Bag of visual words: A soft clustering based exposition

Vinay Garg, Sreekanth Vempati, C. V. Jawahar
Center for Visual Information Technology,
International Institute of Information Technology
Hyderabad, Andhra Pradesh, India, 500032
Email: vinay.garg@research.iiit.ac.in, v_sreekanth@research.iiit.ac.in, jawahar@iiit.ac.in

***Abstract*—In this paper, we explain the bag of words representation from a soft computing perspective. The traditional Bag of word representation describes an image as a bag of discrete visual codewords. Where histogram of the number of occurrences of these codewords is used for image classification tasks. The drawback of the approach is that every visual feature in an image is assigned to single codeword, which leads to the loss of information regarding the other relevant codewords that can represent the same feature. In this paper, we show how fuzzy and possibilistic codeword assignment improves the classification performance on Scene-15-dataset.**
keywords: codewords, fuzzy assignments, possibilistic assignments, image classification.

I. Introduction

In the recent past, advances in the storage, capturing device and Internet has led to the rapid growth in the number of digital image collections. Automatic classification of images based on the semantic category can be helpful in efficient search and management of these large collections of images. For example, a collection of photos needs to be categorized into semantic categories like "bedroom", "mountain", "night time", etc., to support efficient browsing and search. Recent research shows the success of Bag of Words representation for images in automatic classification and search tasks [1], [2], [3], [4].

Bag of visual word representation is inspired from the word-document representations of images. The first step of the bag of visual words based approach is the computation of local feature descriptors like SIFT [5] for a set of image patches. These patches can be either at the key-point locations or densely sampled on a regular grid of the image. These set of local feature descriptors are quantized using a clustering technique. This step is referred to as vocabulary building step. The resultant set of cluster centers is referred to as visual vocabulary or codebook, and each cluster center is individually called a "visual word" or "codeword". The generated visual vocabulary is then used for assigning the nearest visual word for each of the local feature descriptors in a given image. This step is referred to as assignment step. The histogram of visual words in a given image is used as its representation for image classification, retrieval or recognition tasks. The main drawback of the traditional bag of words approach is the hard assignment of the visual code words to the local image features [12]. As only a single visual word is assigned to a given feature descriptor, the relevance of the feature descriptor to other visual words is lost which leads to poor results in classification tasks. In order to overcome this problem, we use fuzzy set theoretic notions in the traditional bag of words approach. Fuzzy logic allows an object to belong to multiple classes with varying degrees of membership. This helps in modeling the ambiguity of assigning a visual codeword to a local feature descriptor. Introducing fuzziness, leads to better characterization of an image in terms of the distribution of visual words, which in turn helps in the better classification of the images.

In the section 2, we discuss the related work. We describe the traditional Bag of words approach, its shortcomings and fuzzy bag of words approach along with experimental results in section 3. Section 4 contains an alternative approach to Possibilistic BoW by eliminating fuzziness parameter, from which we will derive the equation for soft assignment which is used frequently [12] and finally conclude in section 5.

II. Related Work

Bag of Words representation, motivated from field of documents search, was first used for the problem of texture recognition [6]. It was then popularly used for content based image search and classification tasks [1], [2], [3], [4]. Spatial layout is lost by representing an image as histograms of visual words for image classification tasks. In order to capture spatial layout, Lazebnik *et. al* [1] proposed pyramid histogram of visual words.

Other improvements in bag of words approaches mainly focus on the vocabulary generation. Vogel *et. al* [11] present a semantic vocabulary for scene classification tasks where in each image patch is labeled with a semantic label like sky, water, grass, etc., Winn *et. al* [7] proposed universal code-book vocabulary for object recognition. Perronnin *et. al* [8] presented a class-specific vocabularies for generic visual recognition. Jurie and Triggs [9] compare different clustering techniques for generation of vocabulary. Lazebnik *et. al* [10] have proposed learning of quantization code-books by information loss minimization.

