Unknown Object Identification Using Category Visual Words with Rejection Function

Yuto Tanaka, Tetsuya Takiguchi, Yasuo Ariki
Department of System Informatics, Kobe University
1-1 Rokkodai, Kobe, Japan
ytanaka@me.cs.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

In this paper, we introduce an identification method for unknown category objects. Most popular conventional methods in object recognition use Bag of Features (BoF) that represents the image as an appearance frequency histogram of common visual words by quantizing SIFT features. However, this method is unable to identify unknown objects because the common visual words cannot represent the unknown objects well. From this viewpoint, we introduce an unknown object identification method that creates individual category visual words with a rejection function, which can absorb the features of other objects or the background. As a result of object recognition of 10 classes, the proposed method has improved the recognition rate by 8.0 points, compared with the conventional BoF method.

1 Introduction

Generic object recognition involves recognizing objects by their general name using a computer in a real-world setting. This is one of the most challenging tasks in computer vision. Moreover, due to the popularization of digital cameras and the development of high-capacity hard disk drives in the recent years, it is getting difficult to classify and to retrieve enormous videos and images manually. Therefore, computers are required to automatically classify and retrieve such videos and images. For this reason, generic object recognition is becoming more and more important.

The most popular conventional method of generic object recognition is Bag of Features (BoF) [1]. BoF is the appearance-based method that extracts the local features, such as SIFT (Scale-Invariant Feature Transform) [2][3], from the object images, and classifies them into W clusters using a k-means algorithm. The centroid vector of each cluster is called "visual words", and the number of words, W, is determined empirically. In this way, the object image is represented by the histogram of the common visual words. The representation of the object image using BoF is robust to occlusion because it is expressed as the collection of local features, and it is also robust to the change of appearance because of vector quantization by the k-means algorithm.

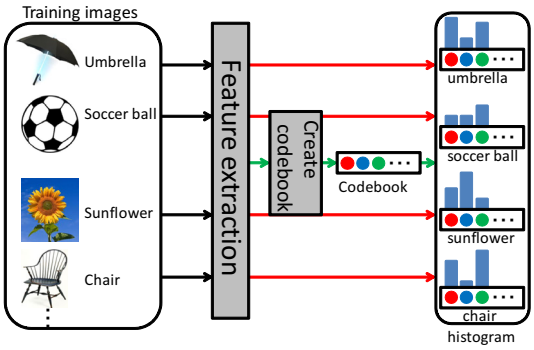

Figure 1. A flow of conventional BoF

Fig. 1 shows a flow of the conventional BoF method. In Fig. 1, training data for object images consist of various categories, and SIFT features are extracted from them. A codebook is created from the features, and the features of the object images are quantized into common visual words in the codebook. However, the BoF method cannot recognize unknown category objects well.

Fig. 2 shows both situations where object A is known or unknown. In this figure, a user says, "Where is object A?". If the system knows object A, it says "This is A" by picking up the object. However, if the system does not know object A, it has to collect the training images of object A (for example, from the Internet) in order to build the model for object A, because object A is unknown to the system.

There are huge images on the Web, and they are tagged with category names. By searching for the training data for images from the keyword "A" and performing supervised learning using the obtained images, the system can learn the model of the unknown object A and pick it up.

As described above, the BoF method creates the common visual words using various training images to represent the underlying known objects. Therefore, in

