Filtering out background features from BoF representation by generating fuzzy signatures

Qiang Qiu
Research Institute of Robotics
Shanghai Jiao Tong University
Shanghai, China
qiu6401@sjtu.edu.cn

Qixin Cao
Research Institute of Robotics
Shanghai Jiao Tong University
Shanghai, China
qxcao@sjtu.edu.cn

Masaru Adachi
YASKAWA Electric Corporation
Fukuoka, Japan
Masaru.adachi@yaskawa.co.jp

***Abstract*—Bag of Features (BoF) approach has gained its popularity in the past decade due to its simplicity and outstanding performance in computer vision tasks. However, the lack of spatial information makes the BoF method sensitive to background noise features in real-world object recognition tasks. This paper presents a method for removing background noise features from the BoF representation of images by generating fuzzy signatures. This technique treats each visual word as a fuzzy set, and defines a membership function to wipe off background features in testing images. The experimental results show that fuzzy signature can remove some background features and improve the performance of BoF method in real-world object recognition tasks.**

Keywords—object recognition; bag of features; fuzzy signatures; background features

I. INTRODUCTION

Object recognition is a hot issue in both computer vision and intelligent robotics. Computational strategies for object recognition can be roughly divided into two categories: shape-based methods and feature-based methods. While shape-based methods cannot distinguish objects with a same shape, feature-based methods are quite qualified for this work. David G. Lowe's SIFT is a popular real-world object recognition algorithm [1]; however, this method can only be used to recognize a particular instance of an object, rather than a broad class. Besides this, bag of Features (BoF) is another popular object recognition algorithm, which is competent to the task of object categorization.

The past ten years have seen the growing popularity of BoF method in object recognition tasks. This method has been a normalized method after Csurka's research [2], and its basic framework is shown in Fig. 1 and summarized as follows.

1) Training module:

a) Detection and description: Extracting local features in training images and defining them as vectors, this vector quantization procedure makes it possible for these visual textures to be processed using mathematical means.

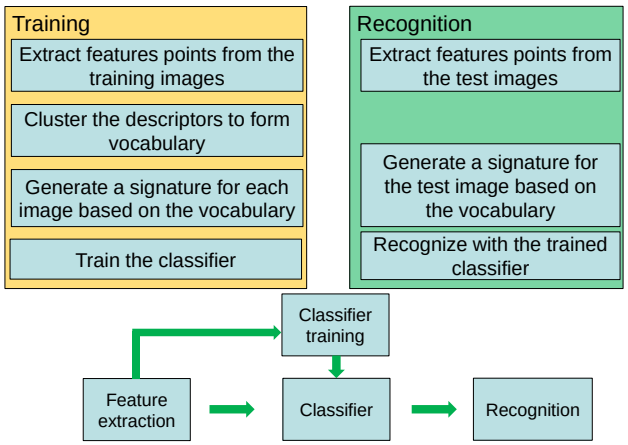

Fig. 1. Basic framework of bag of features method.

b) Vocabulary building: Cluster all training images' "visual features" into "visual words", which form a "visual vocabulary".

c) Generate signatures: Each image can be represented by a vector corresponding to the "visual words" in it. This vector can be a signature for the image.

d) Training the classifier: Using the training images' signatures and labels to train a classifier, support vector machine (SVM), for example.

2) Recognition module:

a) Detection and description: Extracting local features in testing images and defining them as vectors.

b) Generate signatures: Each testing image can be represented by a signature according to the vocabulary formed in the training module.

c) Recognition: Using the trained classifier to classify the testing image's signature, and returning an approximately name to the image.

Numerous studies have verified BoF representation's performance on Caltech 101 [3], PASCAL-2005 [4], COIL-100 [5] and other public image database. Nevertheless, one of the most notorious disadvantages of BoF is that it ignores the

spatial relationships among the local features. And the lack of spatial information makes it difficult to distinguish foreground features from background features. Reference [5] has investigated the influence of background features in PASCAL test set, and found that characteristic backgrounds would provide additional cues for classification. However, the characteristic background is helpful only with regard to the database itself, Torralba and Efros think it's a kind of dataset bias [6]. Besides, in cluttered real-world scenes, an object's background is indeterminate. This is especially true for household objects, that is to say, background features have a negative impact on real-world object recognition tasks.

