PAPER

Incorporating Contextual Information into Bag-of-Visual-Words Framework for Effective Object Categorization*

Shuang BAI[†a)], *Nonmember*, **Tetsuya MATSUMOTO**[†], **Yoshinori TAKEUCHI**[†], **Hiroaki KUDO**[†],
and **Noboru OHNISHI**[†], *Members*

SUMMARY Bag of visual words is a promising approach to object categorization. However, in this framework, ambiguity exists in patch encoding by visual words, due to information loss caused by vector quantization. In this paper, we propose to incorporate patch-level contextual information into bag of visual words for reducing the ambiguity mentioned above. To achieve this goal, we construct a hierarchical codebook in which visual words in the upper hierarchy contain contextual information of visual words in the lower hierarchy. In the proposed method, from each sample point we extract patches of different scales, all of which are described by the SIFT descriptor. Then, we build the hierarchical codebook in which visual words created from coarse scale patches are put in the upper hierarchy, while visual words created from fine scale patches are put in the lower hierarchy. At the same time, by employing the corresponding relationship among these extracted patches, visual words in different hierarchies are associated with each other. After that, we design a method to assign patch pairs, whose patches are extracted from the same sample point, to the constructed codebook. Furthermore, to utilize image information effectively, we implement the proposed method based on two sets of features which are extracted through different sampling strategies and fuse them using a probabilistic approach. Finally, we evaluate the proposed method on dataset Caltech 101 and dataset Caltech 256. Experimental results demonstrate the effectiveness of the proposed method.

key words: *object categorization, bag of visual words, contextual information, hierarchical codebook*

1. Introduction

Research on object categorization is of great application significance in image retrieval and autonomous agents. As a challenging problem in the field of computer vision, it has been attracting more and more attention. The objective of generic object categorization is to recognize the class that an object belongs to rather than the specific object instance. Generally, a qualified object categorization system should be able to cope with view and lighting change, object occlusion and background clutter as well as intra-class dissimilarity and inter-class similarity, all of which are typical for objects in the real world [8].

At first, global features extracted from the whole image, such as colour and texture, were employed to represent images [5], [9]. However, because of the challenges mentioned above, the performance of approaches based on global features is not quite satisfactory. In contrast, object representations obtained by using discriminative local features have demonstrated their superiority over global features in image classification. Promising results to categorize objects by utilizing local features were shown in previous works [1], [4], [6]–[8], [12]. Usually, in approaches based on local features, a set of image patches are extracted first by applying an interest point detector [10], [11] or dense sampling [18]. Then, descriptors like SIFT [10] are employed to describe each extracted local patch, based on which image representations are created in the later stage.

In local feature based approaches, the method called bag-of-visual-words [4], after it was proposed, has become the most notable one. In the bag of visual words framework, first a codebook of visual words is constructed by applying vector quantization to image patch features extracted from training images. After that, images are represented on the basis of the codebook. In the image representation stage, each patch extracted from an image is described and assigned to its nearest visual word in the codebook. As a consequence, an image is represented as a histogram indicating the frequency of each visual word appearing in the image. This procedure has been shown to be able to produce robust and characteristic image representations for object categorization [6], [7].

However, in the bag of visual words framework, since images are represented as a collection of quantized visual words, much information gets lost. Therefore, ambiguity may arise in encoding image patches by assigning them to their nearest visual words in the codebook. For instance, patches which are similar in appearance but of different semantic interpretation may be given the same visual word label. It is quite possible for this ambiguity to make the image representation less discriminative, and result in deterioration in the classification performance.

A well-known technique to reduce quantization error is soft assignment [23], [24], where each image patch feature is assigned to a number of visual words in the codebook. In order to alleviate the ambiguity caused by the reasons mentioned above, we also adopt the soft assignment strategy. However, the main point of our work to use soft assignment is to utilize patch-level contextual information. In the proposed method, the encoding of image patch features is determined not only by the distance between patch features and visual words but also by their patch-level contextual information.

Manuscript received March 13, 2012.
Manuscript revised July 7, 2012.

[†]The authors are with the Department of Media Sience, Graduate School of information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.

[*]This paper was presented at MIRU 2011.

a) E-mail: baishuang@ohnishi.m.is.nagoya-u.ac.jp

DOI: 10.1587/transinf.E95.D.3060

In the proposed method, we use related patches to construct a hierarchical codebook in which visual words in different hierarchies are associated to each other. In addition, in the proposed hierarchical codebook, visual words in the upper hierarchy can provide contextual information to their associated visual words in the lower hierarchy. Thereafter, in the image representation stage, patches extracted at the same sample point are taken as a patch pair. Finally, all the paired patches from an image are encoded using visual words of the hierarchical codebook, with the relationship among associated visual words considered. Moreover, to explore image information effectively, we implement the proposed method on features extracted through regularly sampling and an interest point detector, respectively, and fuse these two kinds of features through a probabilistic approach.

The hierarchical codebook is used to reduce patch assignment ambiguity. Although visual words in our work do not have explicit semantic meaning, in order to make the effect of the proposed method clear, we give the following intuitive explanation. For instance, a coarse scale visual word from patches representing human faces are closely related to fine scale visual words representing eyes and fine scale visual words representing noses. In the process of image patch encoding, a coarse scale image patch is related to a fine scale image patch, and these two patches are encoded together. When the coarse scale patch is assigned to visual words representing human faces, and its corresponding fine scale patches are assigned to visual words representing human eyes, it is highly confident that the two patches are assigned correctly. On the contrary, if the coarse scale patch is assigned to visual words representing human faces, while its corresponding fine scale patches are assigned to visual words that have nothing to do with human, the confidence for the assignment is low. Therefore, based on the hierarchical codebook, relationship among patches can be utilized for reducing encoding ambiguity.