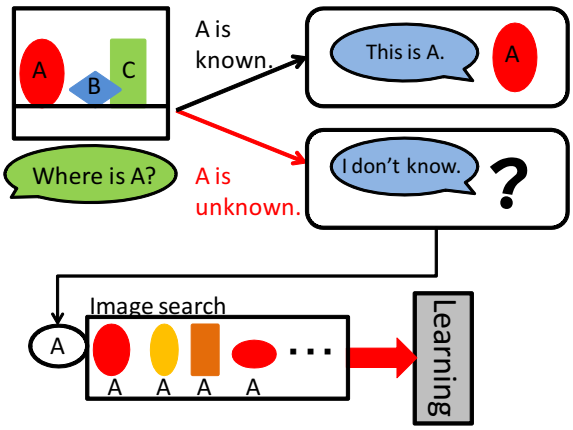

Figure 2. System flow for known objects and unknown objects

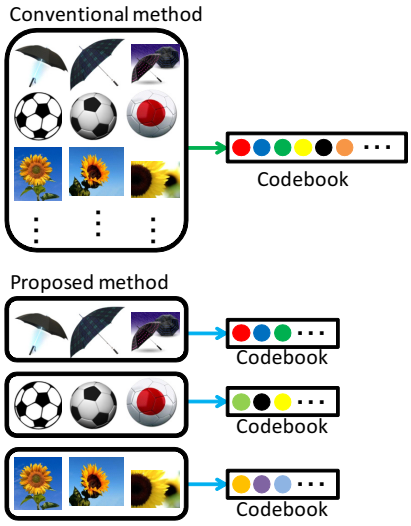

Figure 3. Difference between conventional method and proposed method

order to identify an unknown object, new common visual words suitable for the unknown object have to be re-trained by the k-means algorithm using the SIFT features of the unknown object images and all training images. In this viewpoint, the use of common visual words may not be suitable for training a new category object due to computation cost. To alleviate this problem, we introduce an unknown object identification method that creates the individual category visual words suitable for unknown objects instead of common visual words. Also, to discriminate an unknown object from others, the rejection visual words are introduced, which can absorb the features of other objects and the background.

This paper is organized as follows. In Section 2, the proposed method is described. In Section 3, the performance of the proposed method is evaluated for 10-class image dataset recognition and identification. Section 4 summarizes the paper and discusses future work.

2 Unknown Object Identification

In our proposed method, visual words for each object category are created independently. The difference between the conventional method based on BoF and the proposed method is shown in Fig. 3. The conventional method creates a common codebook using training images of all categories. Since the created histograms are different in each category, the common codebook is effective for their recognition. However, when an unknown object is given as a test image, the common codebook is not so discriminative because there is no guarantee that the specific features of the unknown object are included in the common codebook.

In order to identify an unknown object, new common visual words suitable for the unknown object have to be re-trained using the unknown object images and all training images. Therefore, the use of common visual words may not be suitable for training a new category object due to computation cost. On the other hands, in our method, as the individual codebooks are created using the training images of individual category, the updating of the individual codebooks is not required.

The most important point of the individual codebook is that the BoF (frequency histogram) of the unknown object fits the other category object due to the background features. To address this problem, we introduce rejection visual words that can absorb the background features as well as the features of the other objects.

2.1 Rejection Visual Word

The codebook created for each object category may have a problem in vector quantization, where images include the object features and also unrelated features locating in the background.

Hereafter, we will refer to these unrelated features as "noises". Since there are many noises in a dataset (for example, Caltech-101 for image recognition), there is much more noise in Web images. To solve this problem, we introduce "Rejection visual word" as shown in Fig. 4.

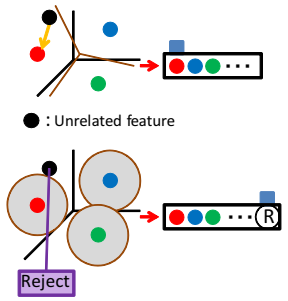

Figure 4. Reject visual word

In the conventional BoF method shown in Fig. 4 (top drawing), the unrelated features (black point) are classified into one of the visual words (red point). As a result, the created histogram may differ from real histogram.

However, if each visual word has its own area enclosed with some radius as shown in Fig. 4 (bottom drawing), and if the outside area of the visual-words circles is regarded as a "rejection visual word", the unrelated features are excluded into the rejection visual word and not counted into the histogram. Therefore, a histogram that better resembles the real histogram is obtained as shown in Fig. 5.

In the figure, two cases are shown, where vector quantization of the umbrella image is carried out with rejection visual word or without rejection visual word. If the codebook has no rejection visual word, the

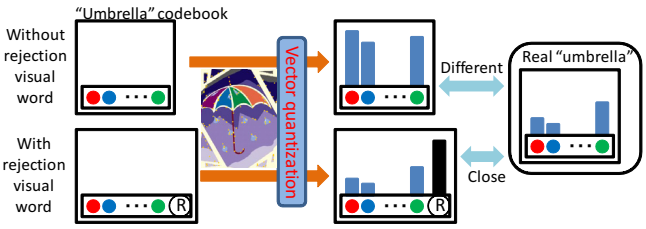

Figure 5. Histogram with/without rejection visual word

obtained histogram differs from a real umbrella histogram. However, if the codebook has a rejection visual word, since the unrelated features are rejected, a histogram closely resembling the real umbrella histogram is obtained.

2.2 Object Identification

As the codebook is created for each object category, the object identification can be carried out easily as shown in Fig. 6.