Philbin *et. al* [12] have presented the use of assignment of a single descriptor to multiple nearest visual words for the problem of particular object retrieval. Our work is in the similar direction of these works, which aim for obtaining

978-0-7695-4599-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NCVPRIPG.2011.15

better representation of the bag of words by taking care of the multiple relevant visual words.

III. Fuzzy Bag of Visual-Words

Traditional bag of words approach assigns a single visual word to each of the features descriptors in an image. This hard assignment gives rise to two issues: codeword uncertainty and codeword plausibility [14]. Codeword uncertainty refers to the problem of selecting the correct codeword out of two or more relevant candidates. The traditional bag of words approach merely selects the best representing codeword, ignoring the relevance of other candidates. Codeword plausibility denotes the problem of selecting a codeword without a suitable candidate in the vocabulary. Traditional bag of words approach assigns the best fitting codeword, regardless of the fact that this codeword is not a proper representative. Both these problems are illustrated in Fig. 1.

We now show how these problems can be handled by

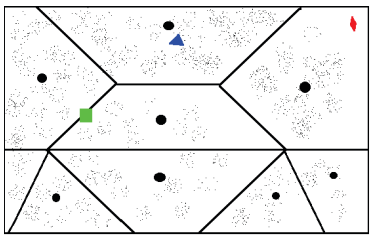


Fig. 1. An example showing the problems of codeword ambiguity. The small dots are image features, the circles are codewords found by hard clustering. Data sample that is well suited to the codebook approach is shown by blue triangle. Problem of codeword uncertainty is shown by the green square, and the problem of codeword plausibility by the red diamond.

introducing fuzziness in the vocabulary building and assignment steps. In general, hard clustering schemes like k-means clustering is used for vocabulary building. Given a set of N visual feature descriptors, k-means algorithm tries to find an optimal set S, having C cluster centers, which minimizes the following objective function:

$$J_{kmeans}(S) = \sum_{i=1}^{C}\sum_{j=1}^{N} ||x_j^{(i)} - c_i||^2 \quad (1)$$

where, x_j represents j^{th} feature, c_i represents the center of i^{th} cluster. The above equation assigns a feature descriptor to the single nearest cluster center without considering the other most nearest cluster centers.

In fuzzy vector quantization framework, instead of assigning each feature with a single codeword, we use an uncertainty term model [13] in which each feature is assigned to multiple codewords with some membership value, which represents its relevance to that codeword. This membership value can either be relative (as in *"Fuzzy C-Means (FCM)"*) or absolute (as in*"Possibilistic C-Means (PCM)"*) [13], [16].

A. Fuzzy/Probabilistic C-Means

In fuzzy c-means we use a membership function u_{ij} to modify the objective function as follows.

$$J_{fcm}(S) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^m ||x_j - c_i||^2 \quad (2)$$

subject to the condition

$$\sum_{i=1}^{C} u_{ij} = 1, \forall j \quad (3)$$

Here u_{ij} is the membership value of the j^{th} feature to the i^{th} codeword. The above equation is minimized by using iterative optimization using the following update equations

$$u_{ij} = \frac{d_{ij}^{\frac{-2}{m-1}}}{\sum_{l=1}^{c} d_{lj}^{\frac{-2}{m-1}}}, \quad (4)$$

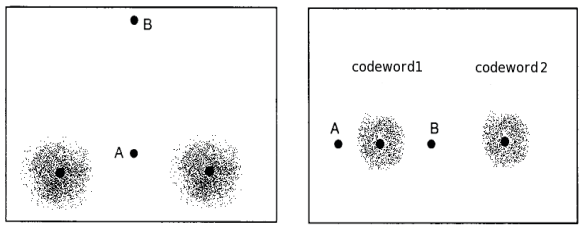
$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m} \quad (5)$$

Here u_{ij} is the membership value and d_{ij} is distance of the j^{th} feature to the i^{th} codeword respectively and m, $(m > 1)$ is called the fuzzifier or the weighting exponent whose value determines the amount of fuzziness that is introduced in the assignments.

