Object Category Recognition by Bag-of-Features using Co-occurrence Representation by Foreground and Background Information

Tomoyuki Nagahashi
Dept. of Computer Science
Chubu University
Aichi, Japan
nagahashi@vison.cs.chubu.ac.jp

Hironobu Fujiyoshi
Dept. of Computer Science
Chubu University
Aichi, Japan
hf@cs.chubu.ac.jp

Abstract

This paper proposes an object category recognition method based on a bag-of-features algorithm that uses co-occurrence expressions of foreground and background information. Since bag-of-features algorithms use histograms to express features, they ignore object position information. They are considered more precise since they only code feature values in foreground regions comprised of the target categories. We investigated a method that first uses image segmentation to extract foreground regions, then codes only the feature values for those regions. We compared this method's recognition rate to the recognition rate of the standard bag-of-features algorithm. Our experimental findings demonstrated that coding feature values from both foreground and background regions resulted in more precise recognition than coding feature values from foreground regions only. Based on these findings, we have proposed a bag-of-features algorithm that focuses on the co-occurrence of local features in the foreground and background, and uses 2+1D vector quantization histograms. Our evaluation testing showed that the proposed algorithm had a recognition rate about 3.8% better than the standard bag-of-features algorithm.

1 Introduction

Object category recognition contained in unconstrained real-world images by their ordinary names is called general object recognition[1]. For the problem of object category recognition, which determines the categories to which the objects in an image belong, the bag-of-features model[2] has been proposed and has been much researched in recent years. The bag-of-features model regards an image as a local feature set and uses local feature histograms to categorize the image features. Csurka *et al,*[2] proposed using feature points that are invariant under affine transform[3] for the local features used in the bag-of-features and using SVM for the classifier. Fei-Fei *et al,*[4] proposed SIFT feature[5] descriptions of grid points as the local features and the LDA (Latent Dirichlet Allocation) generative model as the classifier, reporting a recognition rate of 64% for 13 categories of natural scenery images.

In the bag-of-features approach, image features are described as local feature histograms, so position information on the local features is not used. In recent years, there have been attempts to raise recognition accuracy by using local feature position information in the bag-of-feature approach, and methods have been proposed for describing spatial relations in images[6] and describing features in image foreground regions[7][8][9]. As a method of describing the spatial position information of local features, Lazebnik *et al,* proposed Spatial Pyramid Matching[6] using multi-resolution histograms for representation. Spatial Pyramid Matching divides an image into a grid so that the position relations of the resulting small region units can be described. As a method that uses the local features of foreground regions, Marcin *et al,*[7] proposed a method in which the background information is suppressed by using a masked image in the training data to obtain a weight distribution from the relations of features to match the foreground regions of the input image; that weight distribution is then used to construct a vector quantization histogram. Suzuki *et al,*[8] used SVM to distinguish extracted features as foreground or background, and increased recognition accuracy by describing only the foreground feature. Bosch *et al,*[9] proposed a method in which regions of interest (ROI) are determined on the basis of feature similarity to the training image, and then shape and appearance features from those regions are learned by Multiple Kernel Learning. These methods of describing features from foreground regions are based on the assumption that background information is noise; they seek to improve the recognition rate by excluding the background and describing only foreground features. Nevertheless, a method in which recognition accuracy was improved by using the relations between objects as context[10] and the feasibility of recognition by using information from the background alone[11] have been reported. In this case, objects that are not in the target category are taken to belong to the background. Thus, we can say that, for image classification, it is not clear what kinds of features bag-of-features captures and what features work well.

Therefore, we first investigate what features are effective in category recognition by taking the target category region to be foreground and objects not in the target category to be background, and analyze the foreground and background region features of the bag-of-features. We compare the recognition rates of a method in which the foreground region is known in advance and features are extracted from the entire image, a method in which features are extracted only from the foreground region, and a method in which features are extracted respectively from the foreground region and background region. We use AdaBoost as the classifier and investigate what kinds of features are selected by the weak classifier and describe the effectiveness of foreground information and background information. Next, we use the results of the investigation to cre-

ate a bag-of-features using the co-occurrence local features in the foreground and background, propose 2+1D vector quantization histograms for the foreground and background. We also demonstrate the effectiveness of this approach.