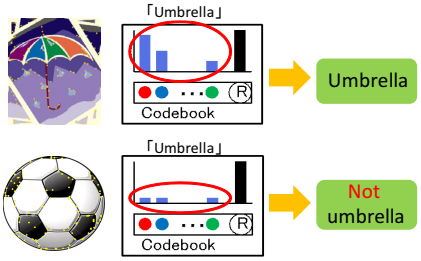

Figure 6. Identification

There are two examples in Fig. 6. The top figure shows the case where the "umbrella" image is quantized by using an "umbrella" codebook. The bottom figure shows the case where the "soccer ball" image is quantized by using the "umbrella" codebook. Since the umbrella image includes many umbrella features, the visual words histogram is clearly represented and it can be recognized as an umbrella. On the other hand, since the soccer ball image includes only a few umbrella features, and many unrelated features ("noises") are excluded to the rejection visual word, the visual words histogram is weakly represented, and it cannot be recognized as an umbrella. In short, the object can be identified by thresholding the total value of the frequency on the visual words excluding the "noise" into the rejection visual word, the visual word in the histogram. If the total value is greater than some threshold, the object is identified as the exact object such as umbrella. On the contrary, if the value is less than the threshold, the object is identified as an "unknown object", namely, an object other than an umbrella.

Conventional methods, such as kNN [5] or SVM [6], have to be trained using training image histograms. However, the proposed method does not need the training image histogram, but uses only the input image's histogram. Thereby, the calculation time can be reduced greatly.

3 Experiments

3.1 Dataset

In order to validate the effectiveness of the proposed method, 30 images were collected from Google Image Search as training images, and 20 images from Caltech-101 dataset as test images for each category.

If the number of training images is less than 30 for each unknown object, inadequate object images trend to be retrieved, so that the number of training images is set for 30.

Three experiments were carried out; the object recognition task, the object identification task and the validation task of the rejection visual word. First, the object recognition task is carried out for 10 classes (dalmatian, dollar bill, hedgehog, pizza, soccer ball, stop sign, sunflower, umbrella, Windsor chair and yin yang). Second, the object identification task is carried out for identification of the same 10 classes. Third, the validation task is carried out for evaluating the circle radius of the rejection visual word, as shown in Fig. 4.

3.2 Experiment Results

The results of the object recognition task are shown in Fig. 7, where the vertical axis and the horizontal axis indicate the recognition rate and the codebook size, respectively.

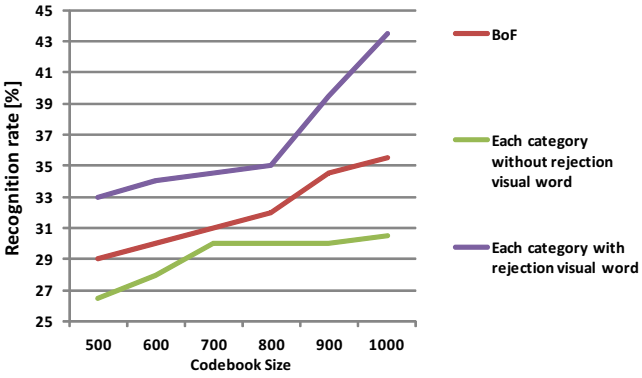

Figure 7. Results of object recognition

In this paper, since the codebook is created for each individual category, the codebook size is thought to be smaller than that of the conventional BoF. Therefore, the codebook size was set from 500 to 1,000 in our experiments.

The purple line in the figure shows the result of the proposed method with rejection visual word. The green line shows the result of the proposed method without rejection visual word. The red line shows the result of the conventional BoF method.

From this figure, it can be said that the recognition rate of the proposed method without the rejection visual word is lower than that of the conventional BoF method. This is because if the rejection visual word is not employed, the created histograms differ from the real histograms. However, the proposed method with the rejection visual word is better than the conventional method. This indicates that the rejection visual word plays an important role in object recognition in this case.

The experiment results of the object identification are shown in Fig. 8. The identification task is defined as answering either Yes or No to the question "Is this A?". The results were evaluated by Precision Recall measure [4] average. We compared the proposed method with kNN method shown in Fig. 6. From Fig. 8, if the rejection visual word is not used in kNN, the result is lowest at F-measure. Compared to kNN with rejection visual word, the proposed method showed good precision and F-measure value. In addition, the computation time was greatly reduced.

