Object Recognition Using Hierarchical Structure

Takumi Toyama*, Koichi Kise*

* *Graduate School of Engineering, Osaka Prefecture University, Osaka*
E-mail: takumi@m.cs.osakafu-u.ac.jp, kise@cs.osakafu-u.ac.jp

Abstract

Object recognition is one of the most challenging problems in the field of computer vision. Although recent approaches have shown promising results, such approaches are specialized in each recognition task. Therefore they cannot be extended many other recognition tasks. To integrate many types of recognition, we propose a novel recognition method which uses hierarchical structure. Our experimental results show the proposed method has advantages on processing time and accuracy compared to a conventional method for generic object recognition.

1. Introduction

The realization of object recognition is one of strong desires of vision researchers. To date, several approaches to object recognition tasks are proposed. In such approaches, it has been shown local features extracted from images allow us robust recognition.

For the task of specific object recognition (recognition of object instances), it is reported that approaches which use "raw" local features are effective [1]. On the other hand, local features are clustered to acquire high accuracy for the task of generic object recognition (recognition of object classes) [2]. It seems these approaches are specialized in each task. Approaches for generic object recognition are not suited for specific object recognition tasks and vice versa. In actual uses, it is more preferable to integrate both recognition tasks into a same framework. For this reason, we need to redefine the specific object recognition tasks as well as the generic object recognition tasks in the same framework.

Specific object recognition handles object instances. These instances are regarded as members of their classes. Moreover, these classes are also members of their super-classes. Thus, as shown in Fig. 1, each class and instance compose hierarchical structure from generic to specific levels.

By using hierarchical structure, we can redefine both recognition tasks in the same framework. At the bottom level, the recognition task is the same as the specific object recognition. It is also the same that we can execute a generic object recognition task at the remaining levels.

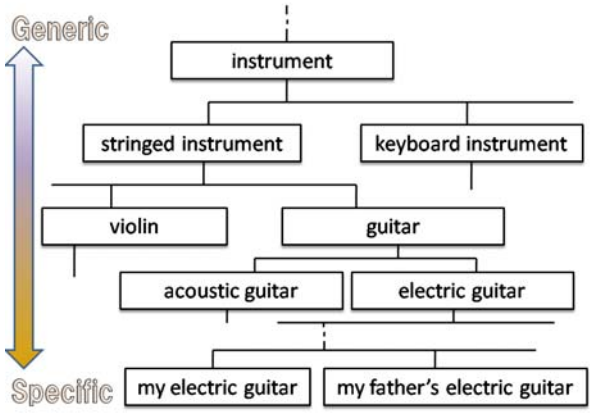


Figure 1: At the bottom level, there are object instances whereas it becomes more generic at the top level in hierarchical structure.

Most current object recognition systems require computational costs for multiclass object recognition because they compare all classes at the same time to output the results. In hierarchical structure, we winnow the candidate classes at each level. Thus, the total number of comparison is smaller than that of the previous systems.

In addition, we emphasize another advantage of recognition system using hierarchical structure. For classification, it is common to represent each image as a histogram (vector). To compare each image, the representation of each histogram must be shared by all classes. This means we have to use the same representation all the time in any recognition tasks. However, the part where we focus on to distinguish between a car and a guitar is quite different from the part which is significant for specific object recognition. In hierarchical structure, we can change the representation of histogram at each level so that we can focus on the significant part for each recognition task.

In this paper, we give the experimental results of comparison our recognition method using hierarchical structure to a representative method for object recognition. The results show the proposed method outperforms to the conventional method.

ISSN 1975-4736

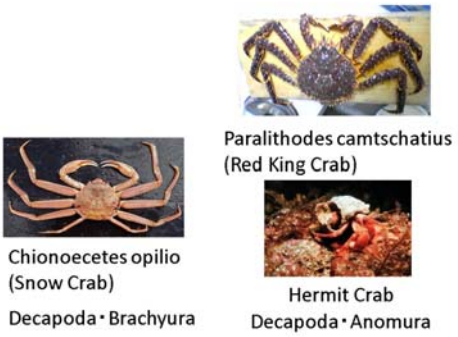


Figure 2: Although the red king crab and the snow crab have similar appearance, they are academically classified into different classes.

2. Recognition Using Hierarchical Structure

We introduce a novel method which is an extension to the representative method for generic object recognition. This method consists of two phases. One is the training phase and the other is the testing phase.

2.1 Training

2.1.1 Construction of Hierarchical Structure. First, we need to consider the construction of hierarchical structure. Although hierarchical structure is already defined in some categories such as biological classification, they are not always suitable for visual recognition as shown in Fig. 2. For this reason, to construct hierarchical structure suitable for the visual recognition is necessary to achieve high accuracy. In this paper, we construct hierarchical structure which is based on their appearance by hand.

