BRINGING ORDER IN THE BAG OF WORDS

Shihong Zhang, Rahat Khan, Damien Muselet and Alain Trémeau
Université de Lyon, F-42023, Saint-Étienne, France
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France

Keywords: Bag-of-words, Object Categorization, Spatial Information.

Abstract: This paper presents a method to infuse spatial information in the bag of words (BOW) framework for object categorization. The main idea is to account the local spatial distribution of the visual words. Rather than finding rigid local patterns, we consider the visual words in close spatial proximity as a pouch of words and we represent the image as a bag of word-pouches. For this purpose, sub-windows are extracted from the images and characterized by local bags of words. Then a clustering step is applied in the local bag of words space to construct the word-pouches. We show that this representation is complementary to the classical BOW. Thus a concatenation of these two representations is used as the final descriptor. Experiments are conducted on two very well known image datasets.

1 INTRODUCTION

In this paper, we deal with the problem of category-level classification in the images. This is a challenging problem in computer vision and one of the successful solutions is the Bag-of-Words (BOW) approach (Csurka et al., 2004), which employs the histogram of particular image patterns (the visual words) in a given image. However, one major limitation of the BOW model is that, it does not retain the spatial relationship among the visual words. Different methods have been proposed to take advantage of the spatial distribution of visual words to improve classification accuracy. For example, Lazebnik et al. employed the pyramid match kernel proposed by (Grauman and Darrell, 2005) into BOW framework to account the global distribution of the visual words among the image and achieved very high classification accuracy (Lazebnik et al., 2006). Among local approaches, Zhang et al. (Zhang and Mayo, 2008) improved the classification performance of the BOW model by discovering intermediate representations for each object class. Specifically, their approach includes the spatial relationships between all the frequent and informative image keypoints in the smaller regions of the image. A group of works intents to model the co-occurrence patterns of visual words. Among them, (Sivic et al., 2005) extended the BOW model using spatial information in their work. The spatial information, which they term as "doublets", is formed from spatially neighboring word pairs. In (Bhatti and Hanbury, 2010), Bhatti et al. introduced the pair-wise relations between image features. In their work, the image is represented by a concatenation of independent visual words with pair-wise visual words. Yuan et al. (Yuan et al., 2007) defined co-occurrence pattern occurring in local proximity as visual phrase and use this information for classification.

Most of the local approaches only consider pairs of visual words and we argue that we should not restrict the number of words accounted in the local neighborhoods. Unfortunately, increasing the number of words considered in each neighborhood tends to increase the dimension of the final descriptor. Hence, we propose an alternative that considers the visual words in close spatial proximity as a pouch of words and represents the image as a bag of word-pouches. The originality of this approach is that it applies a clustering step in the BOW space in order to extract the most representative pouches. Bag of word-pouches is also an orderless representation but interestingly it encodes some spatial information because each pouch is representative of a group of words which reside close to each other in the image space. Unlike the classical methods that introduce spatial information in the BOW, our approach accounts the spatial distribution of the visual words without increasing the dimension of the final descriptor. Furthermore our method, detailed in next Section, is complementary

Zhang S., Khan R., Muselet D. and Trémeau A..
BRINGING ORDER IN THE BAG OF WORDS.
DOI: 10.5220/0003859307230726
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 723-726
ISBN: 978-989-8565-03-7

to BOW representation and when concatenated with, provides superior classification accuracy.

2 BAG OF WORD-POUCHES

For image and object classification, the bag of words approach is the most widely used. The idea consists in characterizing each image by a histogram of quantized descriptors (Csurka et al., 2004), that are extracted from the descriptor space thanks to a clustering algorithm (e.g. k-means). Figure 1 illustrates the bag of words construction of 3 different images. In this example, we can see that we have 3 representative clusters in the descriptor space, each one associated with one visual word (square, circle and star). Then, each image is characterized by the histogram of these "visual words". This toy example underlines that no spatial information is accounted in the final representation. Indeed, whereas the spatial distributions of the visual words are different between the 3 images, their respective bag of words are the same.