Assigning the features in this manner encodes the relevance of a feature to a particular codeword depending upon its distance from the other codewords. Eq.(3) ensures that the sum of the membership degrees for each feature to all the codewords equals 1. This means that each feature receives the same weight in comparison to all other data and, therefore, that all data are (equally) included into the cluster partition. The membership values resemble the probability of a particular feature belonging to a particular codeword, since sum of the membership values of a particular feature for all clusters is 1.

Although this probabilistic fuzzy assignment solves the problem of codeword uncertainty and plausibility but probabilistic fuzzy membership values can be misleading when there is some noise or outliers. Consider, for example,the simple case of two codewords shown in Fig. 2(a) . Feature A has the same distance to both the codewords and thus it is assigned a membership degree of about 0.5. However, the same degree of membership are assigned to feature B even though this feature is further away from both the codewords and should be considered less typical. Because of the normalization (Eq.2) however, the sum of the membership values has to be 1. Consequently B receives fairly high membership values of 0.5 to both the codewords.

Also two features equidistant from a particular codeword, may be assigned with different membership values because the membership value not only depends on the distance of a feature from that particular codeword but also on its distance from other codewords. This is shown in Fig. 2(b), where features A and B, even though are equidistant from codeword



(a) (b)

Fig. 2. (a) Example of a dataset with two features A and B in which the (Probabilistic) fuzzy membership of these features in both the codewords are equal, even though feature B is much less representative of either codeword. (b) Example of a dataset with two codewords in which the membership generated by the Probabilistic assignments for features A and B are different, even though they are equidistant from codeword 1.

1 but will have different membership values. This problem arises from the constraint on the memberships, which forces feature B to give up some membership in codeword 1 in order to increase its membership in codeword 2.

B. Possibilistic C-Means

The problems caused by noise in the probabilistic assignments are mainly because of the normalization constraint (Eq.(3)). By dropping this constraint, we can achieve a more intuitive assignment of degrees of membership and avoid undesirable normalization effects. But dropping the normalization constraint can lead to the trivial solution where all u_{ij} are 0, which will lead to the minimization of the J_{fcm} [16]. This problem is overcome by adding another term in the objective function as follows

$$J_{pcm}(S) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m}||x_j - c_i||^2 + \sum_{i=1}^{C}\eta_i \sum_{j=1}^{N}(1-u_{ij})^m, \quad (6)$$

where $\eta_i > 0 (i = 1, \ldots, c)$. The first term leads to a minimization of the weighted distances, the second term suppresses the trivial solution since this sum rewards high membership (close to 1) that makes the expression $(1 - u_{ij})^m$ approximately 0. Thus the desire for the strong assignments of features to the clusters is expressed in the objective function J_{pcm}.
J_{pcm} is minimized by using the following update equation for membership values

$$u_{ij} = \frac{1}{1 + (\frac{d_{ij}^2}{\eta_i})^{\frac{1}{m-1}}} \quad (7)$$

The update equation for c_i remains the same as Eq. (5).

Here η_i is called the "bandwidth" or "resolution" or "scale" parameter. Considering the case m = 2 and substituting η_i for d_{ij}^2 yields u_{ij} = 0.5. It becomes obvious that η_i is a parameter that determines the distance to the cluster i at which the membership degree should be 0.5. The significance of this parameter can be seen like this, that it is distance beyond which a feature will not be of much relevance to a particular codeword. Its value can either be fixed for every codeword or can be estimated by the fuzzy intra-cluster distance using the fuzzy membership matrix.

A distinguishing characteristic of possibilistic assignment is that the membership values u_{ij} of feature j in codeword i is absolute as depends only on the distance of the feature from the codeword as compared to probabilistic assignments where they are relative.

C. Experimental Results

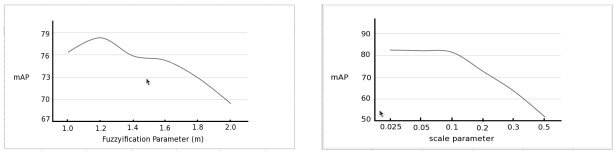
We present the scene classification results using the hard assignment as baseline results and compare them with classification results obtained using fuzzy framework (both fuzzy probabilistic and fuzzy possibilistic assignments). All the experiments are performed on Scene-15 dataset [15]. It consists of 4485 images spread over 15 categories like mountains, forests, kitchen, etc. Mean Average Precision (mAP), calculated using different techniques is used as the evaluation measure for our experiments.

