Aggregating Local Descriptors for Epigraphs Recognition

Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti, Lucia Vadicamo

ISTI-CNR, via G. Moruzzi, 1 – 56124 Pisa (Italy)
Emails: <name>.<surname>@isti.cnr.it

Abstract. In this paper, we consider the task of recognizing epigraphs in images such as photos taken using mobile devices. Given a set of 17,155 photos related to 14,560 epigraphs, we used a k-NearestNeighbor approach in order to perform the recognition. The contribution of this work is in evaluating state-of-the-art visual object recognition techniques in this specific context. The experimental results conducted show that Vector of Locally Aggregated Descriptors obtained aggregating SIFT descriptors is the best choice for this task.

Keywords: Epigraphs Recognition, Object Recognition, Content-Base Image Retrieval, Bag-of-Features, VLAD.

1 Introduction

Because of large availability of digital cameras, especially embedded in smartphones and tablets, there is a growing demand of information retrieval systems able to search by using images as query. The basic idea is to allow the user to make a photo of the object she is interested on in order to receive related information. To his goal, the objects in the photo are compare with the ones stored in a repository in order to recognize it and give back relevant metadata, links, etc.

We conducted the work reported in this paper in the context of the Europeana network of Ancient Greek and Latin Epigraphy (EAGLE) CIP-Best Practice Network. In EAGLE, an object is essentially an ancient epigraph. In this work, we focus on searching for the most similar epigraphs with respect to the one represented in a photo made by the user. The dataset we used consists of 17,155 photos related to 14,560 epigraphs that were made available, within the EAGLE project, by Sapienza University of Rome. This functionality will be integrated on the flagship mobile application, to enable tourists to understand inscriptions they find on location. The application will allow a visitor of a site where one of the stored epigraphy is visible (museum, street, archaeological site, printed reproduction, etc.) to take a picture with a mobile phone, send the picture to the central repository and receive back the enriched information associated with that picture. Moreover, crowd-sourcing functionalities will be developed in the application to accelerate the adoption of the EAGLE and epigraphic content.

For achieving the task of identifying objects in an image or video sequence, usually referred to as object recognition, research conducted in both Computer Vision and Multimedia Information Retrieval fields has focused on local features that provide a

representation that allows matching local structures between images. First, distinctive key points are selected in each image. Second, a description of the selected regions is given. Direct local features matching has been proved to be very effective in recognizing the same objects in two photos. However, to achieve scalability (i.e., to be able to search in large datasets) aggregation techniques are necessary in order to summarize the information reported for each key point.

Traditionally, object recognition has been successfully applied to consumer products, buildings, monuments and landmarks. However, we did not find any specific experiments conducted on ancient epigraphs. Moreover, state-of-the-art techniques are very effective on small sets (tens) of objects while approximate techniques are applied when millions of objects have to be recognized. The number of photos of epigraphs expected in the context of the EAGLE project is in the middle of these two extremes. Thus, it is very interesting to understand what is the best technique for recognizing epigraphs. In the following, we report the results obtained testing various state-of-the art techniques in order to effectively recognizing epigraphs in photos given a medium-large scale set (tens of thousands) of known epigraphs.

The rest of this paper is structured as follows. In Section 2, we discuss related work. In Section 3, we give information about the tested approaches. Then, we specify the experimental settings and discuss the obtained results in Section 4. Finally, in Section 5, we report conclusions and discuss future work.

2 Related Work

In the last few years, research on object recognition has focused on local features [8], [11]. Following this approach, an image is represented by describing the visual content of typically thousands of regions of interest automatic selected. To achieve best effectiveness, images are compared by matching their local features and searching for a geometric transformation that can associate the regions of both images.

In the last few years, the problem of recognizing cultural heritage related objects, in particular landmarks, has received growing attention by the research community. As an example, Google presented its approach to building a web-scale landmark recognition engine [13]. The problem of landmark recognition is typically addressed by leveraging on techniques of automatic classification, as for instances kNN Classification [4], applied to image features.

Between 2007 and 2010, the VISITO Tuscany[1], (VIsual Support to Interactive TOurism in Tuscany) project, has focused on technologies able to offer an interactive and customized advanced tour guide service to visit the cities of art in Tuscany. This project has investigated cultural heritage object recognition, (such as monuments, landmarks, etc.) developing a mobile application and related research papers such as [2]. However, epigraphs recognition was out of the scope of the VISITO Tuscany project.