The novelties of the proposed method are as follows. First, we use related patches to construct a hierarchical codebook, in which visual words in the upper hierarchy contain contextual information of their associated visual words in the lower hierarchy. Secondly, in the image representation stage, patches are assigned to visual words in the codebook with patch-level contextual information considered. Thirdly, we fuse image features extracted based on different sampling strategies through a probabilistic approach.

2. Related Work

The bag of visual words method for object categorization is motivated by the bag of words method for text categorization [14]. In the work by Csurka et al. [4], it is proposed to classify objects by means of bag of key points. In their work, images are represented as a collection of visual words of a codebook which is obtained by applying vector quantization to affine invariant descriptors of image patches. However, since much information from the original image gets lost in this representation, after the method is proposed, many researchers have tried to improve its performance by designing more discriminative representations.

A method to represent images in local visual and semantic concept-based feature space [15] was proposed. In this framework, the authors utilized the intrinsic correlation among visual words of a codebook constructed via self-organizing map and the spatial relationship among patches in an image to make the image representation more discriminative. Additionally, Lazebnik et al. proposed spatial pyramid matching technique. They partitioned an image into increasingly finer sub-regions and computed histograms of local features for each sub-region [1]. Through this approach, spatial information of patches in an image is able to be incorporated into the final image representation.

Besides using spatial information, researchers tried to incorporate other contextual information into bag of visual words framework. How to utilize contextual information appropriately is investigated by Yang [16]. They presented a mechanism to assess roles of context features for different object recognition tasks, by analysing information entropy and data ambiguity. Based on the evaluation result, different weights are assigned to each set of context features, so that useful features will have more impact on the categorization process. In addition, Mirza-Mohammadi et al. [13] presented to incorporate contextual information in the codebook construction step. 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 feature words based on their proximity in contextual space. As a result, contextual-guided bag of visual words are created and used for object categorization.

Another work on contextual information was done by Qin and Yung [17] for using contextual visual words for scene classification. In their work, traditional bag of visual words is extended by incorporating contextual information from coarser scales and neighbourhood regions to local regions of interest. They combined a patch of interest with its contextual information to obtain a feature which is believed to be more discriminative. By utilizing the proposed model, image classification performance has outperformed methods based on traditional visual words significantly. However, none of previous approaches considered to utilize patch-level contextual information to create associated visual words and encode paired patches with their corresponding relationship explored.

Codebooks which are more powerful and compact are also designed in the aim of creating more discriminative image representations. Words selection [21] is one of the popular approaches. In this kind of approaches, the most informative words are selected from all created visual words to remove redundancy and noise based on criteria like mutual information and odds ratio. However, although what is discarded is less informative words, image representations based on codebooks obtained this way becomes less discriminative. As a result, final classification performance demonstrated no satisfactory performance gain. This kind of codebooks are designed to make each individual visual word more discriminative, no additional information is involved in the constructed codebook and image representations.

An approach similar to ours is proposed in the work by Wu et al. [19]. In this paper, a multi-sample, multi-tree approach is adopted. They extracted several complementary patches around the same sample point and took them as a visual packet. Then, for each type of sampled patches, a specific codebook is created. At last, the encoding of a visual packet is determined by all the patches in the visual packet. Their approach is equivalent to a fine partition of the joint feature space of patches in the visual packet. No relationship among visual words was considered in their work. On the contrary, we sample patches of different scales at the same sample point to utilize their corresponding relationship to construct a hierarchical codebook in which visual words in different hierarchies are associated. And in the image representation stage, patches extracted from the same sample point are taken as pairs, and encoded with the relationship among visual words considered. Our approach is designed to reduce patch assignment ambiguity and make resulted image representations more robust and discriminative.

3. Image Categorization by Incorporating Relationship among Patches into Image Representations

In order to construct a hierarchical codebook in which visual words in the upper hierarchy contain contextual information of visual words in the lower hierarchy, we propose to extract a set of patches of different scales at the same sample points. Then, patches extracted from the same sample point are related to each other. After that, we create coarse scale visual words from coarse scale patch features and put them in the upper hierarchy of the codebook, at the same time, create fine scale visual words from fine scale patch features and put them in the lower hierarchy. Furthermore, fine scale visual words and coarse scale visual words created from related coarse scale patch features and fine scale patch features are associated to each other. In the image representation creation stage, patches extracted from the same sample points are encoded with their corresponding relationship explored based on the hierarchical codebook. In this way, relationship among patches is able to be incorporated into image representations.

3.1 Codebook Construction Based on Related Patches

In the construction of the hierarchical codebook in which contextual information of visual words is incorporated, image patches which are related to each other are extracted first. At each sample point, we extract two patches with different scales, i.e. fine scale and coarse scale patches, and take them as a pair. All extracted patches are described by the SIFT descriptor [10]. Then, we combine the coarse scale patch feature with the fine scale patch feature as a single feature in the same way as the work by Qin [17]. Consequently, the combined feature contains information for both the fine scale patch and its context. After that, vector quantization is applied to these combined features, so that context-aware feature clusters are able to be created.