2.1.2 Local Feature Extraction. From each image, we extract local features. In recent research, it is reported that local features are to be effective for object recognition especially in the tasks of recognition whose images contain viewpoint changes, clutter backgrounds and occlusions. Local features are extracted from images by a detector and a descriptor. A detector extracts regions which have special properties in images. Then local features are output as vectors by a descriptor for each detected region.

We use the combination of the Harris-Affine detector and the SIFT descriptor to obtain the robustness for affine and scale transformation. The Hessian-Affine detector [3] extracts corner-like regions which are robust to affine transformation. The SIFT (Scale-invariant Features Transform) descriptor [4] computes a gradient orientation

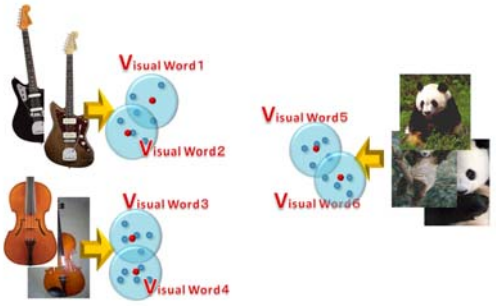


Figure 3: Centroids of clusters are visual words.

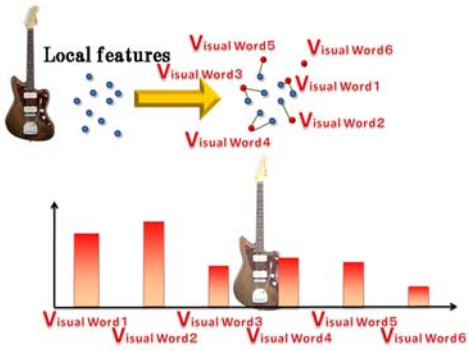


Figure 4: Represent the image as a histogram of local features.

histogram within the detected region. The SIFT descriptor is invariant to scale transformation.

2.1.3 Visual Words. To compute the similarity of each image, we need to describe all images with the same representation. One method for doing this is to define visual words (visual vocabulary). We represent an image as a histogram of frequency of each visual word. It is similar to document representation as a frequency of each word for document retrieval.

We obtain visual words by clustering a set of local features found in each class as shown in Fig. 3. Each centroid is regarded as a visual word. In the n class case, we extract m visual words from each class so that the number of dimension of histogram is $n \times m$.

After obtaining visual words, we represent each image with visual words as shown in Fig. 4. The i th entry of the histogram is the number of all local features in the image which are closest to the i th visual word.

In hierarchical structure, we can set different visual words at each level so that the representation is suitable for each recognition task.

2.1.4 Support Vector Machine. In previous research, several approaches have proposed for classification. The

ISSN 1975-4736 ©MITA 2009

SVM(Support Vector Machine) is well adapted because of its feature that it can divide two classes data effectively.

We obtain models with the SVM. Since the SVM is a binary classifier, models for all possible pairs should be trained for multiclass classification, which is called the one-against-one technique. We use this technique at each level in hierarchical structure.

2.2 Testing

2.2.1 Classification with the SVM. In testing phase, all test images are represented as histograms similarly in the training phase. For each test pattern $\mathbf{x}$, we obtain its label (-1 or +1) by computing the SVM decision function. The majority of answer of each model for all classes is adapted as a final answer.

A kernel function is incorporated into the SVM framework. We use the RBF (Radial Basis Function) in our method.

2.2.2 Classification Using Hierarchical Structure. Our recognition system executes recognition tasks from top to bottom in hierarchical structure. At the first (top) level, we execute generic object recognition such as a guitar or a violin. Following the first level, we execute the second level recognition with an answer at the top level. Although this means misclassification at the first level is never recovered, we found this method is more efficient than the previous methods which handle all bottom classes at the same time in terms of accuracy and execution time.

In previous approaches, the number of computation for each test pattern is enormously large when the number of classes is also large. By using hierarchical structure, the total number of computation becomes smaller because the number of possible pairs at each level is not large as shown in Fig. 5.

3. Experiments

3.1 Experimental Conditions

For our evaluation, we construct the stringed instrument hierarchical structure as shown in Fig. 6 (a). The reason we picked up this category was because objects in this category have distinguishable visual parts which is easily found by anyone if he or she knows the category.

The image dataset consisting of 1885 images were prepared by downloading images of stringed instruments from Google and Flickr. We used SVM^{light} [5] with the RBF for classification. All components of this system were run on a computer with an AMD Opteron 2.8GHz CPU and 64GB RAM.

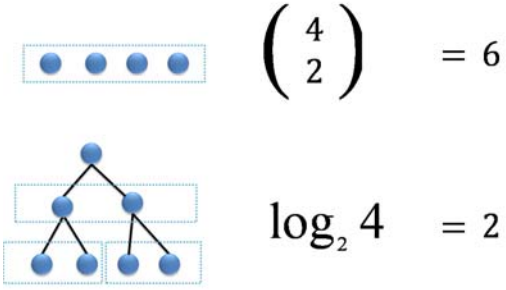


Figure 5: In 4 class case, the number of computation is 6 without hierarchical structure. By using hierarchical structure, we can reduce 4 times computation.