[1] http://www.visitotuscany.it

3 Tested Approaches

In this Section, we report information about the tested approaches. First we discuss the SIFT that we selected as local feature. Then we briefly discuss the Bag-of-Features (BoF) and Vector of Locally Aggregated Descriptors (VLAD) that make use of the local features (SIFT in our case) in order to achieve high efficiency and effectiveness via aggregation of the information they contain.

3.1 SIFT

The Scale Invariant Feature Transformation (SIFT) [7] are extracted from key points selected using difference of Gaussians applied in scale space to a series of smoothed resampled images. The description of the region around this selected points rely on histogram of gradients. SIFT are not only the most important and cited local features ever defined, but they are still almost unbeaten in terms of effectiveness. Recently, binary local features have been proposed in order to improve efficiency of direct local features matching. In our experiments we achieve scalability aggregating features and thus we are more interested in effective representation of the images than efficient comparison of the features themselves.

3.2 Bag-of-Features

The Bag-of-Features (BoF) was initially proposed in [10] and has been studied in many other papers. The goal of the BoF approach is to substitute each local descriptor of an image with visual words obtained from a predefined vocabulary in order to apply traditional text retrieval techniques to CBIR.

The first step is selecting some visual words creating a vocabulary. The visual vocabulary is typically built clustering, using *k*-means, local descriptors of the dataset and selecting the centroids. The second step assigns each local descriptor to the identifier of the nearest word in the vocabulary. 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 (e.g., TF-IDF [9]) is adopted for evaluating the similarity between any two images.

As mentioned in [12], "a fundamental difference between an image query (e.g. 1500 visual terms) and a text query (e.g. 3 terms) is largely ignored in existing index design".

Efficiency and memory constraints have been recently addressed by aggregating local descriptors into a fixed-size vector representation that describe the whole image. In particular, Fisher Vector (FV) and VLAD have shown better performance than BoF. In this work we will focus on VLAD which has been proved to be a simplified non-probabilistic version of FV [6]. Despite its simplicity, VLAD performance is comparable to that of FV.

3.3 Vector of Locally Aggregated Desctipors (VLAD)

The VLAD representation was proposed in [5]. As for BoF, a codebook $\{\mu_1, \ldots, \mu_k\}$ is first learned using a cluster algorithm (e.g., k-means). Each local descriptor x_t in each image is then associated to its nearest visual word $NN(x_t)$ in the codebook. For each codeword, the differences $x_t - \mu_i$ of the vectors x_t assigned to μ_i are accumulated:

$$v_i = \sum_{x_t:\, NN(x_t)=\mu_i} x_t - \mu_i. \qquad (1)$$

The VLAD representation is the concatenation of the accumulated vectors, i.e. $V = [v_1^T \ldots v_K^T]$. Power-law and L^2 normalization are usually applied and for comparing two VLAD description L^2 Euclidean distance has been proved to be effective.

VLAD descriptions have a high dimensionality. Principal Component Analysis has been proposed to have a more compact representation.

4 Experiments

4.1 Dataset

Being partner of the EAGLE project, we had the opportunity to access a dataset of 17,155 photos related to 14,560 epigraphs made available to us by Sapienza University of Roma. For our experiments, we also needed a ground truth, i.e., photos in which we want to automatically recognize the epigraph together with the actual epigraph represented in the image. We constructed this ground truth selecting 70 photos from the whole dataset and removing them from the knowledge base. In other words, we removed these query photos from the ones that are given to the computer in order to understand the visual content of each epigraph. This was only possible for the epigraphs that had more than one photo. We also carefully selected queries that could represent the various types of epigraphs. In **Fig. 1** we report 5 query examples together with the other images for the same object available in the dataset.

