VISUAL WORD PAIRS FOR AUTOMATIC IMAGE ANNOTATION

Nicolas HERVE, Nozha BOUJEMAA

nicolas.herve@inria.fr, nozha.boujemaa@inria.fr
INRIA Paris-Rocquencourt, IMEDIA project
Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France

ABSTRACT

The bag-of-visual-words is a popular representation for images that has proven to be quite effective for automatic annotation. In this paper, we extend this representation in order to include weak geometrical information by using visual word pairs. We show on a standard benchmark dataset that this new image representation improves significantly the performances of an automatic annotation system.

Index Terms— image annotation, visual vocabulary, bag-of-visual-words, co-occurence, visual word pairs, support vector machine (SVM)

1. INTRODUCTION

Automatically assigning labels to images is a very challenging task. This is especially true when dealing with generic databases where no *a priori* information is available on the visual content of the images. The algorithms used have to cope with huge variations both in the scenes and objects depicted as in the technical shooting conditions. When these labels describe globally the content or the context of an image, the annotation task is often referred to as scene categorization. It has been shown that effective approaches are now available for this problem that bring usefull results to end-users [1]. However, the fine labeling of the visual content or the detection of objects is still an open problem. State-of-the-art results are far from being satisfactory.

The bag-of-visual-words, very much inspired by the classical bag-of-words for text, is one of the most popular representation used for images. The main idea behind bag-of-visual-words is to represent images with orderless collection of visual patches and to compute an histogram counting the occurrences of these patches as a global signature. This representation can then be used in any learning framework to manage the automatic annotation problem. It is simple to implement and provides current state-of-the-art performances on several evaluation benchmarks.

One of the main characteristic of bag-of-visual-words is their orderless nature. The spatial position of the visual patches is dropped and never used. On one hand this choice brings flexibility and robustness to the representation as it is able to deal with changes in viewpoint or occlusion. On the other hand, the spatial relations between patches could be useful to describe the internal structure of objects or to highlight the importance of contextual visual information for these objects. Thus, for generic methods, the use of spatial information has to be considered in a slight way to avoid the building of too rigid models. Some work has already been done to use geometrical information in image annotation. Agarwal et al. [2] propose a two step approach. First, some object parts are detected in images, based on a previously generated vocabulary. Then, for the few detected parts, the spatial relations are described by quantizing their relative distances and orientations. The final image signature is a two-part feature vector containing parts occurrences on one side and quantized relations on the other side. In [3], Amores et al. propose the generalized correlogram descriptor that encompass both local and contextual information. They show that using simultaneously both types of information is efficient and faster.

In this paper we investigate the use of similar representation in the framework of automatic image annotation. Section 2 describes our standard annotation framework. In section 3 we introduce the word pairs. Finally, experimental results on the classical Pascal VOC 2007 dataset [4] are reported in section 4.

2. COMMON ANNOTATION FRAMEWORK

We can distinguish four main steps in an annotation framework based on bag-of-visual-words. First, visual patches are extracted from images and their visual signatures is computed with low-level descriptors. A vocabulary of visual words is then built. It will be used to quantize the visual patches. Global representations of images are obtained. Finally, some classifiers are trained and used to predict the desired visual concepts. Our purpose here is not to obtain the best possible scores in our experiments or to compete with state-of-the-art methods. We rather want to emphasize the benefits of visual word pairs and show how they contribute to the global results. Currently, approaches that give the best scores in bench-

The work presented in this paper was partially supported by the European Commission under contract FP6-045389 Vitalas.

mark evaluations often combine several methods for patches sampling, patches description, vocabulary creation and model learning. We prefer here to keep a simple approach with single choices at each step of the processing chain.

2.1. Patches extraction and description

As we need to describe locally the content of images, the first task is to extract visual patches. Several approaches have been proposed : image segmentation, point-of-interest detectors, regular grid and random sampling. Each one of them may also include a multi-scale step. It has been shown that dense sampling (grid or random) is more appropriate [5, 6]. In our experiments, we choose to extract square patches of 16x16 pixels, partially overlapping, on a regular grid at a unique scale. The average image size in the Pascal VOC 2007 dataset is around 500x350 pixels. We extract roughly 1000 patches per image.