This paper proposes a fuzzy signature to remove some background features. We treat each visual word as a fuzzy set, and define a membership function to wipe off background features in a given image. Our own experiments confirm that fuzzy signature can improve the BoF method's performance when the testing images have a cluttered background.

II. Related Work

The signature module can be roughly divided into two parts: assignment step and weighting step. Fuzzy signature and the traditional signature are different both in the assignment step and weighting step; besides, there are some previous works that have introduced fuzzyness into the BoF method, but they are completely different from our method, so we will briefly review these three issues in the following sections.

A. Assignment step

Before we generate signatures, we should assign each descriptor to different visual words according to the vocabulary. Additionally, there are mainly two kinds of assignment approaches, hard assignment and soft assignment.

Hard assignment refers to finding the closest visual word in the vocabulary for each descriptor, and using this visual word to represent the local feature. This approach is widely used in BoF method due to its simplicity. However, if a given feature is almost the same distance from two cluster centers, the hard assignment approach begins to underperform.

Considering the ambiguous features that lie near Voronoi boundaries, some researchers have explored soft assignment. Soft assignment, which is also called multiple assignment, refers to representing each descriptor using the k nearest visual words (cluster centers) in the vocabulary. Jegou et al. utilize this assignment approach in the BoF method. This results in better performance in retrieval accuracy [7]. But one cannot make an omelet without breaking eggs, the cost of the improved accuracy is higher search time. In Jegou's paper, a k=3 multiple assignment implementation requires 7 times the number of multiplications of hard assignment.

B. Weighting step

Since the features are assigned to different visual words in the previous sections, this section aims to represent each image with a signature. The signature of an image serves as the input data for the categorization module, which is a representation format for recording the information on the image.

The simplest way is to generate a histogram of the visual words by occurrences of each word, which is known as the term frequency (TF) representation. However, this representation may lead to poor performance when applied to words that occur too frequently. Under such circumstance, the representation becomes undiscriminating.

To address this problem, Sivic et al. implement the Term Frequency – Inverse Document Frequency (TF-IDF) representation [8]. This weighting method is superior to TF method in BoF algorithm.

In order to address the problem of ambiguous words (described in previous section), some researchers have proposed soft weighting strategies. In a manner similar to the multiple assignment, the k nearest terms are multiplied by a scaling function such that the nearest term gets more weight than the k'th nearest term. In experimental evaluations, this strategy outperforms term frequency, and TF-IDF schemes.

C. Fuzzy vocabulary

This paper is not the first to utilize fuzzyness in BoF, but existing implementation of this idea [9] focuses on generating more robust signatures, rather than on removing background features. These researchers employ the fuzzy c-means (FCM) clustering algorithm in vocabulary creation process, followed by the use of fuzzy centers and membership functions for the assignment and weighting steps. As a result, fuzzy vocabulary is more robust than traditional vocabulary in terms of mean average precision with respect to vocabulary size.

III. Fuzzy Signatures

We first describe the vocabulary building module; some preprocessing for the fuzzy signatures, followed by introduction of the fuzzy signatures generating method.

A. Build vocabulary

We simply utilize k-means clustering algorithm to obtain k clusters C_i (i.e. visual words) from the training features. Although fuzzyness isn't used during the vocabulary generating step, It's possible to define a series of fuzzy sets based on the training features and generated vocabulary.

Since there are l training features belonging to cluster C_i, where $\boldsymbol{x}_{ci}$ is the cluster center, $\boldsymbol{x}_{ij}$ is the j^{th} feature in C_i (i=1,2,…,k; j=1,2,…,l). The membership function of fuzzy set $\tilde{C}_i$ is defined as below:

$$\mu_{\tilde{C}_i}(x)=\begin{cases}\exp\left(-\dfrac{\|x-x_{ci}\|_2^2}{2\sigma_i^2}\right), & \|x-x_{ci}\|_2 \le r_i \\ 0, & \|x-x_{ci}\|_2 > r_i\end{cases}$$

Here $\|\ \|_2$ is a L_2 Euclidean distance;

The radius of fuzzy set $\tilde{C}_i$ is $r_i = \max(\|\boldsymbol{x}_{ij}-\boldsymbol{x}_{ci}\|_2)$;

Thus the standard deviation can be calculated as follow:

$$\sigma_i = \sqrt{\frac{1}{l}\left[\sum_{j=1}^{l}\left(\left\|x_{ij} - x_{ci}\right\|_2^2\right)\right]}$$

From this, we obtain k fuzzy sets, which we can use to generate a fuzzy signature for each image.