Initially, we extract dense SIFT feature descriptors from each of the training images and cluster a subset of them to obtain a visual vocabulary. Then the histogram of the visual words is build for all the images with hard, fuzzy probabilistic and fuzzy possibilistic assignments, which is used to represent an image. We use 33 percent of images for testing and 67 percent of images for training from the available images of each category. Then a 1-vs-all SVM classifier is trained for all the 15 classes. The overall mAP is used as the evaluation measure in our experiment, which is calculated as the mean of the mAP of all the classes. Since clusters were initialized randomly, so to ensure fair experimental evaluation each experiment was performed 10 times with different initialization.

Initially we use, fuzzy approach both in the vocabulary construction and the assignment step but those results were not as good as compared to baseline results. The reason could be, that vocabulary built using the fuzzy probabilistic and possibilistic approaches were not that discriminative since the cluster centers were coming very close to each other. Therefore, we used vocabulary built using the hard k-means, but introduced fuzziness in the membership assignment step.

First three columns of Table I, shows the average of mAP calculated in 10 different experiments along with the variance, for different vocabulary sizes using histograms prepared by hard, fuzzy probabilistic and fuzzy possibilistic assignments. We can observe that both probabilistic assignment and possibilistic assignments give better results as compared to the baseline results of hard assignments. For a given vocabulary size, we choose the fuzzification parameter m (for probabilistic assignments) and scale parameter η (for possibilistic assignments) resulting in the best mAP by experimentation. $m = 1.2$ and $\eta = 0.05$ was giving the best results for our dataset.

It can be seen from graph Fig. 3(a) that in case of probabilistic assignments the classification performance decreases with the fuzzification parameter m (here vocabulary size = 1000). As the fuzzification parameter m increases all the clusters gain equal membership value of $\frac{1}{c}$ for any feature descriptor. This results in less discriminative representations of the bag



(a) (b)

Fig. 3. (a) Effect of fuzzification parameter m on mAP.(b) Effect of scale parameter η on mAP.

of words for images of all the classes, leading to decrease in the performance.

Similar behavior is seen in case of possibilistic assignments Fig. 3(b). The classification performance decreases as we increase the scale parameter η because as value of scale parameter approaches infinity, the membership value of each feature corresponding to every cluster approaches 1, which again results in less discriminative representations and aptly performance decreases.

Though fuzzy possibilistic assignments are performing better than the hard assignments, but are not as good as fuzzy probabilistic assignments. The possible reason could be that, the interpretation of m is different in the case of FCM and the PCM. In the FCM increasing values of m represent increased sharing of points among all the clusters, whereas in the PCM, increasing values of m represent increased possibility of all the points belonging to a given cluster. So it would be better to remove this m completely from the J_{pcm} [16], which will lead to better performance, as shown in the next section.

IV. Possibilistic C-Means Another Approach

The objective function for PCM described above is a particular implementation of the possibilistic approach which is dependent on fuzzifier parameter m. We could eliminate m altogether by choosing alternative formulations of the PCM described below.

$$J'_{pcm}(S) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}||x_j - c_i||^2 + \sum_{i=1}^{C}\eta_i \sum_{j=1}^{N}(u_{ij}\log_{ij} - u_{ij}) \tag{8}$$

The updating equations for the membership degrees can be derived from J'_{pcm} by setting its derivative to zeros as follows

$$\frac{\delta(J'_{pcm})}{\delta u_{ij}} = 0 \Rightarrow d_{ij}^2 + \eta_i \log u_{ij} + 1 - 1 = 0$$

$$\Rightarrow \log u_{ij} = \frac{-d_{ij}^2}{\eta_i} \Rightarrow u_{ij} = e^{\frac{-d_{ij}^2}{\eta_i}} \tag{9}$$

Eq. (9) defines the membership degrees of the j^{th} feature to the i^{th} cluster center which is similar to membership values of the "Descriptor-space soft assignment" used by Philbin *et. al* [12]. Experiments done using Eq. (9) as the membership assignment equation, outperforms all the techniques used above (even fuzzy assignments and possibilistic assignments described above). Table I shows the result where mAP of each techniques mentioned in section 3 is compared with the "soft assignment" (a particular type of possibilistic assignment).

