Visual Features Selection

Giuseppe Amato, Fabrizio Falchi, and Cladio Gennaro

ISTI-CNR, via G. Moruzzi 1, 56124 Pisa, Italy
{giuseppe.amato,fabrizio.falchi,claudio.gennaro}@isti.cnr.it

Abstract. The state-of-the-art algorithms for large visual content recognition and content based similarity search today use the "Bag of Features" (BoF) or "Bag of Words" (BoW) approach. The idea, borrowed from text retrieval, enables the use of inverted files. A very well known issue with the BoF approach is that the query images, as well as the stored data, are described with thousands of words. This poses obvious efficiency problems when using inverted files to perform efficient image matching. In this paper, we propose and compare various techniques to reduce the number of words describing an image to improve efficiency.

Keywords: bag of features, bag of words, local features, content based image retrieval, landmark recognition

1 INTRODUCTION

During the last decade, the use of local features, as for instance SIFT [Lowe, 2004], has obtained an increasing appreciation for its good performance in tasks of image matching, object recognition, landmark recognition, and image classification. The total number of local features extracted from an image depends on its visual content and size. However, the average number of features extracted from an image is in the order of thousands. The BoF approach [Sivic and Zisserman, 2003] quantizes local features extracted from images representing them with the closest local feature chosen from a fixed visual vocabulary of local features (visual words). Matching of images represented with the BoF approach is performed with traditional text retrieval techniques.

However a query image is associated with thousands of visual words. Therefore, the search algorithm on inverted files has to access thousands of different posting lists. As mentioned in [Zhang et al., 2009], "a fundamental difference between an image query (e.g. 1500 visual terms) is largely ignored in existing index design. This difference makes the inverted list inappropriate to index images." From the very beginning [Sivic and Zisserman, 2003] some words reduction techniques were used (e.g. removing 10% of the more frequent images).

To improve efficiency, many different approaches have been considered including GIST descriptos [Douze et al., 2009], Fisher Kernel [Zhang et al., 2009] and Vector of Locally Aggregated Descriptors (VLAD) [Jégou et al., 2010]. However, their usage does not allow the use of traditional text search engine which has actually been another benefit of the BoF approach.

In order to mitigate the above problems, this paper proposes, discusses, and evaluates some methods to reduce the number of visual words assigned to images. This paper is a summary of a longer paper that will presented at VISAPP 2013 [Amato et al., 2013].

2 PROPOSED APPROACH

The goal of the BoF approach is to substitute each description of the region around an interest points (i.e., each local features) of the images with visual words obtained from a predefined vocabulary in order to apply traditional text retrieval techniques to content-based image retrieval. At the end of the process, each image is described as a set of visual words. The retrieval phase is then performed using text retrieval techniques considering a query image as disjunctive text-query. Typically, the *cosine* similarity measure in conjunction with a term weighting scheme is adopted for evaluating the similarity between any two images.

The proposed words reduction criteria are: *random*, *scale*, *tf*, *idf*, *tf*id*. Each proposed criterion is based on the definition of a score that allows us to assign each local feature or word, describing an image, an estimate of its importance. Thus, local features or words can be ordered and only the most important ones can be retained. The percentage of information to discard is configurable through a score threshold, allowing trade-off between efficiency and effectiveness. The *random* criterion was used as a baseline. It assigns random score to features. The *scale* criterion is based on the information about the size of the region from which the local features was extracted: the larger the region, the higher the score.

The retrieval engine used in the experiments is built as following:

1. For each image in the dataset the SIFT local features are extracted for the identified regions around interest points.
2. A vocabulary of words is selected among all the local features using the *k-means* algorithm.
3. The *Random* or *Scale* reduction techniques are performed (if requested).
4. Each image is described following the BoF approach, i.e., with the ID of the nearest word in the vocabulary to each local feature.
5. The *tf*, *idf*, or *tf*idf* reduction technique are performed (if requested).
6. Each image of the test set is used as a query for searching in the training set. The similarity measure adopted for comparing two images is the Cosine between the query vector and the image vectors corresponding to the set of words assigned to the images. The weight assigned to each word of the vectors are calculated using *tf*idf* measure.
7. In case the system is requested to identify the content of the image, the landmark of the most similar image in the dataset (which is labeled) is assigned to the query image.

3 Experimental results

The quality of the retrieved images is typically evaluated by means of precision and recall measures. As in many other papers, we combined these information by means of the mean Average Precision (mAP), which represents the area below the precision and recall curve.

