Assessing the Distinctiveness and Representativeness of Visual Vocabularies

Leonardo Chang[1,2(✉)], Airel Pérez-Suárez[1], Máximo Rodríguez-Collada[1], José Hernández-Palancar[1], Miguel Arias-Estrada[2], and Luis Enrique Sucar[2]

[1] Advanced Technologies Application Center (CENATAV), 7A # 21406, Siboney, Playa, CP 12200 Havana, Cuba
{lchang,asuarez,mrodriguez,jpalancar}@cenatav.co.cu
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro # 1, Tonantzintla, CP 72840 Puebla, Mexico
{lchang,ariasmo,esucar}@ccc.inaoep.mx

Abstract. Bag of Visual Words is one of the most widely used approaches for representing images for object categorization; however, it has several drawbacks. In this paper, we propose three properties and their corresponding quantitative evaluation measures to assess the ability of a visual word to represent and discriminate an object class. Additionally, we also introduce two methods for ranking and filtering visual vocabularies and a soft weighting method for BoW image representation. Experiments conducted on the Caltech-101 dataset showed the improvement introduced by our proposals, which obtained the best classification results for the highest compression rates when compared with a state-of-the-art mutual information based method for feature selection.

Keywords: Bag of visual words · Visual vocabulary · Object categorization · Object recognition

1 Introduction

One of the most widely used approaches for representing images for object categorization is the Bag of Words (BoW) approach. BoW-based methods have recently obtained good results and they even attained the best results for several classes in the recent PASCAL VOC Challenge 2011 on object classification.

The key idea of BoW approaches is to discretize the entire space of local features (e.g., SIFT, SURF) extracted from a training set at interest points or densely sampled in the image. With this aim, clustering is performed over the set of features extracted from all the images of the training set in order to identify features that are visually equivalent. Each cluster is interpreted as a visual word, and all clusters form a so-called visual vocabulary. Later, in order to represent an image from the training set, each feature extracted from the image is assigned to a visual word of the visual vocabulary; from which a histogram of occurrences of each visual word in the image is obtained. When an unseen image arrives, it is represented using the visual vocabulary and then it is processed by the classifier.

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 331–338, 2015.
DOI: 10.1007/978-3-319-25751-8_40

One of the main limitations of the BoW approach is that the visual vocabulary is built using features that belong to both the object and the background. This implies that the noise extracted from the image background is also considered as part of the object class description. Also, in the BoW representation, every visual word is used and they contribute in the same way to the histogram of an image, regardless of their low representative and discriminative power. All these elements may limit the quality of further classification processes.

Several methods have been proposed in the literature to overcome the limitations of the BoW approach. These include recent works aimed to build more discriminative and representative visual vocabularies. Authors in [1–3] use the class labels of images in the vocabulary training stage in order to obtain a more discriminative vocabulary. Also, since with a typical hard assignment features that lie near Voronoi boundaries are not well-represented by the visual vocabulary, researchers have explored multiple assignments and soft weighting strategies to address this problem. E.g., [4,5] proposed methods for multiple assignment where a feature is matched to k nearest terms in the vocabulary and these terms are weighted by a scaling function such that the nearest terms obtain a higher value. The related work that is closest to ours is the recent work of Zhang *et al.* [6]. In their paper authors propose a supervised Mutual Information (MI) based feature selection method. Their algorithm uses MI between each dimension of the image descriptor and the image class label to compute the dimension importance. Finally, using the highest importance values, they reduce the image representation size. Their method achieve higher accuracy than feature compression methods such as Product Quantization [7] and BPBC [8].

In this paper, we propose three properties to assess the ability of a visual word to represent and discriminate an object class in the context of the BoW approach. We define three measures in order to quantitatively evaluate each of these properties. Besides, we propose two methods for ranking and filtering the visual vocabulary and a new soft weighting method for representing an image from this vocabulary. One of the ranking methods is based on a *tf.idf* weighting scheme while the other one as well as the soft weighting method are based on the above mentioned measures. Experiments conducted on the Caltech-101 [9] dataset showed the improvement introduced by our proposals, which obtained the best classification results for the highest compression rates when compared with a state-of-the-art mutual information based method.