We choose to describe these patches with structural descriptors including texture and shape informations. We set aside the color signal. We use common low-level descriptors that were usually computed on the full images to obtain global signatures and apply them to our small patches. Texture information is gathered by a 16-bins Fourier Histogram [7]. Shape information is gathered by a common 16-bins Edge Orientation Histogram [8]. The final signature is then a 32-bins vector, much smaller than the widely used 128-bins SIFT descriptor. The L1 distance is used to measure similarity between signatures.

2.2. Vocabulary and bag-of-visual-words representation

The bag-of-visual-words is largely inspired by the bag-of-words used to describe textual documents. One of the main differences is that a global vocabulary is already known for texts as it is composed of all the words encountered in a given language. These notions of word and vocabulary initially do not exist for images. The purpose of creating such a vocabulary is to provide a common basis for the representation of images in order to be able to compare them. This step is crucial as all further image representations depend on the quality of this visual vocabulary. A common method to obtain a vocabulary of visual words is to apply a clustering algorithm on a training dataset. The characteristics, and thus the suitability, of the generated vocabulary heavily relies on the characteristics of the chosen clustering algorithm.

K-means is a common clustering algorithm that is often used. It has been shown that k-means is suboptimal as it tends to promote dense regions of the visual space to produce clusters. Algorithms based on fixed radius clusters are reported to provide better results [5]. However, although we tested such approaches, we did not notice any specific improvements over k-means. We believe this is due, in our case, to the small size of the low-level descriptors used. Thus, we only report results obtained with k-means in the following. In our experiments, we use the Pascal VOC 2007 trainval set to build the vocabulary. We extracted approximately 4.8 million patches from its 5011 images.

The vocabulary size is one of the most important parameter. As it is not determined by the clustering algorithm, we report results with varying number of clusters. Another parameter to be determined when choosing a clustering algorithm is whether or not we want to use supervision. Clustering without any supervision will provide a generic vocabulary that could be used for any visual concept. On the other hand, using supervision will produce vocabularies that are more specific, and often more suited, to some given concepts. In [9], Perronnin used both generic and specific vocabularies. We will also report results using a similar approach.

Once a visual vocabulary is available, it is used to quantize the patches of an image. We choose hard assignment of each patch to its closest word. The global image representation is then a simple histogram counting the number of occurrences of each word of the vocabulary. Thus, the signature size is the size of the vocabulary. As the number of patches extracted from each image is slightly different, we normalize the histograms to have the sum of bins equal to one.

2.3. Learning strategy

We use soft margin support vector machines as classifiers with a triangular kernel. They haven proven to perform really well [1]. The triangular kernel has the great advantage of being non-parametric. We choose a one-against-all configuration and train one SVM per visual concept. As the dataset is unbalanced, we weight the training samples so as to have both positive and negative sets represent the same global weight. Moreover, as we noticed in previous experiment that the relaxation parameter of the SVM was almost always set to the same value, we choose to use a value of 1. This way, we don't have any optimization phase and the learning is really quick.

3. VISUAL WORD PAIRS FRAMEWORK FOR AUTOMATIC ANNOTATION

In natural images, the objects are almost always present in a scene they are related to. The contextual information is of great importance when trying to detect these objects [3, 1]. The global image representation provided by a bag-of-visual-words already encompass informations about the objects and their context. We targeted a more precise description of the relations between the objects in an image, and more specifically their relative location to each other. As in [2], the relations between patches can also encode internal structure of complex objects. We believe that this information has to be incorporated directly in the description and treated at the same level as the low-level signatures. This is done in [3], but we found the spatial encoding too restrictive and prefer

to use a less complex approach. We choose to consider the co-occurrence of words in a predefined local neighborhood of each patch. Thus, we only consider the distance between two patches, whatever their relative orientation. Sivic et al. [10] introduce the equivalent notion of *doublets* to refine object localization. Figure 1 shows the 12 pairs containing the central patch. The radius is a parameter of the algorithm.