The third experiment of the validation task for rejection visual word is defined as shown in Fig. 9. The

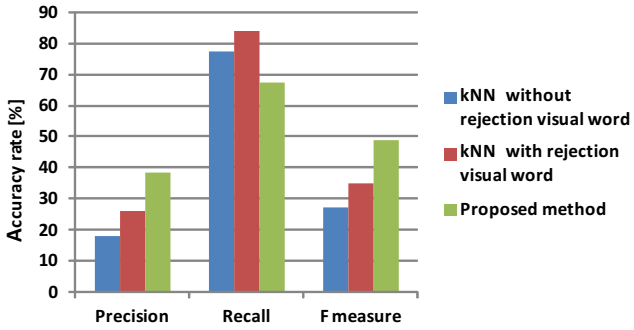

Figure 8. Results of the object identification

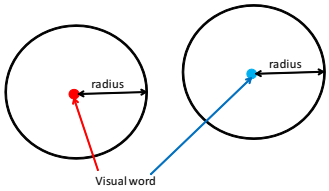

Figure 9. Reject radius

radius controls the variation of each visual word. If the radius is zero, all features are rejected. If the radius is too large, all features are classified into one of the visual word so that they are not rejected. The radius was set from 0 to 500. The codebook size was 1,000. It was the best size as shown in Fig. 7.

The result of the validation task is shown in Fig. 10. The best recognition rate is 43.5% at radius 300. The

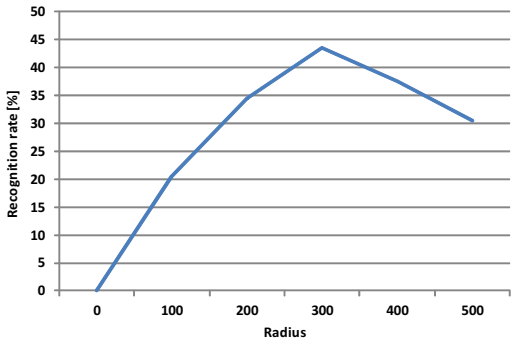

Figure 10. Results of the validation task for rejection visual word

worst rate is naturally at radius 0 because all the features are rejected and the created histogram is all zero. If the radius is set too large, it almost equals no rejection visual word method, like the green line in Fig. 7. Therefore, the radius needs to be decided for each category properly.

From those results of three experiments, the effectiveness of the proposed method is showed. But the recognition and identification accuracies are still low, because of false training images collected from Web, as shown in Fig. 11. These images disturb creating accurate visual words. Since a method for improving image search is extensively studied in [7], by collecting images close to the object, the recognition/identification accuracy will be improved.

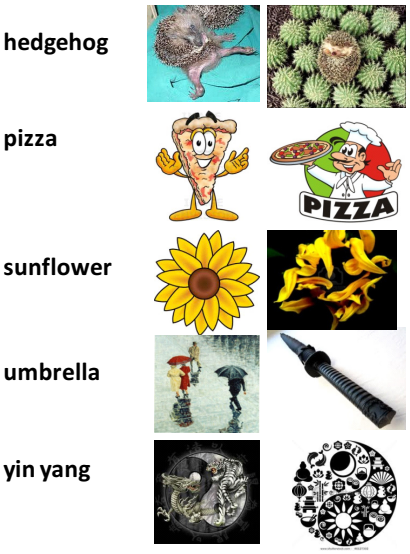

Figure 11. Examples of false images collected from Web

4 Conclusion

In this paper, we introduced an identification method that creates the individual words as well as the rejection visual words, which can absorb the features of other object or the background.

Moreover, we introduced an identification method without training images histograms. The results of our experiments led to the recognition rate being improved by 8.0 points, and the identification rate of the new method was 14.1 points better than the conventional method. In addition, the processing time was greatly reduced.

In the future, we are planning to work on automatically determining the rejection radius threshold and codebook size.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[2] D. G. Lowe, "Object recognition from local scale-invariant features," *Proc. IEEE International Conference on Computer Vision*, pp. 1150–1157, 1999.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[4] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," Technical report #1551, University of Wisconsin Madison, 2006.

[5] B. Dasarathy, "Nearest Neighbor Pattern Classification Techniques," *IEEE Computer Society Press*, Los Alamitos CA, 1991.

[6] J.A.K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, 9, pp. 293–300, 1999.

[7] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, vol. 87, no. 3, pp. 316–336, 2010.