Encoding Spatial Arrangement of Visual Words

Otávio A.B. Penatti, Eduardo Valle, and Ricardo da S. Torres*

Recod Lab, Institute of Computing, University of Campinas (Unicamp), Campinas, Brazil
penatti@ic.unicamp.br, mail@eduardovalle.com, rtorres@ic.unicamp.br

Abstract. This paper presents a new approach to encode spatial-relationship information of visual words in the well-known visual dictionary model. The current most popular approach to describe images based on visual words is by means of bags-of-words which do not encode any spatial information. We propose a graceful way to capture spatial-relationship information of visual words that encodes the spatial arrangement of every visual word in an image. Our experiments show the importance of the spatial information of visual words for image classification and show the gain in classification accuracy when using the new method. The proposed approach creates opportunities for further improvements in image description under the visual dictionary model.

Keywords: spatial-relationship, visual words, visual dictionaries.

1 Introduction

Automatically understanding the content of multimedia data has become very important since there is an exponential growth of multimedia information available recently. The scientific and industrial communities have reached many advances in this field in the latest years. A very popular and effective technique for multimedia information description is by using visual dictionaries [14], which are mainly used in tasks of scene and object categorization.

The main idea of using visual dictionaries is to consider that the visual patterns present in images are similar to textual words present in textual documents. Therefore, an image is composed by visual words as a textual document is composed by textual words.

The process to generate visual dictionaries takes several steps. To obtain the visual words of images, usually interest point detectors, like Hessian-Affine and Harris-Laplace [9] detectors are used; the detected points are described by descriptors like SIFT [8]; and the points in feature space are then clustered to create the visual words. The words thus obtained are more general than the low level descriptors, since the clustering step will tend to quantize the descriptor space into "similar looking" regions.

* Authors thank CNPq, Capes, and Fapesp (2009/10554-8, 2009/05951-8, 2009/18438-7) for the financial support.

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 240–247, 2011.

When the visual dictionary is created, an image can be described by their visual patterns (visual words). The most traditional image descriptor based on visual words is the *bag-of-words*. It is simply a histogram of the visual words in the image. Therefore, when using visual dictionaries we can still have only one feature vector per image, even capturing local information.

The use of visual dictionaries is very popular and new approaches for improving the use and generation of them constantly appear in the literature [1, 3, 12]. As the traditional bag-of-words descriptor does not encode spatial information of images, some works try to overcome this weakness [2, 5, 7]

This paper presents an approach to encode the spatial information of visual words into the feature vector. Our approach captures the spatial arrangement of every visual word in an image. Its basic model is at the same time very simple and easily adaptable, opening the opportunity for a whole family of methods to represent the spatial relationship of visual words.

The remainder of the paper is organized as follows: Section 2 shows the importance of spatial information of visual words for image description. Section 3 presents our approach to encode the spatial arrangement of visual words. Section 4 shows the experiments and results. Section 5 concludes the paper.

2 Spatial-Relationship Information in Visual Dictionaries

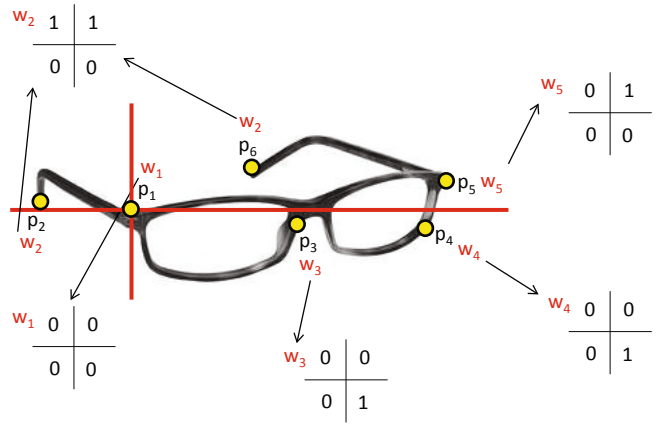
Spatial information of visual words is very important for the characterization of images and objects. Different objects and scenes may be composed by the same visual appearances in different spatial compositions, making that spatial distribution critical to their discrimination.

The traditional bag-of-words descriptor used to describe images based on visual words does not encode spatial information. The need to encode the spatial information of visual words has motivated the creation of some new approaches to tackle the problem. One of the most popular is the spatial pyramid [7] which splits the image into hierarchical cells and computes bags-of-words for each cell, concatenating the results at the end. Other approaches employ the co-occurrence of pairs of visual words [14] or correlograms of visual words [13]. The method presented in [2] proposes image splitting by linear and circular projections, generating one bag for each projection. Most of these approaches suffer from the problem of generating huge amounts of data.