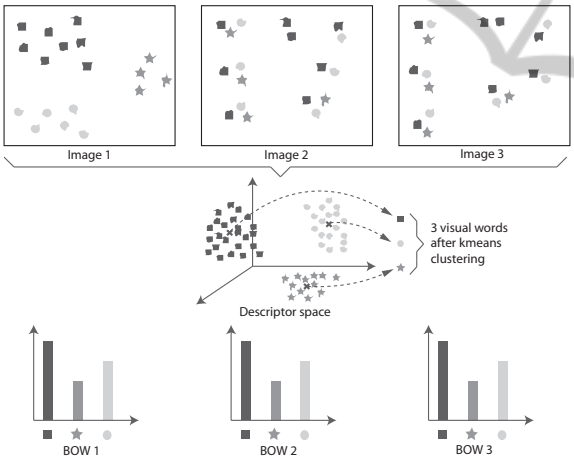

Figure 1: The bag of words (BOW) representation of 3 different images.

In order to infuse some spatial information in the bag of words approach, we propose to account the spatial distributions of the visual words in local areas of the images. The idea is to check if some sets of visual words often occur in the same neighborhood for some object classes and not for some others in order to add this discriminative information in the final representation of the images. Therefore, we propose to evaluate local bags of words from sub-windows extracted from the images and to characterize each image by a bag of *local bags of words*. Following the metaphor of the bag of words, our approach consists in bringing some order by adding some pouches to the bag, so that we put the words into these pouches (that are inside the bag) instead of mixing them in the bag. Consequently, the representatives of the most frequent *local bags of words* are called word-pouches.

The original part of this work is in the creation of the word-pouches. These word-pouches are the histograms of visual words that often occur in the same neighborhood. In order to define these word-pouches, we extract several sub-windows from all the images and for each of these sub-windows, we evaluate their bag of words (local BOW). Then, we create the local BOW space whose dimension is the number of words and in which each local BOW is one point. Finally, we apply clustering in the local BOW space and the cluster representatives are called the word-pouches. Once the word-pouches have been defined, for one image, we extract several sub-windows, evaluate the local BOW of each of these sub-windows and associate it with the nearest word-pouch. Hence, the image is characterized by its bag of word-pouches.

Figure 2 displays the image 2 (see figure 1) from which we have extracted 3 sub-windows. For each of these sub-windows, we have evaluated its BOW and shown the corresponding position in the local BOW space.

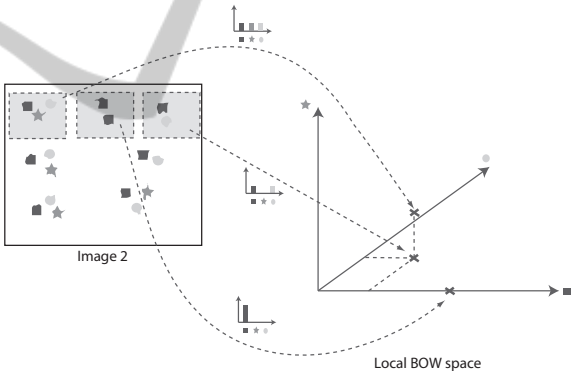

Figure 2: 3 sub-windows extracted from image 2 of figure 1, their respective local BOW and the corresponding points in the local BOW space.

Figure 3 shows the BOWP representations obtained for the 3 images of figure 1. In figure 3, we can see, in the local BOW space, the points associated with the sub-windows extracted from the 3 images. After clustering, we obtain 5 word-pouches that are the bins of the bags of word-pouches. We note that the BOWP is more representative of the image contents than the classical BOW of figure 1since the BOWP of the images 2 and 3 are similar to each other while being different from this of the image 1.