2 Bag-of-features and image local features

In this section, we describe application of bag-of-features to object category recognition. The bag-of-features method applies the concept of bag-of-words[12] that is used for document classification to images. The bag-of-words approach takes as features the set of words in a sample of text and their frequency of occurrence, disregarding word order; similarly, bag-of-features performs image recognition on the basis of a set of local image features, disregarding feature position information. In the bag-of-features approach, local features are first extracted from the image and then vector quantization.The vector quantization results are used to prepare in advance a code book called a visual word dictionary. The vector quantized features are used to create vector quantization histograms. The vector quantization histograms are then input to the classifier as image features to compute the object categories.

In this paper, SIFT features[5] are used as local features. SIFT features are used to discover points of correspondence between images; they take the extreme values of the DoG as keypoints. For image classification, however, increased precision by using points sampled on a grid as keypoints has been reported[4]. Therefore, in this method, the keypoints are taken to be points on a 10-pixel grid, and SIFT features are extracted from scales of 10, 20, 30, 40, and 50 relative to those points. SIFT features can be used for local region description that is robust to rotation and luminance, and so serve as effective local features for the bag-of-features method.

The classifier used for the bag-of-features is an ordinary multi-class SVM. For the classifier in the work reported in this paper, we used AdaBoost[13], which allows easy analysis of the features used for discrimination. The AdaBoost classifier distinguishes only two classes, which are whether or not the category is the target category. In multi-class discrimination with AdaBoost, a classifier is constructed for each category, and the category is decided from the maximum value of the normalized responses of the respective classifiers as One-Versus-Rest.

3 Investigation of effective features for bag-of-features

In this section, we investigate the recognition rate for the method in which only foreground region features are described in the bag-of-features and the method in which the feature description includes background information. We also analyze the weak classifiers selected by AdaBoost and investigate what features are effective in recognition from the results of feature visualization.

3.1 Overview

To confirm how the foreground and background can be captured in bag-of-features, we compared the different feature extraction methods described below.

- **Entire Image** Extract local features from the entire image and create vector quantization histograms,
- **Foreground only** Construct vector quantization histograms using only local features that were extracted from the foreground only and whose positions are within the foreground region.
- **Foreground and background** Separate the foreground region and background region according to the positions of local features and create vector quantization histograms for those regions.

For the SIFT features, we describe the features by these methods and use AdaBoost to construct One-Versus-Rest multi-class classifiers.

For the evaluation, we used the Caltech256 database for image classification[14]. Caltech256 includes no segmentation data, so segmentation regions were prepared manually for each image. Because it was difficult to prepare complete segmentation data, we used nine categories for the evaluation experiments. The categories, which are butterfly, elephant, hawksbill, helicopter, motorbikes, airplanes, car-side, faces-easy, and toad images from Caltech256. Many of the images contained foreground and background. For the evaluation value, we used the f-measure. Thirty images were selected from each category randomly to serve as training images; the remainder were used for evaluation. Evaluation was based on the mean f-measure for three trials. We chose 100 as the size of the visual word dictionary used in vector quantization from the results of preliminary experiments.

This paper investigates the following points.

- **Effective features obtained from recognition rates** We compare the recognition rates of the different feature extraction methods to test the effectiveness of using background information.
- **Effectiveness of background information** We investigate the visual words of the vector quantization histograms obtained from the training data to determine how foreground and background are processed in bag-of-features.

3.2 Effective features obtained from recognition rates

The recognition rates for methods that describe features from the entire image, from the foreground only, and from the foreground and background are presented in Table 1. We can see from those results that describing features from the foreground only produces a recognition rate that is 7% higher than describing features from the entire image. Next, comparing foreground-only feature description with description of features from both foreground and background, we see that the latter method results in 5.6% higher recognition accuracy. The difference between the methods of feature extraction from foreground only and feature extraction from both foreground and background is only whether or not there are background features. Therefore, adding background information improves recognition accuracy and features that are effective for recognition are also present in the background.

Table 1. Recognition rate.

Entire Image	Foreground only	Foreground and background
0.64	0.71	0.77

4 Object category recognition by feature co-occurrence in foreground and background

Object category recognition by bag-of-features uses background information as well as foreground information, and we know that the method of describing features in those respective regions is effective. We therefore propose a method that increases image classification precision by using foreground and background co-occurrence. It is already known that recognition accuracy is improved by describing visual word co-occurrence[15]. The proposed method is intended to raise recognition accuracy by describing the co-occurrence of visual words in the foreground and background. The processing flow of the proposed method is illustrated in Fig. 1. In the training, a keypoint classifier and a category classifier are constructed. Foreground and background keypoints from masked images are used as training data to construct the keypoint classifier. The category classifier is constructed using 2D and 1D vector quantization histograms from foreground and background keypoints as features. For unknown input images, foreground and background are distinguished by using a pre-constructed keypoint classifier for keypoints on a grid. Then, the keypoints of the distinguished foreground and background are looked up in the visual word dictionaries of the foreground and background and 2D and 1D vector quantization histograms are constructed. The vector quantization histograms are used by the pre-constructed category classifier to distinguish the image categories.

