 500, 200, 50 visual words are created.

Object categorization results on Caltech 101 are demonstrated in table 1 and fig. 3. As we can see from the categorization result, for most of the categories used the proposed method and the contextual words method outperformed the traditional bag of visual words approach. This result demonstrated that in the bag of visual words framework combining patch of interest and its contextual information in codebook construction and image representation can improve categorization performance. However, for different categories the improvement is different, which indicates that the usefulness of contextual information in different categories can vary. And we observe that some categories gained better performance based on our proposed method while other categories can have superior result on contextual words method. In fact, in the proposed method the combination of the patch of interest and its contextual information is loose, while in contextual words approach the combination of the patch of interest and its contextual information is tight. The experiment results indicate that in different categories the patch of interest and its contextual information should be combined to different extent .

In fig. 4 and table 2, categorization performance on Caltech 256 is demonstrated. The general categorization performance has deteriorated than that of Caltech 101. This is because Caltech 256 has higher intra-class variability, higher object location variability within the image and more cluttered background. We can also see that the improvement for contextual words than traditional bag of visual words method is less than results of dataset Caltech 101. And superiority of our method to contextual words approach is more obvious. This is because variation in images of dataset Caltech 256 is larger. Therefore, our method which combines patch of interest and its contextual information in a more flexible way than that of contextual words method achieves even better performance. Average accuracy results in table 1 and table 2 shows that our proposed method achieves higher categorization accuracy than the other two baseline methods. It confirms the effectiveness of the proposed method. And under the same experiment condition, we evaluated the selected subset using Naïve-Bayes Nearest-Neighbor approach [24]. However, results even worse than the contextual words method were observed for the Naïve-Bayes Nearest-Neighbor approach. We believe this is caused by the property that the performance of Naïve-Bayes Nearest-Neighbor approach may be affected by the training data easily.

5. Conclusion

In this paper, we proposed to construct a codebook in which not only appearance of visual words but also their contextual information is contained. The codebook is constructed by first clustering coarse scale patches. And then, based on the clustering result another two sets of patches with intermediate scale and fine scale would be sorted, and used to create final visual words. These obtained final visual words are combined into the hierarchical codebook, with higher level visual words containing contextual information of lower level visual words, and visual words in different levels are related. After that, in image representation stage we assign patches extracted from the same interest point to the hierarchical codebook in pair, so that patch assignment ambiguity exiting in the bag of visual words framework could be alleviated. We compared the proposed method with two baseline procedures. Experiment results have demonstrated the effectiveness of our proposed method and indicated that for different categories the patch of interest and its contextual information should be combined to different extent to get better result.

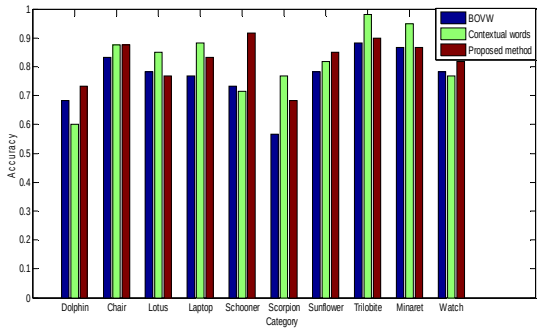

Fig. 3 Performance comparison on Caltech 101 dataset

Table 1 Average accuracy on Caltech 101 dataset

Traditional BOVW	76.80
Contextual words	82.04
Proposed method	82.38

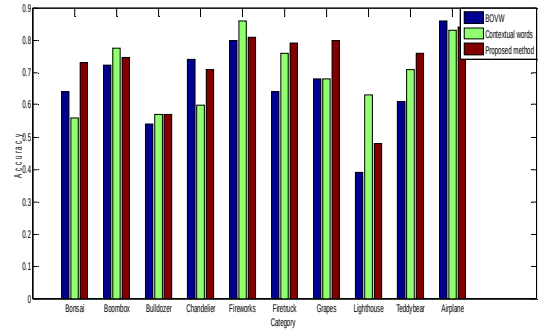

Fig. 4 Performance comparison on Caltech 256 dataset

Table 2 Average accuracy on Caltech 256 dataset

Traditional BOVW	66.32
Contextual words	69.75
Proposed method	72.35

References

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. IEEE. Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp.2169-2178, 2006.