After we clustered these sampled points into groups on the basis of the combined features by using k-means, the fine scale patch features and the coarse scale patch features are also grouped based on their corresponding relationship with the combined features. As a consequence, each group of combined features corresponds to a group of fine scale patch features and a group of coarse scale patch features. In order to reduce outliers in each group, K-nearest neighbour is adopted to detect patch features that are singular in each patch feature group. At last, we apply K-means to each fine scale patch feature group to get a set of fine scale visual words and each coarse scale patch feature group to get a set of coarse scale visual words. In this process, fine scale visual words and coarse scale visual words corresponding to the same combined feature cluster are associated to each other.

In this way, relationship existing in these visual words can be recorded and utilized in the later stage. Finally, all obtained visual words are combined to form a hierarchical codebook in which visual words created from coarse scale patches are put in the upper hierarchy and visual words created from fine scale patches are put in the lower hierarchy. Figure 1 shows the process of the hierarchical codebook construction. The structure of the obtained hierarchical codebook is given in Fig. 2.

3.2 Image Representation Construction Based on the Hierarchical Codebook

To create image representations on the basis of the hierarchical codebook, we take the coarse scale patch and the

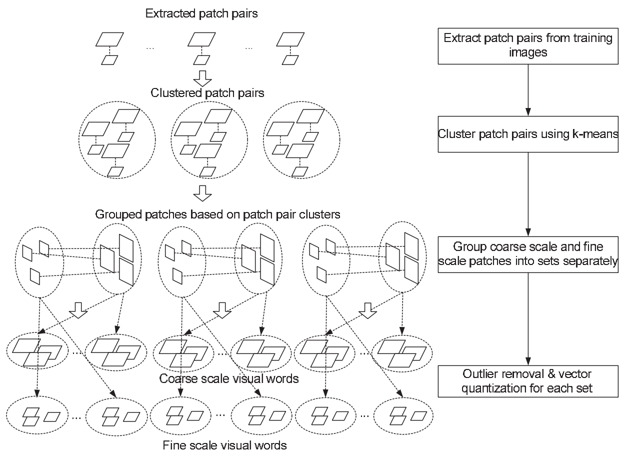

Extracted patch pairs
Clustered patch pairs
Grouped patches based on patch pair clusters
Coarse scale visual words
Fine scale visual words
Extract patch pairs from training images
Cluster patch pairs using k-means
Group coarse scale and fine scale patches into sets separately
Outlier removal & vector quantization for each set

Fig. 1 Hierarchical codebook construction based on related patches. Coarse scale patches and fines scale patches are combined first. Then combined patch pairs are clustered. Corresponding to each cluster, coarse scale patches and fine scale patches are grouped into sets. Finally, visual words are created from each patch group. And fine scale visual words and coarse scale visual words corresponding to the same patch pair cluster are associated with each other.

fine scale patch extracted from the same sample point as a patch pair. Then, we propose a method to assign each pair of patches to the codebook with patch-level contextual information considered. At the same time, in the patch pair assignment, relationship among visual words which are associated in the construction stage is also exploited.

In order to assign a pair of patches to the codebook, first we find the N nearest visual words for each patch in the pair, respectively. The N nearest visual words of a patch in the pair are searched from the hierarchy of corresponding scale in the codebook. After that, by using the corresponding relationship among visual words of different hierarchies, for each found nearest visual word, visual words associated to it are also available. For instance, for a coarse scale visual word, a number of fine scale visual words are associated with it, and for a fine scale visual word, a number of coarse scale visual words are associated with it. The weights assigned to the nearest visual words of a patch in the pair are calculated based on both itself and its paired patch. Specifically, for the calculation of the assignment weights of a coarse scale patch, we use the following equations:

$$vw_{ci} = \alpha \cdot 1/d_{ci} + \beta \cdot 1/d_{fe_i} \tag{1}$$

$$\beta = \begin{cases} 1 & \text{If word } e_i \text{ is one of the N nearest words} \\ & \text{of the fine scale patch in the pair.,} \\ 0 & \text{Otherwise.} \end{cases} \tag{2}$$

where vw_{ci} is the weight factor assigned to the i_{th} nearest visual word ci of the coarse scale patch out of its N nearest visual words. d_{ci} is the Euclidean distance between the coarse scale patch feature and visual word ci. e_i is the nearest visual word of the fine scale patch in the patch pair out of the set of fine scale visual words associated with visual word ci. d_{fe_i} is the Euclidean distance between the fine scale patch feature in the pair and visual word e_i. The value of β is determined as follows. For the i_{th} nearest visual word of the coarse scale patch feature, from its associated fine scale visual words, the one e_i which is the nearest to the fine scale patch feature in the patch pair is found. Then, e_i is compared with the N nearest visual words of the fine scale patch in the codebook. If e_i is one of N nearest visual words of the fine scale patch, β is set to be one, otherwise zero is set. In this way, if two patches in the pair are assigned to associated visual words, a relatively large weight is assigned to the corresponding visual words, otherwise a small weight is assigned. And in the former case, weights assigned to visual words are determined by both patches in the pair. α is a constant value, which is determined through experiments.