B. *Generate fuzzy signatures*

In our method, the assignment step and weighting step are combined. First we calculate the grade of membership of each feature vector $\boldsymbol{x}_n$ in each fuzzy set $(\tilde{C}_i, \mu_{\tilde{C}i})$, where $n = 1,2,\ldots,N$; $i = 1,2,\ldots,k$.

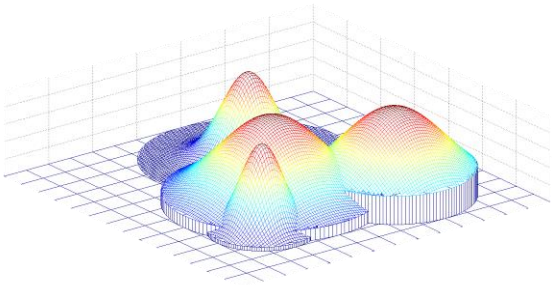

Fig. 2. Example of a two dimensional dataset's 4 fuzzy sets and their membership function distributions.

Fig. 2 shows an example of a two dimensional dataset's four fuzzy sets and their membership function distributions. The feature vector $\boldsymbol{x}_n$ is assigned to any fuzzy set $\tilde{C}_i$ whose membership function $\mu_{\tilde{C}i}(\boldsymbol{x}_n)>0$. Under this condition, each feature vector can be assigned to several visual words. Thus this assignment can be described as a kind of soft assignment scheme. Further, the membership values are used as weight factors to generate signatures. The closer a feature is, the larger its weight is, hence this weighting method is similar to soft weighting strategies. For example, a vocabulary contains 3 visual words $(\boldsymbol{C}_1,\mu_1(\boldsymbol{x}))$, $(\boldsymbol{C}_2,\mu_2(\boldsymbol{x}))$ and $(\boldsymbol{C}_3,\mu_3(\boldsymbol{x}))$, 3 features $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$ are extracted from a testing image; and the membership values $\mu_1(\boldsymbol{x}_1)=1$, $\mu_2(\boldsymbol{x}_1)=0$, $\mu_3(\boldsymbol{x}_1)=0$, $\mu_1(\boldsymbol{x}_2)=0.1$, $\mu_2(\boldsymbol{x}_2)=0.3$, $\mu_3(\boldsymbol{x}_2)=0.4$, $\mu_1(\boldsymbol{x}_3)=0.2$, $\mu_2(\boldsymbol{x}_3)=0.7$, $\mu_3(\boldsymbol{x}_3)=0.4$. Then we obtain this image's signature as follows:

$$\text{Signature} = [\mu_1(\boldsymbol{x}_1)+\mu_1(\boldsymbol{x}_2)+\mu_1(\boldsymbol{x}_3),\ \mu_2(\boldsymbol{x}_1)+\mu_2(\boldsymbol{x}_2)+\mu_2(\boldsymbol{x}_3),\ \mu_3(\boldsymbol{x}_1)+\mu_3(\boldsymbol{x}_2)+\mu_3(\boldsymbol{x}_3)] = [1.3,\ 1,\ 0.8].$$

Furthermore, some features may be assigned to none of these fuzzy sets, in other word, their grades of membership to all fuzzy sets are zeroes. This is where the most important part of our method comes into play According to our definition of the membership function, all features in the training images can be assigned to at least one visual word, but in the testing images, none of these features have contributed to generation of the vocabulary, and only the features similar to training features can be assigned to one or several visual words, the other features will be wiped off! Consider, for example, a testing image that contains both target object and background, it's obvious that the foreground features extracted from the target object are probably similar to the dictionary's visual words, and they can contribute to the signature of the image; meanwhile, the background features are more likely to be far away from the visual words, so some of them will be wiped off as noise features.