The paper is organized as follows: Section 2 introduces the proposed properties and measures for the evaluation of the representativeness and distinctiveness of visual words, the methods for ranking and filtering the visual vocabulary as well as the soft weighting method for image representation. The performance of our proposed methods on the Caltech-101 dataset and a discussion of the obtained results are presented in Section 3. Finally, Section 4 concludes the paper with a summary of our findings and a discussion of future work.

2 Proposed Method

In this section we propose three properties and their corresponding quantitative evaluation measures to assess the ability of a visual word to represent and discriminate an object class. Besides, we propose two ranking and filtering methods, one based on the above mentioned measures and the other one, based on a *tf.idf* weighting scheme. Finally, we propose a new soft weighting method for image representation which is also based on the proposed measures.

2.1 Inter-class Representativeness Measure

A visual word could be comprised of features from different object classes, representing visual concepts or parts of objects common to those different classes. These common parts or concepts do not have necessarily to be equally represented inside a visual word because, even when similar, object classes should also have attributes that differentiate them. Therefore, we can say that, in order to represent an object class the best, a property that a visual word must satisfy is to have a high representativeness of this class. In order to measure the representativeness of a class c_j in visual word k, the measure $\mathcal{M}_1$ is proposed:

$$\mathcal{M}_1(k, c_j) = \frac{f_{k,c_j}}{n_k}, \tag{1}$$

where f_{k,c_j} represents the number of features of class c_j in visual word k and n_k is the total number of features in visual word k.

2.2 Intra-class Representativeness Measure

A visual word could be comprised of features from different objects, many of them probably belonging to the same object class. Even when different, object instances from the same class should share several visual concepts. Taking this into account, we can state that a visual word best describes a specific object class while more balanced are the features from that object class comprising the visual word, with respect to the number of different training objects belonging to that class. Therefore, we could say that, in order to represent an object class the best, a property that a visual word must satisfy is to have a high generalization or intra-class representativeness over this class.

To measure the intra-class representativeness of a visual word k for a given object category c_j, the measure μ is proposed:

$$\mu(k, c_j) = \frac{1}{O_{c_j}} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{f_{k,c_j}} - \frac{1}{O_{c_j}} \right|, \tag{2}$$

where O_{c_j} is the number of objects (images) of class c_j in the training set, o_{m,k,c_j} is the number of features extracted from object m of class c_j in visual word k, and f_{k,c_j} is the number of features of class c_j that belong to visual word k.

The term $1/O_{c_j}$ represents the ideal ratio of features of class c_j, that guarantees the best balance, i.e., the case where each object of class c_j is equally represented in visual word k.

The measure μ evaluates how much a given class deviates from its ideal value of intra-class variability balance. In order to make this value comparable with other classes and visual words, μ could be normalized using its maximum possible value, which is $\frac{2 \cdot O_{c_j} - 2}{O_{c_j}^2}$. Taking into account that μ takes its maximum value in the worst case of intra-class representativeness, the measure $\mathcal{M}_2$ is defined to take its maximum value in the case of ideal intra-class variability balance and to be normalized by $\max(\mu(k, c_j))$:

$$\mathcal{M}_2(k, c_j) = 1 - \frac{O_{c_j}}{2 \cdot (O_{c_j} - 1)} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{s_{k,c_j}} - \frac{1}{O_{c_j}} \right|. \tag{3}$$

2.3 Inter-class Distinctiveness Measure

$\mathcal{M}_1$ and $\mathcal{M}_2$ provide, under different perspectives, a quantitative evaluation of the ability of a visual word to describe a given class. However, we should not build a vocabulary just by selecting those visual words that best represent each object class, because this fact does not directly imply that the more representative words will be able to differentiate well one class from another, as a visual vocabulary is expected to do. Therefore, we can state that, in order to be used as part of a visual vocabulary, a desired property of a visual word is that it should have high values of $\mathcal{M}_1(k, c_j)$ and $\mathcal{M}_2(k, c_j)$ (represents well the object class), while having low values of $\mathcal{M}_1(k, \{c_j\}^C)$ and $\mathcal{M}_2(k, \{c_j\}^C)$ (misrepresents the rest of the classes), i.e., it must have high discriminative power.

In order to quantify the distinctiveness of a visual word for a given class, the measure $\mathcal{M}_3$ is proposed. $\mathcal{M}_3$ expresses how much the object class that is best represented by visual word k is separated from the other classes in the $\mathcal{M}_1$ and $\mathcal{M}_2$ rankings.