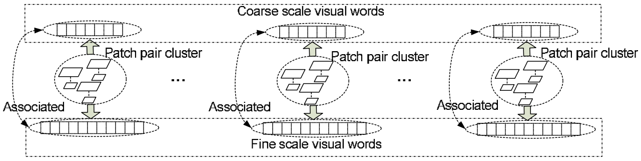


Fig. 2 Hierarchical codebook structure. In the hierarchical codebook visual words created from coarse scale patches are put in the upper hierarchy, while visual words created from fine scale patches are put in the lower hierarchy. At the same time, fine scale visual words are associated to coarse scale visual words which are created corresponding to the same patch pair cluster.

In Fig. 3, the process of creating an image representation based on the proposed method is shown. Figure 3 indicates patch pair assignment to visual words with corresponding relationship using solid arrows and indicates assignment to visual words without corresponding relationship using dashed arrows. By using the above procedure, the relation between the coarse scale patch features and fine scale patch features is incorporated into the bag of visual words framework. The assignment weights of a patch feature is calculated based on both itself and its related patch feature. When all the N nearest visual words have been assigned weights to, these weights are normalized to the range $(0, 1)$ as follows

$$w_{ci} = \frac{vw_{ci}}{\sum_{k=1}^{N} vw_{ck}}. \tag{3}$$

Similarly, for the assignment of the fine scale patch in the patch pair, the same procedure is used. After all extracted image patches are assigned to the hierarchical codebook, the obtained histograms corresponding to the upper

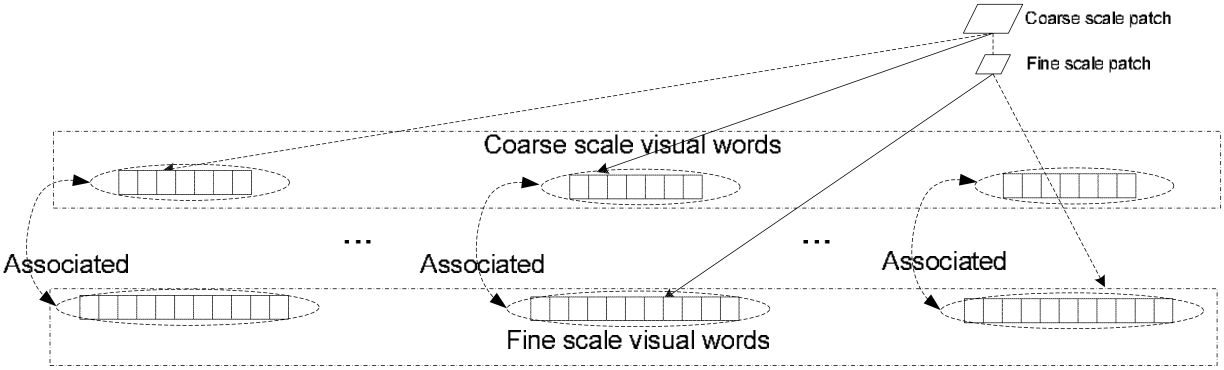


Fig. 3 Image representation creation on the basis of the hierarchical codebook using extracted patch pairs. Each patch in the patch pair is assigned to two nearest visual words. Solid arrows indicate that the assigned nearest words of the two patches are associated, while the dashed arrows indicate the assigned nearest visual words are not associated.

hierarchy and the lower hierarchy of the codebook are concatenated as a single vector, which is taken as the final image representation and used for classification.

4. Image Classification through a Probabilistic Approach

After we proposed the hierarchical codebook containing contextual information of visual words and the approach to represent images based on the proposed codebook, we try to create discriminative image representations, so that final categorization performance can be improved. Since different sampling strategies may be biased towards different aspects of image contents. To utilize image information effectively, we apply our method to both regularly sampled features and features extracted based on an interest point detector. After that, we propose a mechanism to fuse these two image representations created from different sets of features in the image classification stage.

In our research, we adopt SVM for image classification. Usually, classification results predicted by SVM do not have confidence value. To overcome this shortcoming, we adopt a technique for multi-class probability estimates [22], so that the probability for a test feature to belong to each class can be provided. This probability is taken as the confidence for the feature to belong to the corresponding class. As stated above, from each image we extracted two kinds of features through the regularly sampling strategy and an interest point detector, respectively. Thereafter, we create an image representation from each kind of features. As a consequence, for each image we obtained two kinds of image representations, which are used to train an individual multi-class SVM classifier, separately. In the image categorization stage, each test image is also represented as two different representations, as in the training stage. Then, each kind of image representations is classified by its corresponding SVM classifier. Through the multi-class probability estimates technique, for each image representation, its corresponding multi-class SVM classifier can output a vector of probabilities, whose elements represent the probabilities for this representation to belong to the corresponding classes. Since we represented each image based on two kinds of representations, we can obtain two vectors of probabilities. Finally, we sum the two obtained probability vectors to obtain the final confidence value vector for the image, and take the class corresponding to the largest confidence value in the final vector as the class for the test image to belong to.

5. Experiments and Results

In this section, we present experiments and results for evaluating the performance of the proposed method and comparing it to other baseline methods. In experiments, we used dataset Caltech 101 [3] and dataset Caltech 256 [20]. From each dataset, 10 categories of objects are selected randomly. For each object category of dataset Caltech 101, around 60 images are employed, while for each category of dataset Caltech 256, around 100 images are used. Furthermore, five-fold cross validation is adopted, and obtained average value is taken as the final result. SVM with RBF kernel function is used as the classifier. We used LIBSVM [2] to perform classification in our work.