Above all, fuzzy assignment method is a combination of both soft assignment and soft weighting strategy for features within the radius of visual words, and a filter for noise features outside the radius of visual words.

IV. Experiments

In this section, we report our experimental results derived from our canned drink database. We extract SURF [10] features and descriptors from the grayscale images, and utilize k-means to generate a dictionary. Next a hard assignment method and a fuzzy signature method are both realized and their performances are compared in four experiments. Besides, the multi-class classification is done with a "one against one" support vector machine (SVM) supplied by LIBSVM [11].

A. *Image database*

Our canned drink database contains 500 images in ten classes: coca_cnmark, coca_enmark, coca_text, Fanta_cnmark, Fanta_enmark, Fanta_text, pepsi_cnmark, pepsi_enmark, pepsi_text and background.

As shown in Fig. 3, coca_cnmark is the side of the Chinese logo of Coke, coca_enmark is the side of the English logo of Coke, coca_text is the side of nutrition label of Coke; the same nomenclature is applied to Fanta and Pepsi to produce the other 6 classes; besides, all background images without objects make up the last class. And each class contains 50 images, among which 25 images are taken in a plain background, and the rest are taken in a cluttered background. Most images are medium resolution, i.e. 366 × 274 pixels, and as mentioned before, only grayscale images are used in our experiments.

Fig. 3. Example images from our canned drink database.

TABLE I. EXPERIMENT SETTINGS AND RESULTS FOR THE INFLUENCE OF TRAINING AND TESTING IMAGES

Group	Experiment Settings			Experiment Results (%)	
	Training images	*Testing images*	*Classes*	*Hard assignment*	*Fuzzy assignment*
A:plain-plain	60% of plain background images	40% of plain background images	9	98.89 ±0.91	99.12 ±0.45
B:plain-cluttered	all plain background images	all cluttered background images	9	49.06 ±2.35	78.36 ±0.60
C:both-both	60% of both plain background and cluttered background images	40% of both plain background and cluttered background images	9	98.09 ±0.63	--
D:background	all plain background images and 60% of cluttered "background"	all cluttered background images and 40% of cluttered "background"	10	--	48.13 ±0.21

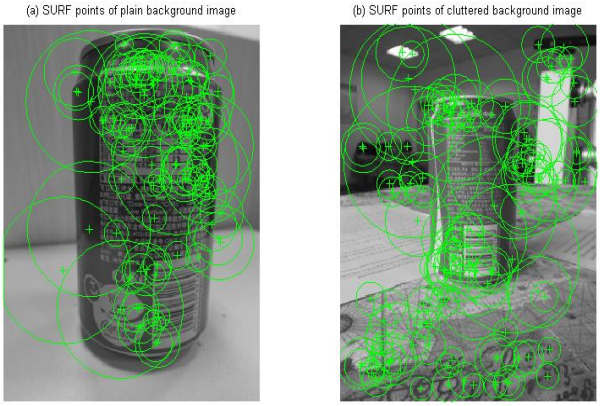

Fig. 4. The difference between the plain background image and cluttered background image. (a) shows the SURF points detected from the plain background image, (b) shows the SURF points detected from the cluttered background image.

The plain-background and cluttered-background images exhibit significant dissimilarity in terms of SURF Features. Fig. 4 shows the SURF points detected from both the plain-background and cluttered-background images. In the plain-background image, almost all feature points are detected from the object; while in the cluttered-background, some feature points are detected from the background.

B. Experiment results

The main purpose of fuzzy assignment is to remove background features from testing images and improve the BoF representation's performance in real-world object recognition. Experiments are performed to evaluate the fuzzy assignment scheme. We design four groups of experiments. As shown in table 1, group A is called "plain-plain": here 60% of the plain-background images are used as training data with the remaining served as testing data, group B is called "plain-cluttered": we use all plain-background images as training data and all cluttered-background images as testing data, and group C is called "both-both": here 60% of both plain-background images and cluttered-background images are used as training data, and the rest as testing images. Additionally, for these 3 groups, we only use 9 classes of our database (except the "background" class). At last, group D is called "background": it's similar to group B, but cluttered "background" is added as the 10^{th} class. Besides, both hard assignment and fuzzy assignment schemes are realized, the vocabulary size is set to 1000, and each experiment repeated 10 times.