Let $\Theta_{\mathcal{M}}(K, c_j)$ be the set of values of a given measure $\mathcal{M}$ for the set of visual words $K = \{k_1, k_2, ..., k_N\}$ and the object class c_j, sorted in descending order of the value of $\mathcal{M}$. Let $\Phi(k, c_j)$ be the position of visual word $k \in K$ in $\Theta_{\mathcal{M}}(K, c_j)$. Let $P_k = \min_{c_j \in C}(\Phi(k, c_j))$ be the best position of visual word k in the set of all object classes $C = \{c_1, c_2, ..., c_Q\}$. Let $c_k = \arg\min_{c_j \in C}(\Phi(k, c_j))$ be the object class where k has position P_k. Then, the inter-class distinctiveness (measure $\mathcal{M}_3$), of a given visual word k for a given measure $\mathcal{M}$, is defined as:

$$\mathcal{M}_3(k, \mathcal{M}) = \frac{1}{(|C| - 1)(|K| - 1)} \sum_{c_j \neq c_k} (\Phi(k, c_j) - P_k). \tag{4}$$

2.4 On Ranking and Reducing the Size of Visual Vocabularies

In this subsection, we present two methods for ranking and reducing the size of the visual vocabularies, towards more reliable and compact image representations. The first one, named MMM, is based on the measures proposed in

Sections 2.1-2.3. Let $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$ be the rankings of vocabulary K, using measures $\mathcal{M}_3(K, \mathcal{M}_1)$ and $\mathcal{M}_3(K, \mathcal{M}_2)$, respectively. $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$ provide a ranking of the vocabulary based on the distinctiveness of visual words according to inter-class and intra-class variability, respectively.

In order to find a consensus, $\Theta(K)$, between both rankings $\Theta^{\mathcal{M}_1}(K)$ and $\Theta^{\mathcal{M}_2}(K)$, a consensus-based voting method can be used; in our case, we decided to use the Borda Count algorithm [10] although any other can be used as well. The Borda Count algorithm obtains a final ranking from multiple rankings over the same set. Given $|K|$ visual words, a visual word receive $|K|$ points for a first preference, $|K| - 1$ points for a second preference, $|K| - 2$ for a third, and so on for each ranking independently. Later, individual values for each visual word are added and a final ranking obtained. From this final ranking a reduced vocabulary can be obtained by selecting the first N visual words.

The second ranking and filtering method we propose is based on a *tf.idf* weighting scheme, named FRM (Frequency-based Ranking Method). Our proposal is based on the definitions introduced in [11]. Traditionally, *tf.idf* has been used as a weighting scheme for image representation. However, in our proposal we use *tf.idf* for ranking and/or filtering the set of visual words. Let $D = \{m_1, m_2, \cdots, m_N\}$ be the image training set from which the visual vocabulary has been built. According to [11], the *term frequency* and the *inverse document frequency* of a visual word v_i in an image m_j, denoted by tf_{v_i,m_j} and idf_{v_i,m_j}, respectively, are defined by the following expressions:

$$tf_{v_i,m_j} = \frac{K_1 \cdot O_{ij}}{O_{ij} + K_2 \cdot \left(1 - b + b \cdot \left(\frac{|\{v_q | O_{qj} > 0\}|}{V_{avg}}\right)\right)} \quad idf_{v_i,m_j} = \log \frac{|D| - |D_{v_i}| + 0.5}{|D_{v_i}| + 0.5}, \tag{5}$$

where K_1, K_2 and b are constants, O_{ij} is the occurrence of v_i in m_j, V_{avg} is the average number of visual words representing the images of the training set, and D_{v_i} is the set of images in which the occurrence of v_i is greater than zero.

Taking into account the way in which the histogram of occurrences of the visual words is built for each image, it will be highly probable that any visual word *occurs* in almost all images. This will have a negative influence in the computation of the *idf* expression. For solving this issue, we propose to redefine D_{v_i} as the set of images in which the occurrence of v_i is greater than the average occurrence of v_i in all the images of the training set.