However, since the clustering step applied in the local BOW space tends to loose details about the visual words and since spatial information is more or less discriminative, depending on the considered cate-

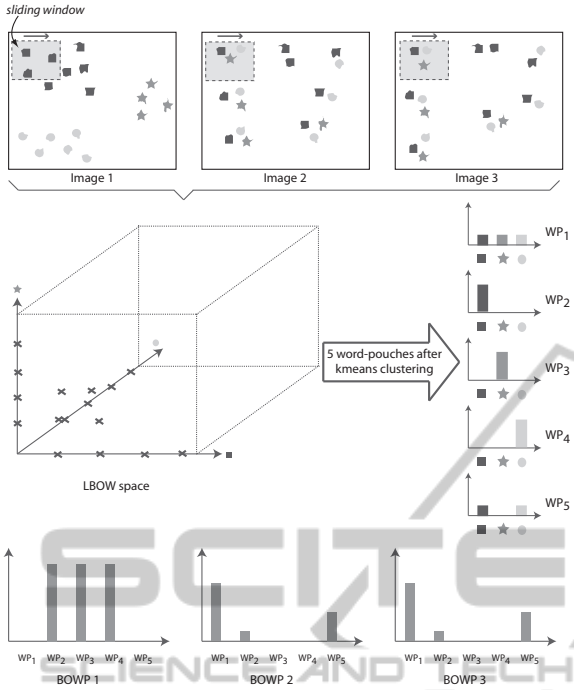

Figure 3: The bag of word-pouches (BOWP) representation of the 3 images from figure 1.

gory, we propose not to restrict the image description to the BOWP and to create a descriptor BOW&WP that is a concatenation of the BOW and the BOWP. This follows the intuition that rather than putting D words in a single bag without pouches, it is better to use less words ($D_w < D$) and put them in 2 bags, the first bag without pouches (BOW representation) and the second one with pouches (BOWP representation). The number of pouches in the second bag is equal to D_p and the relative numbers of words (D_w) and pouches (D_p) in the two bags are discussed in the next Section.

3 EXPERIMENTAL RESULTS

For the experiments, we have used two datasets. The first one is the Caltech101 image dataset (Fei-fei, 2004) from which we consider only the 10 most frequent categories (Caltech10). The second dataset is the Graz01 image dataset (Opelt et al., 2004) that contains two object categories namely bikes and persons and a background class. In this section, we present the average classification accuracy over 10 individual runs for both datasets.

16×16 patches are densely extracted, with 8 pixels overlap, and SIFT is used as descriptor (Lowe, 1999). We run a K-means clustering on a random subset of $50,000$ descriptors to construct the visual vocabulary. To create the bag of word-pouches, we empirically find that a square window size of 48×48 pixels with 8 pixels overlap works the best for the considered datasets. This size of window can accommodate 25 visual words given the parameter we have used in the dense sampling step. A SVM classifier with intersection kernel (Lazebnik et al., 2006) is used for all the experiments with the cost parameter (C) value set to 1.

For the first experiment, the used descriptor BOW&WP is constituted by D values and is a concatenation of a BOW and a BOWP. Then, we evaluate the average accuracy of this descriptor while varying the relative numbers of words (D_w) and word-pouches (D_p) within it, i.e. the k's in the two k-means algorithms. Since we want to compare descriptors with the same size, we choose D_w and D_p so that $D_w + D_p = D$. Hence, we show the result evolution while increasing D_p from 0 to $0.75 \times D$ ($D = 800$) in figure 4. The aim of this experiment is to define the best relative amounts of words and word-pouches for the considered dataset and for a constant-size descriptor. We can see that, whatever the considered dataset, accounting spatial information increases the average accuracy even if the number of words is decreasing. Since the best overall trade-off is obtained for equal numbers of words and word-pouches, for the rest of the experiments, we choose these ratios ($0.5 * D_w + 0.5 * D_p$) and we call this descriptor BOW&WP.

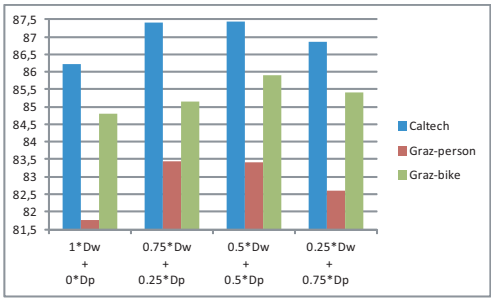

Figure 4: Average accuracy obtained with 800 dimensional descriptors on Caltech101 and Graz01 datasets.