We have proposed a hierarchical codebook, in which visual words created from related patches are associated with each other. Based on the hierarchical codebook, we take patches from the same sample point as a patch pair and assign them to the codebook with their corresponding relationship explored. By using the proposed method, in the obtained image representations, values corresponding to coarse scale visual words are closely related to values corresponding to fine scale visual words. To show the effect of the method, we give Fig. 4 which shows the distribution of accumulated image patch assignment values corresponding to visual words of the hierarchical codebook. To create this figure, 200 coarse scale visual words and 200 fine scale visual words are used. For the conventional method, coarse scale visual words and fine scale visual words are independent.

The correlation among patches from the same image is intrinsic. In the traditional method, each image patch is encoded independently without considering relation among patches. In our work, visual words are related to each other based on the relation of patches from which they are created. And in the stage of image patch encoding, patches extracted from the same sample points are also related to each other. Then related patches are encoded together, so that when related patches are assigned to related visual words, high weight values are given and vice versa. Through this manner, relationship among image patches can be incorporated into image representations, where the ambiguity for individual image patch encoding can be relieved. The results in Fig. 4 demonstrate that the proposed method is effective for incorporating relationship among different patches into

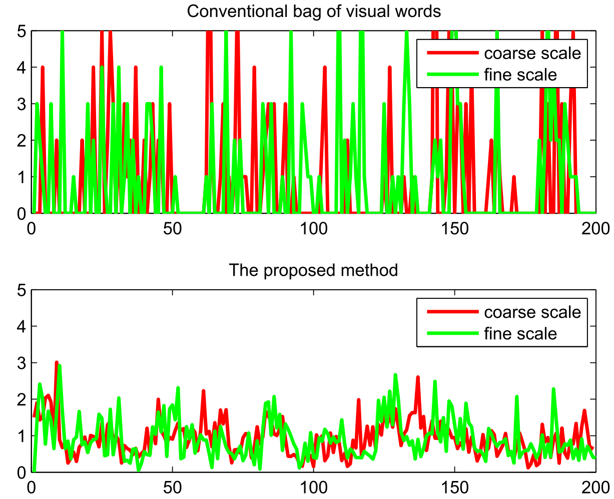


Fig. 4 Distribution of values corresponding to coarse scale visual words and fine scale visual words. In the figure, x-axis is visual words ordered in sequence, y-axis is accumulated patch assignment value for visual words.

Table 1 Categorization performance of the proposed method under different numbers of patches on datasets Caltech 101 and Caltech 256.

	500	1000	2000	3000
Caltech 101	0.7208	0.7416	0.7500	0.7625
Caltech 256	0.7237	0.7225	0.7437	0.7387

image representations.

In the following experiments, we first investigate influence of parameters of the proposed method on the classification performance. In our work, we implemented the proposed method on patch features extracted based on regularly sampling and an interest point detector, respectively. For the regularly sampling strategy, the number of patches used for creating image representations is evaluated first, where patches are first extracted with a sampling step of 4 pixels. Then, a specified number of patches are selected randomly for creating image representations. In this experiment, coarse scale patches are of the size of 60 × 60, while fine scale patches are of the size of 20 × 20. These patch sizes are determined experimentally. The number of nearest visual words for a patch to assign to is set to be ten. The number of coarse scale visual words is 1000 and the number of fine scale visual words is 2000. These visual words are created from 50 combined feature clusters. The obtained result is given in Table 1.

We evaluated the performance of the proposed method under different numbers of patch features, when regularly sampled patches are used for creating image representations. From the results given, it can be seen that larger numbers of patch features give better performance. The reason is that when a larger number of image patch features are used, the obtained image representations contain more information about the original images, which makes them more discriminative. This result is in agreement with observations in previous works [18]. However, when more image patches are extracted from an image, the computation burden will also increase. According to the results, the proposed method is robust against the number of patch features. Therefore, only a limited number of image patch features are used in our work.

For the proposed method, the number of nearest visual words to which a patch is assigned to is also an important parameter. In order to test how this parameter influences classification performance, the subsequent experiment is conducted. In this experiment, we implement the proposed method on both regularly sampled image features and features extracted based on an interest point detector. For regularly sampled image patches, the patch sampling step and patch sizes are set to be the same as in the previous experiment. On the other hand, for patches sampled based on the interest point detector, after we detected interest points using Harris-Laplace detector, two patches of different scales are extracted from each detected point. These patches are of the sizes of *interest scale* and 1.5 × *interest scale*. The number of visual words in the codebook is set to be the same as in the previous experiment for both regularly sampled features and features extracted from interest points. Experiment results are given in Fig. 5.

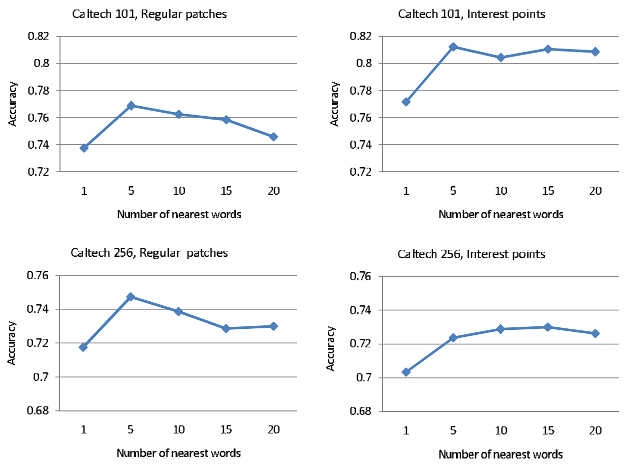


Fig. 5 Experimental results on the number of visual words a patch is assigned to. Datasets Caltech 101 and Caltech 256 are used.