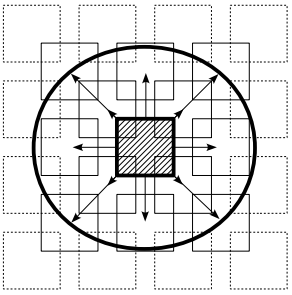

Fig. 1. Visual word pairs.

First results on segmentation obtained in [10] tends to indicate that this approach is pertinent to focus more on the objects. An easy way to implement this method is to create a new vocabulary containing all possible pairs of words from a base-vocabulary. For a base-vocabulary containing n words, the pairs-vocabulary will have $n(n+1)/2$ words. Then, obtaining a representation of an image is quite straightforward. After having quantized all the patches with the base-vocabulary, we simply accumulate the pairs in an histogram for which the distance between the two patches centers is below the given radius. Unlike Sivic [10], we keep all the words from our base-vocabulary to build the pairs. The signatures obtained are really sparse. In our experiments we choose the radius so as to keep only 1% of all the possible pairs in an image.

4. EXPERIMENTAL RESULTS

All the experiments have been conducted on the Pascal VOC 2007 dataset [4]. It has the advantage of being quite generic and freely available [1]. 20 visual concepts have been manually annotated. The set is split in a training and validation set, containing 5011 images, and a test set containing 4952 images. We keep these predefined sets. As for the official evaluation campaign, we use the Mean Average Precision measure (MAP) to evaluate the performances.

4.1. Common approach results

The dataset is heavily unbalanced. The number of images containing a concept is really different from one concept to another. In order to have a first idea of the object detection task difficulty, we compute the MAP for a random ranking of the test set. We obtained 0.0133. We also evaluate our low-level descriptors computed globally on the full images as well as our learning strategy. This is a good way of obtaining a baseline as these global descriptions will provide interesting clues on the importance of contextual information. We obtained 0.2271. We report in figure 2 the MAP for

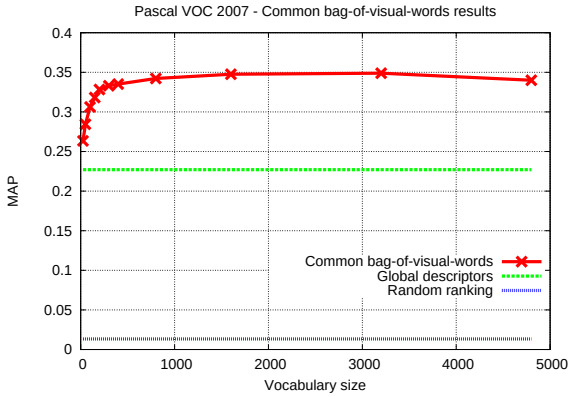

Fig. 2. Standard bag-of-visual-word results

different base-vocabulary sizes. We obtain a classical curve shape, growing fast for small and medium size vocabularies and reaching maximum of 0.3489 at 3200 words. Then it decreases slowly as the vocabulary tends to have too many words, which are too precise and fail to generalize. This can also be seen as an aspect of the curse of dimensionality as the number of words tends to reach the number of training samples. We also tested the influence of using classes labels in the creation of the vocabulary. Following the work in [9], we create bipartite histograms based on a global vocabulary combined with class-specific vocabularies of the same size. Results are reported in figure 3. As we can see, the results are slightly better than with a simple generic vocabulary.