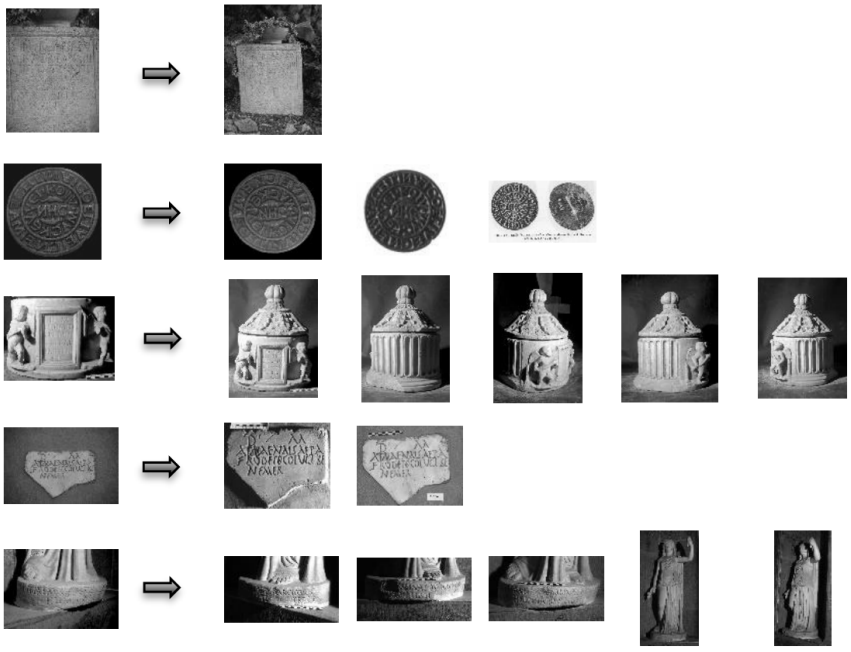

Fig. 1. Examples of query images and associated images of the same epigraph

4.2 Quality Measures

In order to recognize the actual object in a query image, we basically perform a visual similarity search between all the images in the dataset. Thus, the main goal is to have one image of the same epigraph as first result. Whenever this is not the case, it is interesting to understand at which position in the result list the most visually similar photo of the same object appears. In fact, while in this paper we are focusing on techniques able to scale up to the size of the dataset, traditional computer vision techniques could be applied on the results obtained in order to achieve better effectiveness. Given this considerations, we decided to report the probability p of finding an image of the same object between the first r results. For $r = 1$, p also equals the accuracy of a classifier that recognizes the query epigraph as the most similar that have been found.

For each technique, we report the probability p of finding an image containing the same epigraph given as query between the first r results varying r between 1 and 100. Results are reported with the r values on a logarithmic scale.

A more common measure of effectiveness is mean-Average Precision (mAP). In this case, not only the first relevant image but all the images associated with the query are considered. This measure reveal how good is the approach in reporting the related images in the top positions of the result list.

4.3 Experimental Setup

We extracted SIFT from images using the OpenCV library[2]. The BoF and VLAD approaches have been implemented by the NeMIS group of the ISTI-CNR in Java as part of our Visual Information Retrieval library publically available on GitHub[3].

Given an image, thousands of local features are extracted. In our case, we obtained an average of 1591 SIFT per image. However, the fact that some of them refer to bigger regions than others allows to select a subset of local features that are in principle more relevant [2]. Thus, in the experiments we also tried to reduce the number of local features selecting only the most important ones up to about 250 local features per image. In the following, we refer to this second approach as reduced-keypoints. We tested all the approaches both on the whole extracted local features and on the ones obtained filtering by region size.

4.4 Results

Bag-of-Features

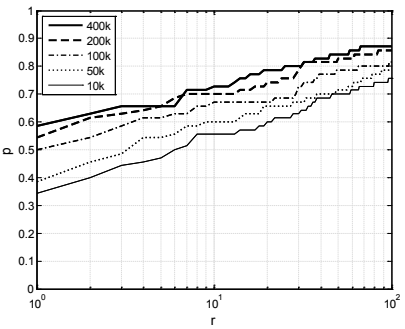

Fig. 2. BoF, cos TF-IDF

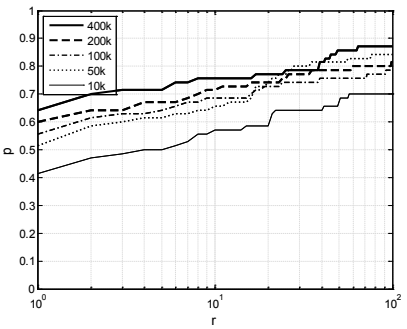

Fig. 3. BoF, cos TF-IDF, reduced-keypoints

Results in **Fig. 2** have been obtained using the BoF feature approach using the cosine TF-IDF similarity measure, varying the size of the vocabulary between 10k and 400k. As expected, the larger the vocabulary the better the results. However, differences between 200k or 400k are only marginal. Thus, we did not tested larger vocabularies.