From Fig. 5, it can be observed that the performance of the multiple visual words assignment is better than the one visual word assignment. However, increasing the number of nearest visual words to which a patch is assigned does not give further improvement. Generally, small numbers give better performance than large numbers for both dataset Caltech 101 and dataset Caltech 256. This is because when a patch is assigned to a large number of visual words, noise will be introduced, which makes the classification performance deteriorate. On the other hand, if the number is too small, the method gets less robust.

Moreover, another important parameter is the size of the codebook used for creating image representations. In our work, the number of fine scale visual words is set to be two times of the number of coarse scale visual words. This ratio is fixed in all the experiments, even if the size of the used codebook is changed. We evaluated the performance of the proposed method using codebooks of different sizes. At the same time, in order to evaluate the contribution of the hierarchical codebook and the context-aware weighting, comparison among the method of multiple assignments on traditional codebook (Baseline), the method of multiple assignments on the hierarchical codebook without context-aware weighting (Unweighted) and the method of multiple assignments on the hierarchical codebook with context-aware weighting (Full-method) is made. For the traditional codebook, the number of visual words is set to be the same as the hierarchical codebook. Its visual words are obtained by applying k-means to image patch features directly. Experiment results are given in Fig. 6. In this figure, the number on the x-axis is the number of fine scale visual words of the codebook, which is two times the number of coarse scale visual words.

Figure 6 demonstrates that image classification performance improves as the size of the used codebook increases for the method on the traditional codebook and the methods on the hierarchical codebook. However, further im-

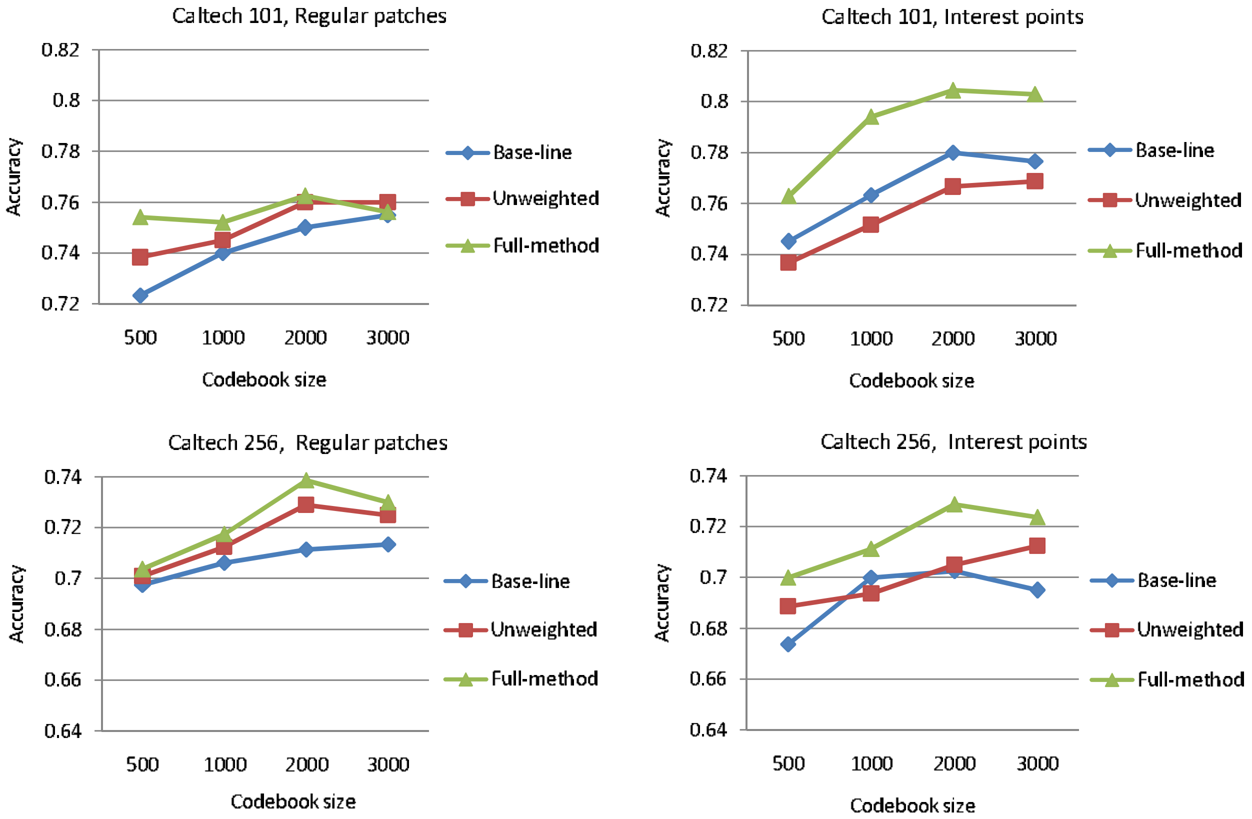


Fig. 6 Experimental results on the size of the codebook and comparison with the baseline methods, datasets Caltech 101 and Caltech 256 are used.