Using the *tf* expression and the new definition of *idf*, we can build for each visual word v_i, a vector containing the product of tf_{v_i,m_j} and idf_{v_i,m_j}, in each image $m_j \in D$. The average of the values contained in this vector, will constitute the ranking of v_i. From this ranking a reduced vocabulary can be obtained by selecting the first N visual words.

2.5 Soft Weighting for Image Representation

Once the visual vocabulary is built, the images are represented through a histogram of the occurrences of the visual words. For building this histogram, the distinctiveness and representativeness of visual words are not taken into account

during the histogram construction. Following, we propose a new soft weighting method for image representation, named SWIR, that tackles the negative effect that the above mentioned problem have over the histogram of occurrences.

Let $\Theta(N)$ be the final raking values of the N selected visual words that constitute the visual vocabulary, obtained using one of the ranking and filtering methods proposed in section 2.4. Let assume these values are normalized such that they are in [0,1]. We will use these values for supporting the presence of the highly discriminative visual words as well as for penalizing the presence of those with low descriptive power. With this aim, we first compute a *pivot* element, denoted as $P_{\Theta(N)}$, as the average raking value of the N visual words. The *contribution weight* of a visual word v_i, denoted as cw_{v_i}, is computed as follows:

$$cw_{v_i} = 1 - P_{\Theta(N)} + \Theta_{v_i}, \tag{6}$$

where Θ_{v_i} is the ranking value of v_i in $\Theta(N)$.

For obtaining the representation of an image m_j, we propose to multiply the occurrences of the visual words in the histogram of m_j by their respective contribution weights. Those visual words having a ranking over $P_{\Theta(N)}$ are considered as more representative and discriminative, and consequently, their presence in the histogram is rewarded (i.e., it is increased). On the other hand, visual words with rankings under the pivot are penalized by reducing their presence in the histogram of occurrences.

3 Experimental Results

The main goal of the experiments we present in this section is to quantitatively evaluate the improvement introduced by our proposals to the BoW-based image representation and to compare with the MI-based method proposed in [6], which obtains the best classification results among the feature selection and compression methods of image representation for object categorization. The experiments were conducted on the well-known Caltech-101 [9] dataset. All the experiments were done on a single thread of a 3.4 GHz Intel i7 processor and 8GB RAM PC.

In the experiments presented here, we used for image representation a BoW-based schema with PHOW features (dense multi-scale SIFT descriptors) [12] and spatial histograms as image descriptors. Elkan's K-means [13] with four different K values (K= 512, 1024, 2048 and 4096) was used to build the visual vocabularies; these vocabularies constitute the baseline. Later, each of the baseline vocabularies is ranked using the MI-based method proposed in [6] and our two proposed ranking methods, MMM and FRM, with and without the new representation method, SWIR. Based on these rankings, nine new vocabularies are obtained by filtering each baseline vocabulary, leaving the 10%, 20%, ..., 90%, respectively. We tested the obtained visual vocabularies in a classification task, using a homogeneous kernel map to transform a χ^2 Support Vector Machine (SVM) into a linear one [14]. We follow the experimental setup of [15], namely, we train on 30 images per class and test on the rest, limiting the number of test images to 50 per class. The classification accuracy results are reported in Fig. 1.

