Journal of Robotics, Networking and Artificial Life, Vol. 1, No. 2 (September 2014), 160-163

Multiple-window Bag of Features for Road Environment Recognition

Shou Morita, Joo Kooi Tan, Hyoungseop Kim and Seiji Ishikawa
Department of Mechanical &Control Engineering, Kyushu Institute of Technology
Sensuicho 1-1, Tobata, Kitakyushu, 804-8550, Janan
E-mail: motira@ss10.cntl.kyutech.ac.jp, {etheltan, kim, ishikawa}@cntl.kyutech.ac.jp

Abstract

The idea of Bag of Features (BoF) is recently often employed for general object recognition. But, as it does not take positional relations of detected features into account, the recognition rate is still not very high for practical use. This paper proposes a method of describing the feature of an object by the BoF representation which considers positional information of the features. Although the original BoF representation is applied to an entire image, the proposed method employs multiple windows on an image. The BoF representation is applied to each of the windows to represent an object in the image interested for recognition. The performance of the proposed method is shown experimentally.

Keywords: Object recognition, bag of features, multiple-windows, VLAD, computer vision.

1. Introduction

The future of a mankind will be more and more complicated and will definitely need the help of an intelligent robot. Then the robot must be equipped with a strong ability of object recognition. On the other hand, a hand-held camera and a wearable computer system which can recognize every object around a blind person may help him/her a lot in living a daily life safely as well as conveniently. Such a system again needs to have a strong ability of object recognition. Various techniques of object recognition have been developed to date. But such techniques normally employ the features depending solely on the objects interested. The features common to every object should be considered to develop a general objects recognition method.

General objects recognition has been paid much attention among computer vision researchers recently. A well-known general object recognition technique is the idea of Bag of Features (BoF) [1]. It is a point-based feature description method and describes every object using a visual word dictionary. But it describes an object as a set of feature points without considering positional information. The positional information, or to know how feature points distribute on an object, is actually important information for its recognition. An idea of spatial pyramid matching (SPM) [2] is proposed in order to take positional information of feature points into account. But it is not very effective, since the method segments an image into 2^n by 2^n regions with no overlap some of which may contain only the background of the image.

The present paper proposes a method of describing the feature of an object by BoF representation which considers positional information of the features. Although the original BoF representation is applied to an entire image, the proposed method employs multiple overlapping windows on an image. The BoF representation is applied to each of the windows to represent an object in an image. In this way, the positional information among obtained BoFs is employed for recognizing an object interested.

2. BoF and VLAD

The idea of BoF is overviewed in the first place followed by giving the concept of VLAD (Vector of Locally Aggregated Descriptors) [3] proposed as another representation of BoF.

Given an object image, the SIFT operator [4] is applied to the image to derive a number of feature

Published by Atlantis Press

points on the object. The point is described by a 128-dimensional vector. It is then projected into a 128-dimensional feature space. A number of object images are respectively transformed into the feature space as a set of feature points. The feature space then contains a large number of the feature points, to which clustering is applied to define some hundreds or thousands of prominent classes. Let the number of the class in the feature space be denoted by M. A class C_i is represented by a feature vector $\boldsymbol{v}_i$ (i=1,2,...,M). The feature space is then defined as a visual word dictionary (VWD) by the set $V=\{\boldsymbol{v}_i | \ i= 1,2,...,M\}$: Vector $\boldsymbol{v}_i$ is referred to a visual word (VW) within the dictionary.

An object image is then described using the VWD. Given an object image, the SIFT operator is applied to the image and the feature points are extracted from the object. Once they are projected into the VWD, they distribute around the VWs which represent the object. Let the number of the feature points distributing around a VW $\boldsymbol{v}_i$ be denoted by f_i. This is actually the frequency of a histogram of the chosen VWs. The object is then characterized by a M-dimensional feature vector

$$\boldsymbol{w} = (f_1, f_2, ..., f_M) \tag{1}$$

This is called Bag of Features (BoF). An object is finally identified by the BoF $\boldsymbol{w}$.

Instead of using the frequency of the VWs, another description of an object [3] is proposed using a VLAD (Vector of Locally Aggregated Descriptors). If a feature point extracted from an object image by SIFT is denoted by $\boldsymbol{x}$, the VLAD feature vector is defined by

$$\boldsymbol{w}_i = \sum_{v(\boldsymbol{x})=i} (\boldsymbol{x} - \boldsymbol{v}_i) \tag{2}$$

After all, the VLAD expression provides a 128M-dimensional feature vector of the form

$$\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M) \tag{3}$$

The magnitude of the component $\boldsymbol{w}_i$ depends largely on the feature points distributing around VW $\boldsymbol{v}_i$.