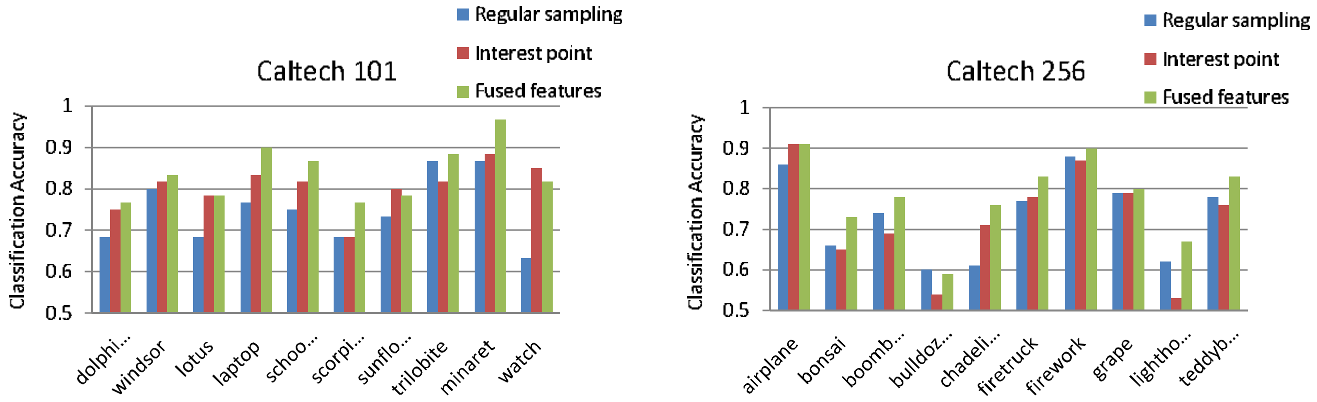


Fig. 7 Comparison of fused features to each individual kind of features, on datasets Caltech 101 and Caltech 256.

provement is insignificant after the used codebook size is large enough. This result indicates that simply increasing the codebook size is not an effective way for classification performance gain. In addition, it is observed that only using the hierarchical codebook does not give obvious improvement. When context-aware weighting based on the hierarchical codebook is adopted, the performance can be improved. The reason is that in the hierarchical codebook visual words are related to each other for encoding paired patch features. If the context-aware weighting is not utilized, relationship among patches can not be incorporated into image representations. Therefore, in order to utilize the relationship among patches for relieving the patch encoding ambiguity, the hierarchical codebook and the context-aware weighting should be utilized together.

For the proposed method, regularly sampled features and features sampled from interest points lead to different performance. It is worthwhile to investigate the performance of both kinds of features on specific categories. Figure 7 gives the performance comparison of the two kinds of features which are extracted based on different sampling strategies. In this result, 1000 coarse scale visual words and

Table 2 Comparison between the proposed method and baseline methods on dataset Caltech 101.

Proposed method on regular SIFT	Proposed method on combined features	Contextual words [17]	Spatial pyramid [1]
0.7625	0.8366	0.8207	0.8524

Table 3 Comparison between the proposed method and baseline methods on dataset Caltech 256.

Proposed method on regular SIFT	Proposed method on combined features	Contextual words [17]	Spatial pyramid [1]
0.7387	0.7800	0.6975	0.7560

2000 fine scale visual words are used. Experiment results on specific categories show that some categories achieved better performance on regularly sampled features, while other categories achieved better performance on features extracted through the interest point detector. The difference in classification performance is because different features are more suitable for representing different aspects of the image categories. Therefore, to utilize extracted image information effectively for object categorization, it is intuitive to fuse these two kinds of features. We propose to fuse these two kinds of features in the image classification stage through a probabilistic mechanism. Performance of the fused features is tested. Experiment results are given in Fig. 7.

From results shown in Fig. 7, it can be observed that for most of the categories fused features give better performance than any individual kind of features. These two kinds of features are able to be used for complementing each other, and the proposed feature fusion method is effective.

Finally, to evaluate the efficiency of the proposed method, we compare it to other methods which have utilized contextual information for creating discriminative representations. The two baseline methods used for comparison are contextual words approach [17] and spatial pyramid approach [1]. In these methods, regularly sampled SIFT features are used for creating image representations. Parameters of the baseline methods are set following previous works [1], [17]. Table 2 and Table 3 give the performance of the proposed method on regular SIFT features and fused features as well as the performance of the baseline methods. From the results, we can see that on the easy dataset Caltech 101 the performance of the proposed method on regular SIFT features and fused features is 76.25% and 83.66% respectively. On the complex dataset Caltech 256, the performance of the proposed method on regular SIFT features and fused features is 73.87% and 78.00%. Compared to the baseline methods, the results indicate that the proposed method can give robust results on both easy and complex datasets. Furthermore, the proposed method gives a new approach to incorporate relationship among image patches into image representations. It is not contradictory to the spatial pyramid baseline method and can be fused to improve the image categorization performance further, which is our future work.