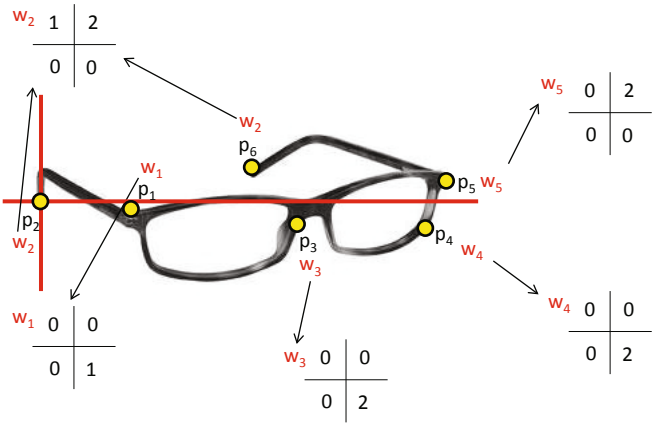
Although the spatial information of visual words is important for visual characterization, their frequency of occurrence, which is captured by the bag, is also very important, as observed in many applications [1, 6, 11]. Therefore, combining frequency of occurrence and spatial information of visual words should be a promising direction for further improvements.

3 Proposed Approach

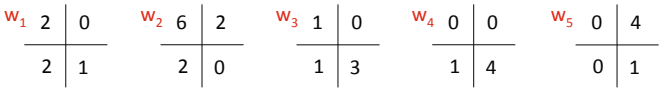
Our approach to encode spatial-relationship information of visual words is based on the idea of dividing the image space into quadrants [10] using each point as



(a)



(b)



(c)

Fig. 1. Example of partitioning and counting. The small circles are the detected points, tagged by their associated visual words (w_i's). We start in (a), putting the quadrant's origin in p_1 and counting in the visual word associated with each other point, where it is in relation to p_1. On the second step (b) the quadrant is at p_2; we add again the counters of the words associated with each other point in the position corresponding to their position in relation to p_2. We proceeded until the quadrant has visited every point in the image. Final counter values are shown in (c).

the origin of the quadrants and counting the number of words that appear in each quadrant. We count how many times a visual word w_i appears in each quadrant in relation to all other points in an specific image. This counting will tell us the *spatial arrangement* of the visual word w_i. Intuitively, the counting will measure the word's positioning in relation to the other points in the image. It reveals that a word w_i tends to be below, at right, or surrounded by other points, for example.

The image space is divided as follows: for each point p_i detected in the image, we divide the space into 4 quadrants, putting the point p_i in the quadrant's origin; then, for every other detected point p_j, we increment the counting of the visual word associated with p_j in the position that corresponds to the position of p_j in relation to p_i. For example, if w_j is the visual word associated with p_j and p_j is at top-left from p_i, the counter for top-left position of w_j is incremented. After all points are analyzed in relation to p_i, the quadrant's origin goes to the next point p_{i+1}, and the counting in relation to p_{i+1} begins. When all points have already been the quadrant's origin, the counting finishes. Figure 1 shows an example of partitioning the image space and counting.

Every word will be associated with 4 numbers. Those numbers tell the spatial arrangement of every visual word in the image. The same visual word can appear in several different locations in an image, however, there is only one set of 4 counters associated with it. The complexity of this method is $O(k^2)$, while the traditional bag is $O(k)$, where k is the dictionary size.

When the counting is finished, each 4-tuple is normalized by its sum. If the word w_i has non-zero values only in its bottom-right counter, for instance, we can say that w_i is a bottom-right word, that is, it appears always at bottom-right position in relation to other points. If w_i has top-left and top-right counters with high values, we can say that w_i is a word that usually appears above other points. If all counters of w_i are equally distributed, w_i is surrounded by other points (middle-word) or it is a word that repeatedly surrounds other points (border-word).

Another advantage of our method is that we do not need to tune parameters for better performance, as no parametrization is necessary.

4 Experiments

The experiments were conducted on the challenging Caltech-256 database [4], including the *clutter* class (257). The visual dictionary was generated using some of the most common parameters in the literature [1]: Hessian-Affine detector, SIFT descriptor, and 1000 aleatory centers. The visual words were hard assigned to the detected points [1]. The training and classification was performed by SVM with RBF kernel.