3. Multiple-window BoF

The proposed method puts K mutually overlapping windows (W_1, W_2, ...,W_K) on an image in order to consider positional relation among extracted feature points, which is a strategy different from the original BoF [1]. It also differs from SPM [2] in the overlap of the windows. The idea of multiple windows is shown in Fig. 1. The proposed method also introduces VLAD for describing BoF. This means to put more emphasis on the VWs which have many feature points around them than the frequency description.

Locating windows on an image has three variations:
(i) Random location; Windows are randomly located on an image;
(ii) Considering feature points distribution: Arranging windows more at the spots where many feature points distribute;
(iii) Combining (i) and (ii): Arranging windows randomly under the condition that the spots where there are many feature points have priority in the arrangement.

Among the above three strategies, (ii) is natural and reasonable, since the present object recognition is feature-points-based recognition. It is, however, important to consider some randomness to escape from over-learning against training images. This is the reason why (iii) is considered. (i) is conducted for the comparison with (ii) and (iii).

The size of the window and the randomness in the windows placement is determined experimentally.
Let the number of the visual words in a window W_k (k=1,2,...,K) be denoted by M_k. In the original idea of BoF, frequency in the BoF histogram is employed for the components of feature vector $\boldsymbol{w}_k$. Instead of using the histogram, the present method introduces VLAD for describing a BoF. The magnitude of the VLAD $\boldsymbol{w}_{km}$ at visual word $\boldsymbol{v}_{km}$ (m=1,2,...,M_k) in window W_k becomes large, if many feature points distribute close to $\boldsymbol{v}_{km}$ in a biased way. After all, the overall feature vector $\boldsymbol{w}$ is defined by

$$\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_k, ..., \boldsymbol{w}_K), \tag{4a}$$

$$\boldsymbol{w}_k = (\boldsymbol{w}_{k1}, \boldsymbol{w}_{k2}, ..., \boldsymbol{w}_{Mk}). \tag{4b}$$

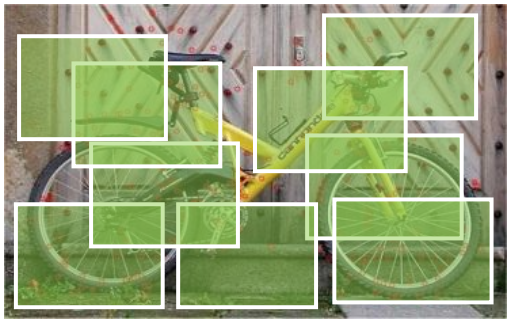


Fig. 1. Multiple-windows set on an object image.

The dimension of the feature vector $\boldsymbol{w}$ is therefore $128M$ ($M=M_1+M_2+...+M_K$).

The recognition strategy employs a nonlinear SVM based on one-versus-rest classification.

4. Experimental Results

An experiment was conducted using the images in an road environment. The employed objects for recognition are a pedestrian, a traffic signal, a car and a bicycle (See Fig. 2). They are all principal objects in the road environment. The number of images used for the training of a SVM is 800; 100 with each object and 400 negative images. On the other hand, the number of images used for test is 500; 100 with each object and 100 negative images. Used PC has a 3.40 GHz CPU with 8 GB memories.

The first experiment, **Exp_1**, was done to examine the performance of VLAD. The original method and SPM employing a frequency histogram for BoF representation were compared to those employing VLAD for BoF representation. The result is given in Table 1.

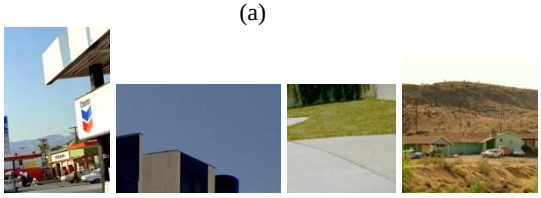
In the second experiment **Exp_2**, the proposed method employing multiple windows is examined its performance with respect to the three cases of windows placement explained in the former section; (i) placement at random, **P_R**, (ii) placement considering feature points distribution, **P_FPD**, and (iii) placement considering feature points distribution and randomness, **P_FPD&R**.

The experimental result is shown in Table 2. In Exp_2, the number of VWs is parameterized and it varies from 50 to 500 per window.

As seen in Table 2, the recognition rate is the maximum when P_FPD&R is adopted for windows placement and 200 VWs are employed with every window. The third experiment, **Exp_3**, was conducted under the employment of multiple windows with P_FPD&R, 200 VWs with each window, and VLAD expression for BoF. The result is shown in Table 3.