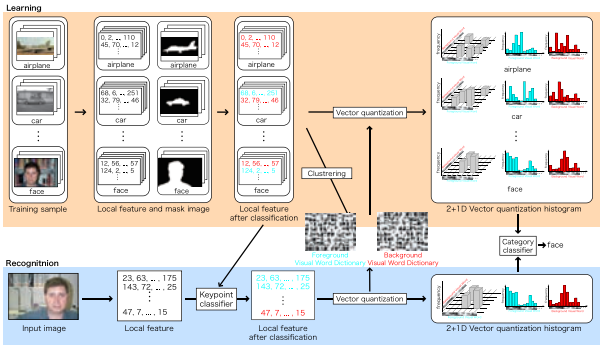

Figure 1. Overview of proposed method.

4.1 Construction of the keypoint classifier

In the training data, keypoints are separated into foreground and background by masked images. For the unknown input images, however, masked images cannot be used, so a way to distinguish keypoints in the foreground and background is required. We therefore used the local feature discrimination proposed by Suzuk *et al*,[8] to construct the keypoint classifier. Manually masked images were used for the training data in constructing the keypoint classifier. The local features of the foreground keypoints and background keypoints obtained from the masked images were used to construct the keypoint classifier by SVM. We used the LIBSVM SVM library[1]. The keypoint classifier built using this SVM discriminates foreground and background keypoints in the input images. Experiments using the 2,411 images used for evaluation produced a recognition rate of 77.9%.

[1]http://www.csie.ntu.edu.tw/čjlin/libsvm/

4.2 Creating a 2+1D vector quantization histogram

The local features separated into foreground and background by the keypoint classifier were used to create the 2D and 1D vector quantization histograms as shown in Fig. 2. First, the discriminated foreground and background keypoints are vector quantized by look-up in the respective visual word dictionaries. Then, a correspondence is made with background keypoints that are present in regions whose scale is an n-multiple of the scale of the keypoints, and 2D vector quantization histograms are created by voting on visual words separated in a 2D space. When doing so, the local features that were not used in the 2D vector quantization histogram voting are used to create 1D vector quantization histograms for the foreground and background. In the training, keypoints distinguished using masked images were used, but for the unknown input images, the keypoint classifier is used to create the 2D vector quantization histogram.

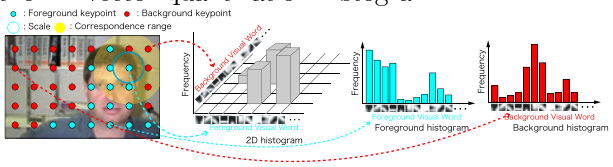

Figure 2. 2+1D vector quantization histogram.

5 Evaluation Experiments

We conducted evaluation experiments to test the effectiveness of the proposed method by comparing the proposed method of bag-of-features using 2+1D vector quantization histograms with previous method 1, which is the conventional bag-of-features approach, and previous method 2, which is bag-of-features with feature extraction from foreground and background regions. The evaluation experiments used the same data sets and evaluation methods as described in section 3.1. For the training data, foreground and background were masked; for the evaluation data, keypoints were used to discriminate foreground and background. The visual word dictionaries were the same for all methods. For comparison with the previous methods, the results for a scale factor of four was used.

5.1 Experimental results

5.1.1 Comparison with previous methods

The recognition rates for the various methods (Fig. 3(a)) show that the recognition rate for previous method 2, which involves foreground and background feature extraction, was 12.6% lower than previous method 1, the ordinary bag-of-features method. When

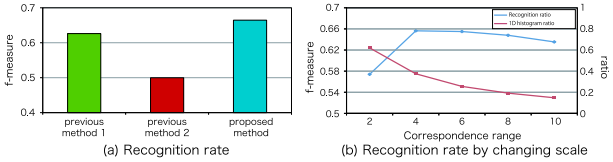

Figure 3. Recognition rate.

the foreground and background are already known, extraction of features from the respective foreground and background regions increased the recognition rate, but inclusion of errors in recognition by the keypoint classifier is considered to result in the overall lower recognition rate. However, the proposed method achieved a 3.8% higher recognition rate than previous method 1.