Table I shows detailed results of recognition experiments using different training images and testing images. First, let us examine the behavior of fuzzy assignment on group A. In this group, traditional hard assignment results in a high MAP (Mean Average Precision) of 98.89%, however, our fuzzy assignment outperforms it with a MAP of 99.12%. In this group, both training images and testing images are taken in the plain background, and in recognition section, most features lie in the radius of visual words, so the fuzzy assignment only plays a role as soft assignment or soft weighting.

Next, consider that we only use plain-background images to train the classifier in real-world object recognition tasks, but in recognition section, the testing images are taken from the environment in real time, the background are probably cluttered. Group B examine the performance of the fuzzy assignment method in this condition. As a result, an obvious performance drop is observed in the traditional hard assignment result. The fuzzy assignment method however maintains an acceptable MAP of 78.36%. This experiment demonstrates the fuzzy assignment method's performance on filtering out the noise features from the background.

Another alternative means of recognizing an object in the cluttered background is to include the object with this cluttered background into training dataset. This is the basis for group C and the result outperforms the fuzzy assignment method. However, in real-world object recognition tasks, the target object may locate in any background, under the circumstances, it would be necessary to include the object with all the possible background into the training dataset, which is impossible. On the other hand, the fuzzy assignment can only train the object with plain background, and return an acceptable result in cluttered background.

In group D, we add the cluttered "background" into training dataset as the 10^{th} class. So the background features are also used to build a visual vocabulary, and the fuzzy assignment should be unable to distinguish between the background features and foreground features. According to the experiment on group D, the fuzzy assignment method loses its ability to filter the noise features as expected.

In general, the fuzzy assignment method is more suitable for real-world object recognition tasks, but this method is sensitive to the training dataset, or its performance will deteriorate considerably, even if only one category contains cluttered features.

V. DISCUSSION

This paper has proposed a fuzzy assignment method for BoF representations; this method has shown promising results on filtering out noise features from the background. Compared to the traditional method by simply adding objects with cluttered background into training dataset, our method is more suitable for real-world object recognition tasks. However, it is sensitive to training dataset; Use of a cluttered training dataset renders this method underperforming. Besides, this method illustrates that the performance of the BoF method can be improved by filtering out the noise features in testing images.

ACKNOWLEDGMENT

This work is supported by the Special Project for International Thermonuclear Experimental Reactor of Most, Research on Intelligent Maintenance for MCF Equipment under grant No.2011GB113005, the National Natural Science Foundation of China under grant No.61273331 and YASKAWA Electric Corporation.

REFERENCES

[1] Lowe, David G. "Object recognition from local scale-invariant features." *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, 1999.

[2] Csurka, Gabriella, et al. "Visual categorization with bags of keypoints." *Workshop on statistical learning in computer vision*, ECCV. Vol. 1. No. 1-22. 2004.

[3] Fei-Fei, Li, Rob Fergus, and Pietro Perona. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." *Computer Vision and Image Understanding* 106.1 (2007): 59-70.

[4] Jiang, Yu-Gang, Chong-Wah Ngo, and Jun Yang. "Towards optimal bag-of-features for object categorization and semantic video retrieval." *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007.

[5] Zhang, Jianguo, et al. "Local features and kernels for classification of texture and object categories: A comprehensive study." *International journal of computer vision* 73.2 (2007): 213-238.

[6] Torralba, Antonio, and Alexei A. Efros. "Unbiased look at dataset bias." *Computer Vision and Pattern Recognition* (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[7] Jegou, Herve, Hedi Harzallah, and Cordelia Schmid. "A contextual dissimilarity measure for accurate and efficient image search." *Computer Vision and Pattern Recognition*, 2007. *CVPR'07. IEEE Conference on*. IEEE, 2007.

[8] Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos." *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

[9] Kogler, Marian, and Mathias Lux. "Bag of visual words revisited: an exploratory study on robust image retrieval exploiting fuzzy codebooks." *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. ACM, 2010.

[10] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." *Computer Vision–ECCV 2006*. Springer Berlin Heidelberg, 2006. 404-417.

[11] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.