[2] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsm.

[3] 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," IEEE Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision, 2004.

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

[5] A. Vailaya, A. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing," IEEE Transactions on Image Processing, vol. 10, no. 1, pp. 117–130, 2001.

[6] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30 no. 7, pp. 1243–1256, 2008.

[7] R. Fergus, P. Perona, A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," Proc. IEEE computer society conference on Computer Vision and Pattern Recognition , vol. 2, pp. 264-271, 2003.

[8] B. Ommer and J.M. Buhmann, "Object Categorization by Compositional Graphical Models," Proc. Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 235-250, 2005.

[9] M. Szummer, R.W. Picard, "Indoor-outdoor image classification," Proc. IEEE Workshop on Content-based Access of Image and Video Databases, pp. 42–50, 1998.

[10] D.G. Lowe, "Distinctive image features from scale-invariant key points," International Journal of Computer Vision, vol. 60, no.2, pp. 91–110, 2004.

[11] K. Mikolajczyk, C. Schmid, "Scale & affine invariant interest point detectors," International Journal on Computer Vision, vol. 60, pp. 63- 86, 2004.

[12] S. Lazebnik, C. Schmid, J. Ponce, "Semi-local affine parts for object recognition," Proc. the British Machine Vision Conference, vol. 2, pp. 959-968, Kingston, UK, September 2004.

[13] M. Mirza-Mohammadi, S. Escalera, P. Radeva, "Contextual Guided Bag of Visual Words Model for Multi-class Object Categorization," CAIP, pp. 748-756, 2009

[14] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classificati on," Proc. the Seventeenth International Conference on Machine Learning, 2000.

[15] M.M. Rahman, P. Bhattacharya and B.C. Desai, "A unified retrieval framework on local visual and semantic concept-based feature spaces", Journal of Visual Communication and Image Representation, Vol.20, no.7, pp. 450-462.

[16] G. Csurka, C. R. Dance, F. Perronnin, J. Willamowski, "Generic Visual Categorization Using Weak Geometry,"

Proc. Toward Category- Level Object Recognition, pp. 207-224, 2006.

[17] L. Yang, N.N. Zheng and J. Yang, “A unified context assessing model for object categorization” Computer Vision and Image Understanding, Vol.115, no. 3, pp. 310-322, 2011.

[18] W. Liu and Y.B. Yang, “Structural Context for Object Categorization,” Proc. the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, pp. 280-291, 2009.

[19] J.Z. Qin and N.H.C. Yung, “scene categorization via contextual visual words,” Pattern Recognition, Vol. 43, no. 5, pp. 1874-1888.

[20] E. Nowak, F. Jurie and B. Triggs, “Sampling Strategies for Bag-of-Features Image Classification,” ECCV (4), pp. 490-503, 2006.

[21] E. Bart and S. Ullman. “Class-based matching of object parts,” CVPR Workshop on Image and Video Registration, 2004.

[22] T. Tuytelaars and C. Schmid, “vector quantizing feature space with a regular lattice,” ICCV, 2007.

[23] Z. Wu, Q. Ke, J. Sun, and H.Y. Shum, “A Multi-sample, Multi-tree Approach to Bag-of-words Image Representation for Image Retrieval,” ICCV, 2009.

[24] O. Boiman, E. Shechtman, M. Irani, “In Defense of Nearest-Neighbor Based Image Classification,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[25] G. Griffin, AD. Holub, P. Perona, The Caltech-256, Caltech Technical Report, 2006.