5.1.2 Range for representing co-occurrence

In the proposed method, the range of correspondence of foreground and background keypoints is determined on the basis of the scale of the foreground keypoints. We therefore investigated the image classification accuracy for correspondence ranges of 2, 4, 6, 8, and 10 times the scale of the foreground keypoints. The recognition rates and ratios of local features used in the 1D histogram for various correspondence ranges are presented in Fig. 3(b). We can see from the figure that larger correspondence ranges mean lower ratios of local features used in the 1D histogram. Also, the keypoint scale of a factor of four produces the highest recognition rate. That increasing the correspondence scale range reduces the recognition rate means that spatial constraints are effective in representing the co-occurrence of local features. Furthermore, because of the drop in recognition rate that comes with reducing the local features used for the 1D histogram, features that cannot be represented by the 2D histogram are considered to be supplemented by the 1D histogram.

5.2 Effects of the 2D histogram

To investigate the effect of feature representation by the proposed method, consider the 1D histograms for foreground and background of previous method 2 for butterfly, elephant, and toad and the 2D histogram of the proposed method shown in Fig. 4. From the 1D

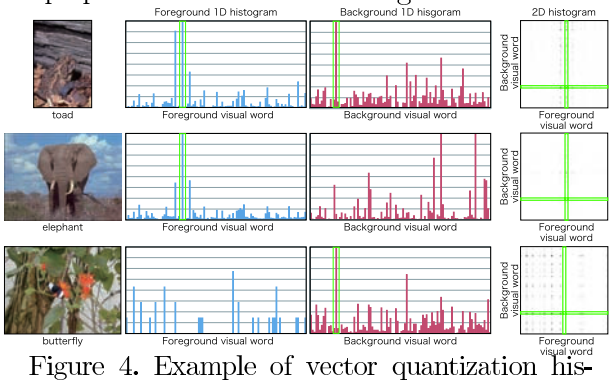

Figure 4. Example of vector quantization histogram.

histogram, we see that the toad and the elephant both have a high frequency of the same foreground visual word, and that the toad and the butterfly both have a high frequency of the same background visual word. Thus, visual words that are toad features, are not very effective in separating the toad from the butterfly and elephant. However, by representing the foreground and background with the 2D histogram, features that can distinguish toads from elephants and butterflies can be represented by visual words of high frequency only in both the foreground and background.

6 Conclusion

We investigated effective features for the bag-of-features model and proposed a bag-of-features method that uses 2+1D vector quantization histograms for representing the co-occurrence of local features in the foreground and background. First, we analyzed the foreground and background region features to investigate what features are effective for category recognition in bag-of-features. We used the results to improve recognition accuracy by description that separates the foreground and background.

On the basis of the results, we proposed a bag-of-features method that represents the co-occurrence of local features in the foreground and background with 2+1D vector quantization histograms. Experiments showed that the proposed method increased recognition accuracy by 3.8% relative to previous methods. Experiments in which the range of co-occurrence representation was varied showed that spatial constraints are effective for co-occurrence representation.

In future work, we will aim for higher precision through co-occurrence representation that takes spatial position relations into account.

References

[1] Yana, K.: Current state and future directions on generic object recognition. IPSJ Transaction on Computer Vision and Image Media **48** (2007) 1–24

[2] Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision. (2004) 1–22

[3] Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: ECCV (1). (2002) 128–142

[4] Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR. Volume 2. (2005) 524–531

[5] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110

[6] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178

[7] Marszałek, M., Schmid, C.: Spatial weighting for bag-of-features. In: CVPR. Volume 2. (2006) 2118–2125

[8] Suzuki, K., Matsukawa, T., Kurita, T.: Bag-of-features car detection based on selected local features using support vector machine. Technical report of IEICE. PRMU **108** (2009) 7–12

[9] Bosch, A., Zisserman, A., Munoz, X.: Image classification using rois and multiple kernel learning. IJCV 2008 (2008)

[10] Okabe, T., Kondo, Y., Kitani, K.M., Sato, Y.: Recognizing multiple objects based on co-occurrence of categories. Progress in Informatics (2010) 43–52

[11] Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV **73** (2007) 213–238

[12] Manning, C.D., Schtze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (1999)

[13] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the Second European Conference on Computational Learning Theory. (1995) 23–37

[14] Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)

[15] Herve, N., Boujemaa, N.: Visual word pairs for automatic image annotation. In: ICME, Piscataway, NJ, USA, IEEE Press (2009) 430–433