We compared our method with the traditional bag-of-words descriptor (BoW), which has only the frequency of occurrence of the words in the image. Our method is here called as WSA (words spatial arrangement). In our method, the feature vector also contains the frequency of occurrence of the words in the

image, like BoW. Therefore, the feature vector of WSA is composed by 5 values per visual word. We also compared BoW to a variation of WSA that does not contain the word frequency of occurrence (WSA-noBag).

The validation was performed by increasing the number of training samples per class. The training samples were randomly selected. All samples that were not in the training set were used in the testing set. Each experiment was repeated 10 times (varying randomly the training set). Figure 2 summarizes the results, showing the average accuracies obtained.

The curves show that WSA is superior to BoW in classification accuracy. This superiority is clear from training sets larger than 5 samples per class. The larger the training set, the larger the difference in favor of WSA. This indicates that the spatial arrangement of visual words aggregates important information to distinguish images and object categories. The results for WSA-noBag are below BoW showing that the frequency of occurrence of visual words is a little more important than only their spatial arrangement. However, the spatial arrangement is almost as discriminant as the frequency of occurrence of a visual word, demonstrating the importance of encoding spatial information of visual words. The superior performance of WSA indicates that combining frequency of occurrence and spatial arrangement of visual words is effective.

To better understand how the spatial information affects recognition results, we have performed a detailed (per class) analysis of classification accuracy considering a training set size of 30 samples per class. Table 4 shows the results obtained for the classes where the differences between BoW and WSA is large (greater than or equal to 0.1). Comparing BoW and WSA, we notice how promising is the use of spatial information together with frequency of occurrence information. WSA is superior in most of the classes and, in some of them, the spatial arrangement makes a large difference (more than 0.1 in absolute improvement of classification rate).

It is worth noting that for a few classes the spatial information was so important that even WSA-noBag (without frequency information) had performances remarkably superior to BoW. It was the case, for example, of classes 15 (bonsai), 25 (cactus), 44 (comet), 137 (mars), 156 (paper-shredder), 234 (tweezer), and 252 (car-side). This shows, in itself, the discriminating power of words spatial configurations.

For a few classes, interestingly, adding spatial information actually perturbs the classification. Those classes were few enough to be enumerated: 3 (backpack), 20 (brain), 24 (butterfly), 26 (cake), 103 (hibiscus), 129 (leopards), 142 (microwave), 241 (waterfall) and 250 (zebra). We are still investigating this phenomenon, but we believe that in some situations of very stereotyped textures (waterfalls, leopards, zebras, butterflies) with lots of detected points, the spatial configuration might confuse the descriptor.

In general, the spatial arrangement of visual words aggregates important discriminant information to the traditional bag-of-words.

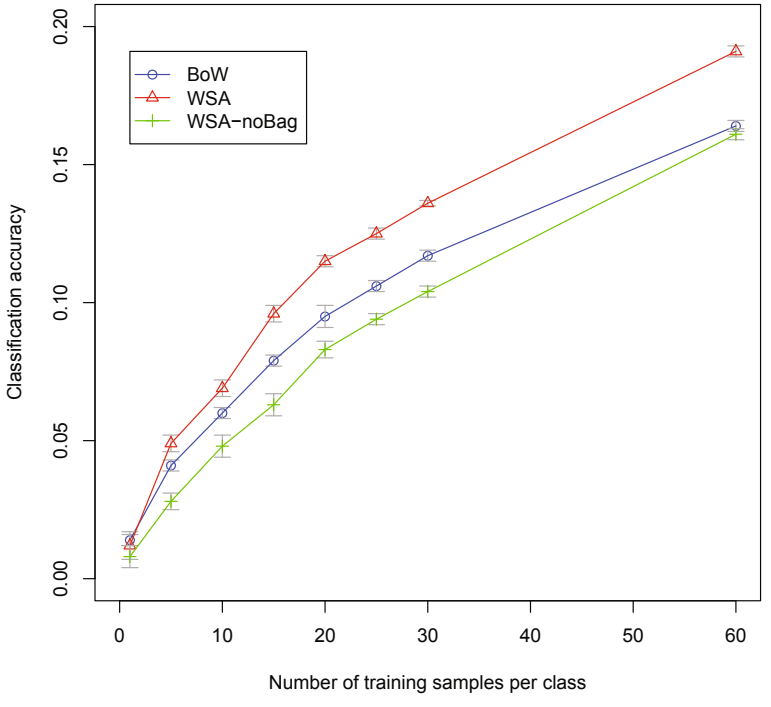


Fig. 2. Overall classification accuracy of the methods in Caltech-256. Each data point is the average for 10 runs, and the error bars are confidence intervals for alpha=0.05.