TABLE I
Comparison of classification results using Hard, Probabilistic and Possibilistic Assignments.

Vocabulary Size	Hard Assignment	Fuzzy Assignment	Possibilistic Assignment I	Possibilistic Assignment II
200	71.21±0.20	75.07±0.13	73.29±0.15	78.38±0.13
500	75.32±0.19	76.85±0.21	75.92±0.10	80.27±0.12
1000	75.99±0.13	77.43±0.12	76.23±0.14	81.25±0.15
2000	76.02±0.11	78.98±0.16	76.89±0.16	82.46±0.19
4000	76.20±0.10	79.80±0.08	77.02±0.09	84.13±0.08

V. Conclusion

Due to the hard assignment of visual features to codewords there is a loss of relevance of other equally relevant codewords in traditional BoW. We have shown that fuzziness in the assignment step when used with the vocabulary built by hard k-means, can result in better "Bag of Words" representation for image classification tasks. The fuzzification parameter m and the scale parameter η are data dependent parameters, and their value needs to be selected experimentally for a given data. Experimental results on Scene-15 dataset demonstrate superiority of Fuzzy BoW over traditional BoW. Also, absolute fuzziness performs better as compared to relative fuzziness, in which absolute fuzziness leading to exponential membership values gives the best result.

References

[1] S. Lazebnik, C. Schmid, J. Ponce,: *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, CVPR, 2006
[2] A. Bosch, A. Zisserman, X. Munoz, *Scene classification using a hybrid generative/discriminative approach.*, IEEE Trans. Pattern Analysis and Machine Intelligence, 30, 2008
[3] R. Fergus, P. Perona, A. Zisserman, *Object class recognition by unsupervised scale-invariant learning* , CVPR, 2003
[4] J. Sivic, A. Zisserman, *Video google: A text retrieval approach to object matching in videos*, ICCV, 2003
[5] D. G. Lowe, *Distinctive image features from scale-invariant key points*, IJCV, 60, 2004
[6] T. leung, J. Malik, *Representing and recognizing the visual appearance of materials using three dimensional textons*, IJCV, 43, 2001
[7] J. Winn, A. Criminisi, T. Minka, *Object categorization by learned universal visual dictionary*, ICCV, 2005
[8] F. Perronnin, C.R. Dance, G. Csurka, M. Bressan, *Adapted vocabularies for generic visual categorization*, ECCV, 2006
[9] F. Jurie, B. Triggs, *Creating efficient codebooks for visual recognition*, ICCV, 2005
[10] S. Lazebnik, M. Ragninsky, *Supervised learning of quantizer codebooks by information loss minimization*, IEEE Trans. Pattern Analysis and Machine Intelligence, 30(7), 2008
[11] J. Vogel, B. Schiele, *Natural scene retrieval based on semantic modeling step*, CIVR, 2004.
[12] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, *Lost in quantization: Improving particular object retrieval in large scale image databases*, CVPR, 2008
[13] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981
[14] J. C. V. Germet, J.M. Geusebroek, C.J. Veenman, A.W.M. Smeulders, *Kernel codebooks for scene categorization*, ECCV, 2008.
[15] S.Lazebnik, C. Schmid, J. Ponce, *Beyond Bag of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*, CVPR, 2006.
[16] R. Krishnapuram and J. M. Keller, *The Possibilistic C-Means Algorithm: Insights and Recommendations*, IEEE Transactions on Fuzzy Systems, Vol. 4, No. 3, August 1996