6. Conclusion

In this paper, we have proposed to construct a hierarchical codebook in which visual words created from coarse scale patches are put in the upper hierarchy, and visual words created from fine scale patches are put in the lower hierarchy. The hierarchical codebook is constructed based on related patches which are extracted from the same sample point. By utilizing corresponding relationship among these patches, visual words in the codebook are also associated. In the image representation stage, patches extracted from the same sample point are taken as a pair and encoded by exploring the relationship among associated visual words. Finally, we implement our methods on both regularly sampled features and interest point features and fuse obtained image representations from these features through a probabilistic approach. We obtained categorization accuracies of 83.66% on dataset Caltech 101 and 78.00% on dataset Caltech 256, respectively, which are comparable to those of the baseline methods. The experiment results show the effectiveness of the proposed method. The future work is to incorporate spatial information into the proposed method to improve the classification performance further.

References

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

[2] C. Chang and C. 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 Trans. Image Process., vol.10, no.1, pp.117–130, 2001.

[6] F. Perronnin, “Universal and adapted vocabularies for generic visual categorization,” IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.7, pp.1243–1256, 2008.

[7] R. Fergus, P. Perona, and 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 and 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," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[11] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," Int. J. Comput. Vis., vol.60, pp.63–86, 2004.

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

[13] M.M. Mohammadi, S. Escalera, and 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 classification," Proc. 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," J. Vis. Commun. Image Represent., vol.20, no.7, pp.450–462, 2009.

[16] 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.

[17] J.Z. Qin and N.H.C. Yung, "Scene categorization via contextual visual words," Pattern Recogniti., vol.43, no.5, pp.1874–1888.

[18] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," ECCV, vol.4, pp.490–503, 2006.

[19] 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," Proc. ICCV, pp.1992–1999, 2009.

[20] G. Griffin, A.D. Holub, and P. Perona, The Caltech-256, Caltech Technical Report, 2006.

[21] F. Jurie, B. Triggs, and P. Perona, "Creating efficient codebooks for visual recognition," Proc. Tenth IEEE International Conference on Computer Vision, pp.604–610, 2005.

[22] T.F. Wu, C.J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pair wise coupling," J. Machine Learning Research, pp.975–1005, 2004.

[23] J.C. van Gemert, A.W.M. Smeulders, and J.M. Geusebroek, "Visual word ambiguity," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.1, pp.1271–1283, 2010.

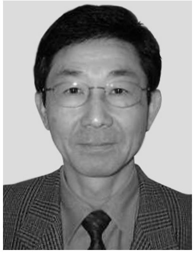
[24] J. Philbin, O. Chum, M. Isard, and J. Sivic, "Lost in quantization: Improving particular object retrieval in large scale image databases," Proc. CVPR, 2008.

Shuang Bai received the degrees of B. Eng. and M. Eng. from the School of Electrical Engineering and Automation of Tianjin University, Tianjin, China in 2007 and 2009, respectively. In 2010, he became a Dr. Eng. candidate in the Graduate School of Information Science of Nagoya University. His research interests include computer vision and pattern recognition.

Tetsuya Matsumoto received the degrees of B. Eng., M. Eng. and Dr. Eng. from Nagoya University, Nagoya, Japan, in 1982, 1984 and 1996, respectively. From 1984 to 1989, he worked in Toshiba Corporation, Fuchu, Japan, where he engaged in research and development of the control system of nuclear power plant. In 1993, he joined Education Center for Information Processing, Nagoya University. In 1998, He moved to Department of Information Engineering, Graduate School of Engineering, Nagoya University. His current interests include neural networks, image processing and machine learning. He is a member of JNNS of Japan and JSAI of Japan.

Yoshinori Takeuchi received the degrees of B. Eng., M. Eng. and Dr. Eng. from Nagoya University in 1994, 1996 and 1999, respectively. In 1999, he was a Research Fellow of the Japan Society for the Promotion of Science. In 2000, he was a member of the Graduate School of Engineering, Nagoya University. Currently, he is an Associate Professor at the Information Security Promotion Agency, Nagoya University. His research interests include computer vision and computer audition. He is a member of IEEE, and RSJ.

Hiroaki Kudo received the degrees of B. Eng., M. Eng. and Dr. Eng. at Nagoya University, Japan in1991, 1993 and1996, respectively. In April 1996, he was a faculty member of the School of Engineering, Nagoya University as a Research Associate. In April 1997, he was a faculty member of the Graduate School of Engineering, Nagoya University. In April 1999, he was an Assistant Professor. In August 2000, he was an Associate Professor at Center for Information Media Studies, Nagoya University. Since 2003, he has been a member of Graduate School of Information Science, Nagoya University. He was a Research Fellow of the Japan Society for the Promotion of Science in1995. His research interests include visual perception and computer vision. He is a member of IEEE, ITE, and IEEJ.



Noboru Ohnishi received the B. Eng., M. Eng. and D. Eng. degrees from Nagoya University, Nagoya, Japan, in 1973, 1975 and 1984, respectively. From 1975 to 1986 he was with the Rehabilitation Engineering Center under the Ministry of Labor. From 1986 to 1989 he was an Assistant Professor in the Department of Electrical Engineering, Nagoya University. From 1989 to 1994, he was an Associate Professor. Since 1994, he is a professor in Nagoya University. From 1993 to 2001, he concurrently holded a Head of Laboratory for Bio-mimetic Sensory System at the Bio-mimetic Control Research Center of RIKEN. He is now in the Graduate School of Information Science. His research interests include computer-vision and -audition, robotics, bio-cybernetics, and rehabilitation engineering. Dr. Ohnishi is a member of IEEE, IEEJ, IPSJ, SICE, JNNS, ITE and RSJ.