5 Discussion

This paper presents a simple and effective approach to encode spatial-relationship information of visual words. Our approach is based on the partition of the image space and in the counting of the occurrences of the visual words in relation to the other visual words positions. It is able to capture the spatial arrangement of every visual word in an image. Experiments show that aggregating the spatial arrangement of visual words to the traditional bag-of-words increases classification accuracy.

Our approach is also promising in the sense that the encoded information can be used in different ways. In this paper we directly use the spatial arrangement of words, however, more elaborated ideas can be applied over this spatial information. For example, the encoded information can categorize visual words spatially, like top-word, right-word, etc. The categorization can be used in many different ways, like, for instance, computing one bag for each category of visual word. We are already investigating the use of one bag for interior-words and other for border-words.

Table 1. Classes of Caltech-256 where the differences in accuracy between the different methods tested were large

Class	Class name	WSA-noBag	BoW	WSA
2	american-flag	0.13	0.12	**0.21**
3	backpack	0.07	**0.22**	0.08
15	bonsai	0.25	0.15	**0.25**
20	brain	0.15	**0.30**	0.14
21	breadmaker	0.04	0.06	**0.33**
24	butterfly	0.19	**0.39**	0.20
25	cactus	**0.27**	0.09	0.18
26	cake	0.04	**0.23**	0.07
44	comet	**0.53**	0.33	0.48
53	desk-globe	0.13	0.10	**0.26**
67	eyeglasses	0.40	0.34	**0.48**
75	floppy-disk	0.18	0.13	**0.36**
100	hawksbill	0.24	0.16	**0.30**
103	hibiscus	0.20	**0.42**	0.22
112	human-skeleton	0.09	0.05	**0.16**
123	ketch	0.22	0.15	**0.34**
127	laptop	0.14	0.10	**0.22**
129	leopards	0.62	**0.87**	0.60
137	mars	0.56	0.46	**0.61**
142	microwave	0.07	**0.23**	0.08
146	mountain-bike	0.29	0.24	**0.40**
156	paper-shredder	**0.25**	0.06	0.22
177	saturn	0.47	0.52	**0.62**
182	self-p.lawn-mower	0.27	0.26	**0.45**
234	tweezer	**0.86**	0.38	0.54
238	video-projector	0.06	0.13	**0.25**
241	waterfall	0.08	**0.38**	0.17
248	yarmulke	0.04	0.15	**0.26**
250	zebra	0.09	**0.25**	0.15
251	airplanes	0.34	0.30	**0.57**
252	car-side	0.50	0.38	**0.51**

Other improvements in the encoding of the spatial arrangement are also under investigation. A prior investigation is being made in the following scenario. We have the same object in different locations in two different images with clutter background. As the current counting schema considers all points in the image, in this case, the counting will change considerably from one image to another. To avoid this, we are investigating the use of windows around the point when counting. Other improvements are being tested, like a change in the partitioning schema. Instead of using 4 quadrants, we are trying to partition the space horizontally and vertically independently. This way of partitioning is more robust to rotation. Another possibility of use of our approach is for segmentation purposes, like, for instance, using the middle-words as seeds for some segmentation methods.

References

1. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR, pp. 2559–2566 (2010)
2. Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L.: Spatial-bag-of-features. In: CVPR, pp. 3352–3359 (2010)
3. van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. TPAMI 32(7), 1271–1283 (2010)
4. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
5. Hoíng, N.V., Gouet-Brunet, V., Rukoz, M., Manouvrier, M.: Embedding spatial information into image content description for scene retrieval. Pattern Recognition 43(9), 3013–3024 (2010)
6. Wenjun, L., Min, W.: Multimedia forensic hash based on visual words. In: ICIP, pp. 989–992 (2010)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal of Comp. Vis. 60(2), 91–110 (2004)
9. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. Journal of Comp. Vis. 60, 63–86 (2004)
10. Penatti, O.A.B., Torres, R.da.S.: Spatial relationship descriptor based on partitions. REIC 7(3) (2007) (in Portuguese)
11. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR (2008)
12. Jianzhao, Q., Yung, N.: Category-specific incremental visual codebook training for scene categorization. In: ICIP, pp. 1501–1504 (2010)
13. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: CVPR, vol. 2, pp. 2033–2040 (2006)
14. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: ICCV, vol. 1, pp. 370–377 (2005)