The number of used windows in Exp_2 and Exp_3 is 10 whose size is approximately 1/3 of the entire image. They are arranged as shown in Fig. 3, in which windows are placed where many feature points exist in (a), whereas randomness is considered in addition to the feature points distribution in (b). These windows placements are kept unchanged through Exp_2 and Exp_3.

(a)



(b)

Fig. 2. Objects for recognition: (a) Positive samples; a pedestrian, a traffic signal, a car and a bicycle, (b) negative samples.

Table 1. Result of Exp_1: Original BoF & SPM without/with VLAD.

Methods	O-BoF*	O-BoF +VLAD	SPM	SPM +VLAD
Rec. rate [%]**	63.0	67.4	69.8	72.8

* Original BoF

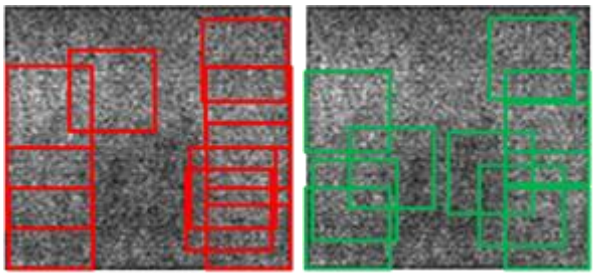
** Recognition rate

Table 2. Result of Exp_2: Multiple window BoF.

	P_R			
No. VWs / window	50	100	200	500
Recogn' rate [%]	68.6	72.8	69.4	68.4
	P_FPD			
No. VWs / window	50	100	200	500
Recogn' rate [%]	71.0	69.4	69.4	68.8
	P_FPD&R			
No. VWs / window	50	100	200	500
Recogn' rate [%]	71.8	73.0	73.2	67.6

Table 3. Result of Exp_3: Multiple window BoF with VLAD.

	Proposed method
Recognition rate [%]	74.8



(a) (b)

Fig. 3. The windows employed in Exp_2 and Exp_3: (a) Case (ii), (b) case (iii).

Table 4. Average recognition rate with respect to three strategies of windows placement.

	P_R	P_FPD	P_FPD&R
Ave. rec. rate*	69.8	69.7	71.4

*Average recognition rate [%]

5. Discussion

In Exp_1, the positive performance of the VLAD was recognized as shown in Table 1. The idea of VLAD is to put emphasis on the VW which is characteristic to a particular object more than frequency, and it worked affirmatively in the recognition of the 10 objects employed in the experiment.

On the other hand, in Exp_2, the maximum recognition rate was 73.2% when strategy (iii) in section III was employed as windows location and 200 visual words were used with each of the 10 windows. When a single window, an image itself, is employed, which is the original way of using the BoF, the recognition rate is 63.0% as seen in Table 1. Although SPM is a multiple window method, the recognition rate is worse, 69.8%, than the proposed method which is more flexible in windows placement than SPM. This fact indicates the effectiveness of the proposed use of multiple windows in the BoF-based object recognition.

Finally, the proposed method, employing multiple windows and VLAD expression, achieved 74.8% of the recognition rate as given in Table 3. This is the best result at the moment.

As for the three strategies of windows placement, strategy (iii) seems to act better than the other two, which is seen in Table 4. It shows average recognition rates with respect to each windows placement in Table 2. It may, however, be necessary to perform more experiments to make the superiority certain, since the difference is not very large.

In the employment of BoF, various weights could be considered including frequency of the SIFT feature points [1], VLAD [3], TF-IDF [5] and weighted BoF [6]. But they don't give very high recognition rates to general objects. One may need to improve this in some way.

6. Conclusion

In this paper, multiple window bag of features was proposed which considered positional relation of the feature points on an object. For BoF representation, the vector of locally aggregated descriptors, VLAD, was also employed for recognizing ten familiar objects in a traffic environment. By effective placement of the multiple windows on an image, 74.8% of recognition rate was achieved, which is satisfactory for general object recognition. However, the research should be continued to raise the recognition rate more in order to put the method into practical use.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 25350477.

References

1. G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in *Proc. ECCV: Learning in Computer Vision* (2004), pp. 1-22.
2. S. Lazebnik, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in *Proc. IEEE Computer Vision and Pattern Recognition* (2006), pp. 2169-2178.
3. H. Jégou, et al., Aggregating local descriptors into a compact image representation, in *Proc. CVPR* (2010).
4. D.G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, 60(2) (2004), pp. 91-110.
5. W. B. Frakes, R. Baeza-Yates, Information Retrieval: Data Structures & Algorithms. *Prentice-Hall, Engelwood Clifs*, (1992).
6. T. Manabe, J. K. Tan, H. Kim, S. Ishikawa, Recognizing indoor objects by weighted bag of features, in *Proc. IEEE Tensymp* (2014). (submitted)