Then, we propose to compare the results of our BOW&WP descriptor and the classical BOW for different values of D. We recall that the dimensions of BOW&WP and BOW are both equal to D. Figure 5 shows the results for D varying from 200 to 1000. We can see that for the 3 tested datasets and whatever the value of D the proposed BOW&WP outperforms the classical BOW.

Finally, we propose to analyze the relative average accuracy improvement (RAAI) of BOW&WP re-

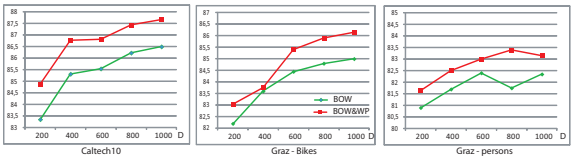

Figure 5: Average accuracies on Caltech10 and Graz01.

gards to classical BOW for each category of CALTECH10. The higher the value of RAAI the more the BOW&WP outperforms the BOW. Figure 6 shows the mean RAAI over all the values of *D* from 200 to 1000 for each category. The categories are ranked in the order of decreasing RAAI, so that BOW&WP performs better (compared to BOW) on the categories on the left (in green) than on the categories on the right (in red). This figure shows that, depending on the category, the improvement provided by the proposed BOW&WP varies a lot. It can be very high (more than 10%) for some categories such as chandeliers or bonsais that are characterized by rigid and stable structures and can be negative for some others such as leopards that are non-rigid objects whose poses and viewpoints are highly varying between the images.

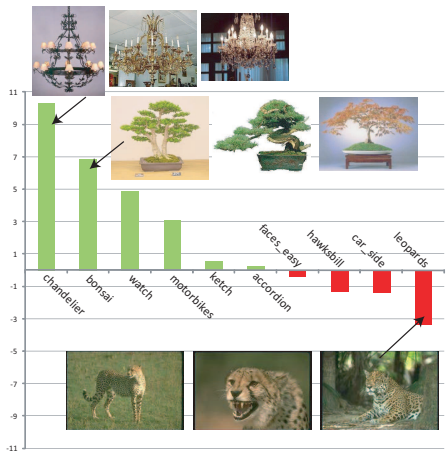

Figure 6: Relative average accuracy improvements of BOW&WP regards to classical BOW for the 10 most frequent CALTECH101 categories.

4 CONCLUSIONS AND FUTURE WORKS

In this paper, we have proposed an original and efficient way to account spatial distribution of the visual words in the image representations. The idea consists in accounting the way the visual words are locally organized. We have shown that adding this information in the bag of words can help in decreasing the number of visual words accounted for the construction of the descriptor while improving the average accuracies.

REFERENCES

Bhatti, N. A. and Hanbury, A. (2010). Co-occurrence bag of words for object recognition. In *Proceedings of the 15th Computer Vision Winter Workshop (CVWW)*.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*.

Fei-fei, L. (2004). Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision, CVPR*.

Grauman, K. and Darrell, T. (2005). The pyramid match kernel: discriminative classification with sets of image features. In *International Conference of Computer Vision*, pages 1458–1465.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*.

Opelt, A., Fussenegger, M., Pinz, A., and Auer, P. (2004). Weak hypotheses and boosting for generic object detection and recognition. In Pajdla, T. and Matas, J., editors, *ECCV (2)*, volume 3022 of *Lecture Notes in Computer Science*, pages 71–84. Springer.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering objects and their location in images. In *IEEE Intl. Conf. on Computer Vision*.

Yuan, J., Wu, Y., and Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhang, E. and Mayo, M. (2008). Pattern discovery for object categorization. In *23rd International Conference Image and Vision Computing New Zealand 2008(IVCNZ 2008)*.