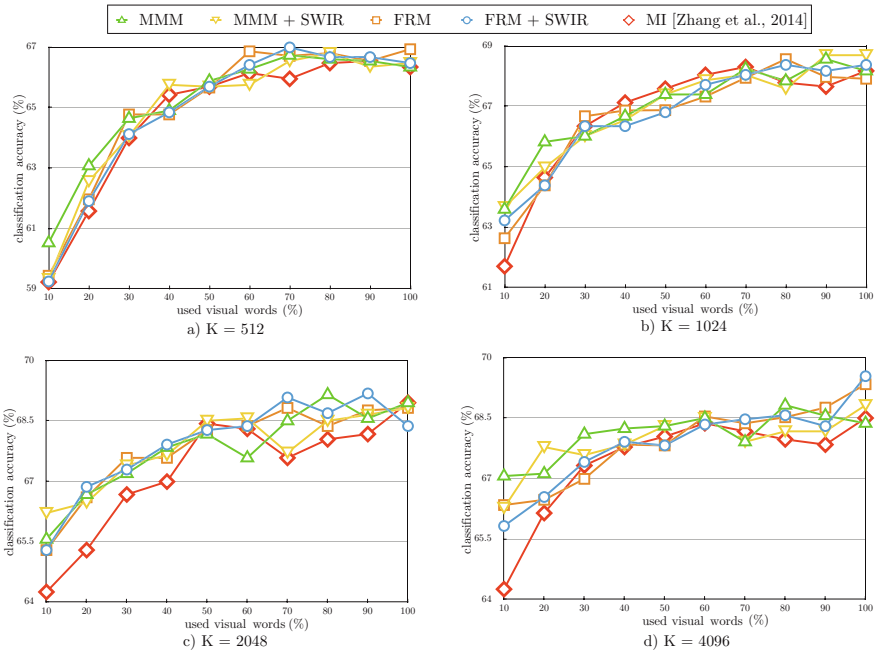

Fig. 1. Mean classification accuracy results on the Caltech-101 dataset.

As it can be seen in Fig. 1, for each value of K used in the experiment, our proposals obtain the best classification accuracy results for the highest compression rates (10, 20%), being MMM the best method. Moreover, for the other filtering sizes our proposals attain comparable and even better results than the MI-based method. Besides, in almost all values of K the combination of the MMM and FRM methods with SWIR method gets the highest classification accuracy results, being the combination between FRM and SWIR the best. Therefore, we can assert that taking into account the distinctiveness and representativeness of visual words in the image representation improves the accuracy of the classifier.

4 Conclusions

In this paper we have introduced three properties and their corresponding quantitative evaluation measures to assess the ability of a visual word to represent and discriminate an object class, in the context of the BoW approach. We also devised two methods for reducing the size of visual vocabularies that allow to obtain more distinctive and representative visual vocabularies for BoW image representation. Finally, we introduced a soft weighting method for image representation. The experiments conducted over the well-known Caltech-101 dataset showed that i) using the more discriminative and representative visual words, and ii) their properties quantitative measures it is possible to obtain more accurate

and compact visual vocabularies and improved BoW-based image representations. Future work will focus on defining a method that, based on the proposed measures, help us to automatically choose the filter size that maximizes the classification accuracy.

Acknowledgments. This work was supported in part by CONACYT Project No. 215546. L. Chang was supported in part by CONACYT scholarship No. 240251.

References

1. Kesorn, K., Poslad, S.: An enhanced bag-of-visual word vector space model to represent visual content in athletics images. IEEE Transactions on Multimedia **14**(1), 211–222 (2012)
2. Lopez-Sastre, R., Tuytelaars, T., Acevedo-Rodriguez, F., Maldonado-Bascon, S.: Towards a more discriminative and semantic visual vocabulary. Computer Vision and Image Understanding **115**(3), 415–425 (2011)
3. Jiu, M., Wolf, C., Garcia, C., Baskurt, A.: Supervised learning and codebook optimization for bag of words models. Cognitive Computation **4**, 409–419 (2012)
4. Chang, L., Duarte, M.M., Sucar, L.E., Morales, E.F.: A bayesian approach for object classification based on clusters of sift local features. Expert Systems with Applications **39**, 1679–1686 (2012)
5. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: CVIR, pp. 494–501. ACM (2007)
6. Zhang, Y., Wu, J., Cai, J.: Compact representation for image classification: To choose or to compress? In: CVPR 2014 (2014)
7. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Pattern Analysis and Machine Intellingence **33**(1), 117–128 (2011)
8. Gong, Y., Kumar, S., Rowley, H.A., Lazebnik, S.: Learning binary codes for high-dimensional data using bilinear projections. In: CVPR 2013 (2013)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. **106**(1), 59–70 (2007)
10. Emerson, P.: The original borda count and partial voting. Social Choice and Welfare **40**(2), 353–358 (2013)
11. Moulin, C., Barat, C., Ducottet, C.: Fusion of tf.idf weighted bag of visual features for image classification. In: Qunot, G. (ed.) CBMI, pp. 1–6. IEEE (2010)
12. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: ICCV, vol. **23**(1), pp. 1–8 (2007)
13. Elkan, C.: Using the triangle inequality to accelerate k-means. In: Fawcett, T., Mishra, N., (eds.) ICML, pp. 147–153. AAAI Press (2003)
14. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. Pattern Analysis and Machine Intellingence **34**(3) (2011)
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)