Table 1: Evaluation of hierarchical structures

	Hierarchical structure shown in Fig. 6 (a)	Hierarchical structure shown in Fig. 6 (b)
Accuracy for leaf classes	45.0%	36.5%

Table 2: Accuracy using different visual words settings at each level

	The different visual words settings	The visual words setting shared by all levels
Level 1	80.1%	74.4%
Level 2	73.3%	61.4%
Level 3-A	52.9%	42.3%
Level 3-B	46.8%	25.9%
Level 4-A	32.4%	18.6%
Level 4-B	31.9%	27.9%

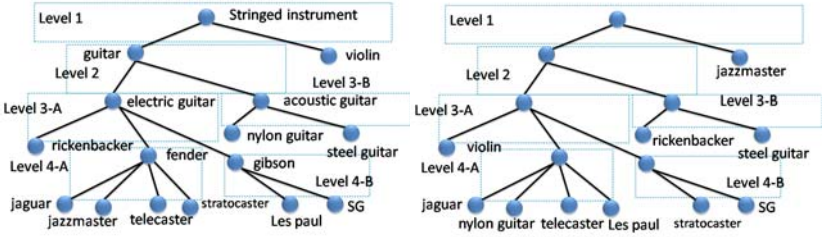
Table 3: Comparison the proposed method to the conventional method

	The proposed method	The conventional method
Accuracy	45.0%	34.4%
Processing time	0.45s	1.67s

3.2 Evaluation of Hierarchical Structures

First, we show the results of evaluating the performance of different hierarchical structures. We investigated the significance to construct a proper hierarchical structure in this evaluation.

ISSN 1975-4736 ©MITA 2009



(a) (b)

Figure 6: Hierarchical structure

We compared two hierarchies. One was the structure shown in Fig. 6 (a) which is constructed by hand based on the appearance of these classes. The other is shown in Fig. 6 (b) which has the same tree but with random assignment of categories to nodes.

Table 1 shows accuracy of recognition for leaf classes. We can see the hierarchical structure (a) performs better than (b). The main reason that (b) was inferior to (a) is that it tended to fail classification at the 3rd level, especially classification of violin.

From this result, we have confirmed the significance of construction of a proper hierarchical structure.

3.3 Evaluation of Visual Words Settings

Next, we evaluated the performance of visual words settings. We compared the different visual words settings at each level to the visual words setting shared by all levels. We used hierarchical structure (a).

Table 2 shows the results at each level. For comparison, these results are the best accuracy obtained in the experiments. We can see the different visual words settings acquired higher accuracy at all level.

Consequently, it has proven to be effective to change the visual words settings at each recognition levels.

3.4 Comparison to a conventional method

Finally, we compared the proposed method to a conventional method. In the conventional method a single visual words setting is employed. Objects are recognized based on the one-against-one technique. To compare these two methods, we evaluated the performance for leaf classes.

Table 3 shows the results. The proposed method outperformed to the conventional method on both accuracy and processing time. In terms of accuracy, the conventional method had difficulty to obtain enough votes for the correct class because the number of models is so large compared to the proposed method.

As we stated in Section 2, the number of application of SVMs in the proposed method is smaller than that in the conventional method. Therefore, we can reduce processing time for recognition.

3. Conclusion

In this paper, we have introduced a novel object recognition system using hierarchical structure. By using hierarchical structure, we can vary the levels of specificity so that the recognition system can be extended from generic to specific tasks.

Experimental results show that this method can reduce the computational cost and acquire higher accuracy than the conventional method. In addition, we confirmed the significance of construction of hierarchical structure. Future work is to construct hierarchical structure automatically and to acquire higher accuracy.

Acknowledgements This work was supported in part by the Grant-in-Aid for Scientific Research (B) (19300062) from Japan Society for the Promotion of Science (JSPS).

5. References

[1] K. Kise, K. Noguchi and M. Iwamura: “Memory efficient recognition of specific objects with local features”, Proc. of the 19th International Conference of Pattern Recognition (ICPR2008), pp.1-4 (2008).

[2] J. Zhang, M. Marszalek, S. Lazebnik and C. Schmid: “Local features and kernels for classification of texture and object categories: A comprehensive study”, International Journal of Computer Vision, Vol. 73, No. 2, pp. 213-238 (2006).

[3] D. Lowe: “Distinctive image features from scale invariant keypoints”, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110 (2004).

[4] K. Mikolajcyk and C. Schmid: “Scale and Affine invariant interest point detector”, International Journal of Computer Vision, Vol. 60, No. 1, pp. 63-86 (2004).

[5] T. Joachims: “SVM*light* Support Vector Machine”, http://www.cs.cornell.edu/People/tj/svm%5Flight/.

ISSN 1975-4736 ©MITA 2009