4.2. Visual word pairs approach results

We report in figure 3 the MAP obtained with the word pairs vocabularies. The gain is obvious. The performances of pairs vocabularies start to decrease when the base-vocabulary has a size of 200, leading to word pairs histograms of size 20100. As previously mentioned, we face here the curse of dimensionality problem. We can notice that the maximum score obtained with a base-vocabulary of 150 words (MAP 0.3545) is already higher than the maximum reached in the standard configuration. We also tried to use generic and per-class vocabularies at the same time to produce word pairs vocabularies. Here it brings absolutely no improvement and we obtained the same results as with the generic base-vocabulary. So these results are not reported. We study the influence of the radius parameter on the performances. Results reported in table 1. We increase the radius so as to have the same amount

[1] http://www.pascal-network.org/challenges/VOC/voc2007

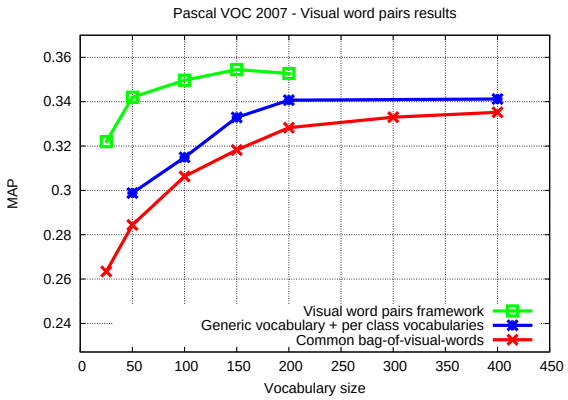

Fig. 3. Visual word pairs results

of pairs extracted per image (always 1%). The best results are obtained for the smallest radius.

Radius	MAP
$r \leq 1\%$	0.3220
$1\% < r \leq 2\%$	0.3176
$2\% < r \leq 3\%$	0.3065

Table 1. Influence of radius choice - 25 words vocabulary

The local co-occurrence of words included in our representation brings significant improvements. As the weak geometrical informations they represent are of a different nature than the simple words presence, we also choose to evaluate a combined representation. We mix the standard signatures obtained with a 1 600 words vocabulary and the pairs signatures for varying base-vocabulary sizes. We normalize the resulting histogram so that both have the same weight in the final representation. The best score with the combined representation (0.3839) is reached with a base-vocabulary of 100 words and is 10% higher than the best score of the standard representation (0.3489). We believe that some visual words never co-occur with the same neighbours and thus are not represented in our pairs representation. But their presence is still a valuable information and should not be discarded. In the same way, some words may not be significative alone, but often appear with the same neighbour. This information is encoded in our new representation. The two representations are complementary.

5. CONCLUSION

Embedding the word pairs in a standard bag-of-visual-words representation brings very significant improvement for an automatic annotation task. The weak geometrical information they encode is complementary to the standard words occurrences histogram. Moreover, as they are based on the same base-vocabulary, these representations are obtained with a low additional computational cost. Furthermore, we noticed that small vocabularies may be used with word pairs, thus also contributing to lower computational costs. Several improvements could be achieved : multi-scale schema, as well as a more refined definition of the radius that impact the words neighborhood. Other low-level descriptors could be considered to describe visual patches. Pairs of words separately described by different visual descriptors can also be built. The influence of patches belonging to objects or to the background in the word pairs can be studied.

6. REFERENCES

[1] N. Hervé and N. Boujemaa, "Image annotation : which approach for realistic databases ?," in *ACM International Conference on Image and Video Retrieval (CIVR'07)*, July 2007.

[2] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, 2004.

[3] J. Amores, N. Sebe, and P. Radeva, "Efficient object-class recognition by boosting contextual information," in *IbPRIA*, 2005.

[4] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," .

[5] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *IEEE International Conference on Computer Vision*, 2005.

[6] N. Hervé, N. Boujemaa, and M.Houle, "Document description : what works for images should also work for text ?," in *IS&T/SPIE Electronic Imaging, Multimedia Processing and Applications*, January 2009.

[7] M. Ferecatu, *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*, Ph.D. thesis, University of Versailles Saint-Quentin-En-Yvelines, 2005.

[8] A. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29(8), 1996.

[9] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *European Conference on Computer Vision*, 2006.

[10] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *IEEE International Conference on Computer Vision*, 2005.