For evaluating the performance of the various reduction techniques approaches, we use the Oxford Building datasets that was presented in [Philbin et al., 2007] and has been used in many other papers. The dataset consists of 5,062 images of 55 buildings in Oxford. The ground truth consists of 55 queries and related sets of results divided in best, correct, ambiguous and not relevant. The vocabulary used has one million words.

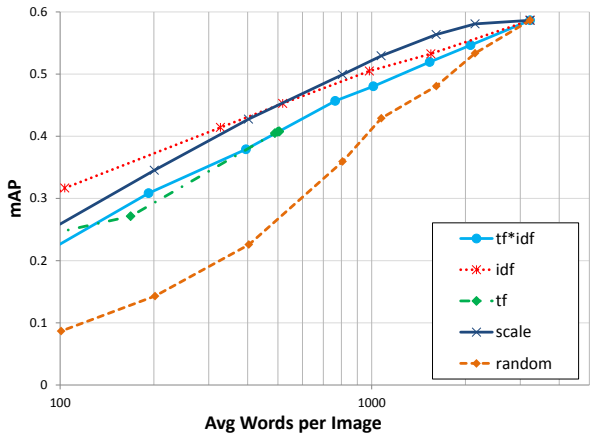

Fig. 1. Mean average precision of the various selection criteria obtained on the Oxford Buildings 5k dataset.

We first report the results obtained in a content based image retrieval scenario using the Oxford Building dataset using the ground truth given by the authors [Philbin et al., 2007]. In Figure 1 we report the mAP obtained On the x-axis we reported the average words per image obtained after the reduction. Note that the x-axis is logarithmic. We first note that all the reduction techniques significantly outperform naive *random* approach and that both the *idf* and *scale* approaches are able to achieve very good mAP results (about 0.5) while reducing the average number of words per image from 3,200 to 800. Thus, just taking the 25% of the most relevant words, we achieve the 80% of the effectiveness. The comparison between the *idf* and *scale* approaches reveals that *scale* is preferable for reduction up to 500 words per image. Thus, it seems very important to discard small regions of interest up to 500 words.

While the average number of words is useful to describe the length of the image description, it is actually the number of distinct words per image that have

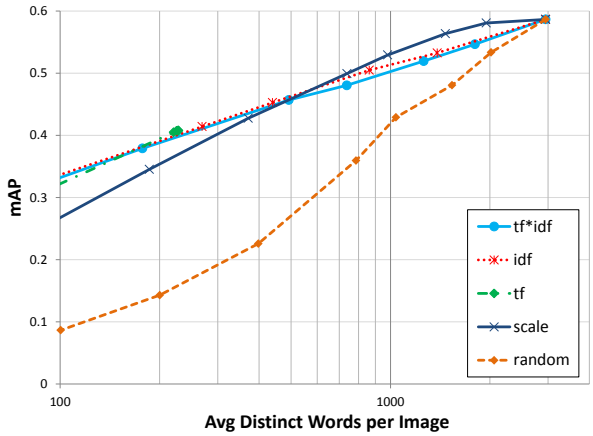

Fig. 2. Mean average precision of the various selection criteria obtained on the Oxford Buildings 5k dataset.

more impact on the efficiency of searching using inverted index. Thus, in Figure 2 we report mAP with respect to the average number of distinct words. In this case the results obtained by *tf*idf* and *tf* are very similar to the ones obtained by *idf*. In fact, considering *tf* in the reduction results in a smaller number of average distinct words per image for the same vales of average number of words.

References

[Amato et al., 2013] Amato, G., Falchi, F., and Gennaro, C. (2013). On reducing the number of visualwords in the bag-of-features representation. In *VISAPP 2013 - Proceedings of the International Conference on Computer Vision Theory and Applications*.

[Douze et al., 2009] Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., and Schmid, C. (2009). Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 19:1–19:8, New York, NY, USA. ACM.

[Jégou et al., 2010] Jégou, H., Douze, M., and Schmid, C. (2010). Improving bag-of-features for large scale image search. *Int. J. Comput. Vision*, 87:316–336.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

[Philbin et al., 2007] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–, Washington, DC, USA. IEEE Computer Society.

[Zhang et al., 2009] Zhang, X., Li, Z., Zhang, L., Ma, W.-Y., and Shum, H.-Y. (2009). Efficient indexing for large scale visual search. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1103 –1110.