In **Fig. 3** we report the same type of results considering only the most important local features. Results show that the SIFT reduction is useful.

[2] http://opencv.org/

[3] https://github.com/ffalchi/it.cnr.isti.vir

VLAD

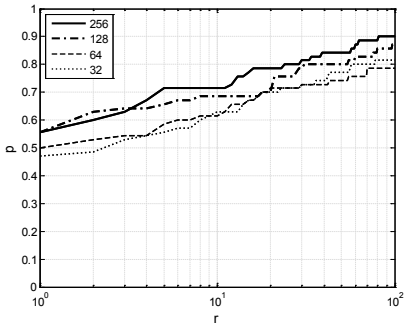

Fig. 4. VLAD

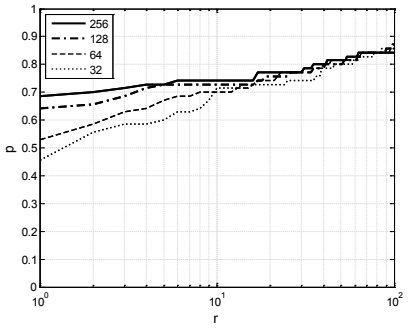

Fig. 5. VLAD, reduced-keypoints

In Fig. 4 and Fig. 5, we report the results obtained by the VLAD approach varying the size k of the codebook between 32 and 256 using all the extracted SIFT and the reduced ones respectively. It is interesting to see that the SIFT selection is useful when k is higher and r is smaller. In the other cases it can even decrease the quality of the results.

Considering that we are more interested on small r, which means having relevant images on the very first positions, the overall best results are the ones obtained for $k = 256$ and reducing number of local features.

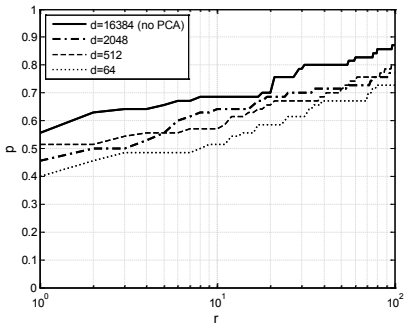

Fig. 6. VLAD-PCA, k=128

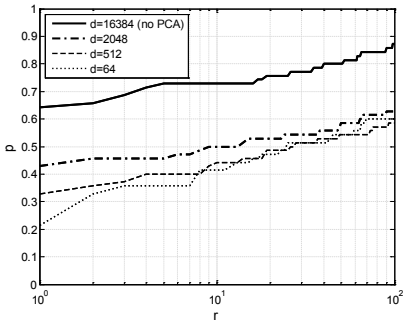

Fig. 7. VLAD-PCA, k=128, reduced-kp

We also tried to apply PCA on the VLAD vectors, especially for trying to reduce the complexity of the comparison. We then selected $k = 128$ and applied PCA in order to reduce the dimensionality of the vector (16,384). Results are reported for VLAD vector reduced to 2048, 512 and 64 dimensions. Unfortunately, the dimensionality reduction significantly reduces the quality of the results.

Comparison

In **Table 1**, we summarize the results obtained, ordered with respect to the mAP quality measure. Only the best approaches are shown. In the first column, we report a brief text about the approach. In the second column, the average number of SIFT considered is shown (i.e., 235 when local features reduction was applied and 1,591 otherwise). The third column reports the number of words used. While the words have been selected both for BoF and VLAD using k-means, their use is very different. Thus, in the "bytes" column, we computed the average size in bytes of the resulting representation. As quality measures, we used the probability p of having at least one related image between the first r results for $r = 1, 10, 100$ and the mAP.

In case we use these approaches to recognize the query image relying on the nearest image in the dataset, the best approach is the VLAD for a codebook size of 256 and selecting the 250 most relevant local features. In this case, the accuracy obtained is .69. The more traditional BoF-cos TF-IDF approach obtained good results when a large codebook (i.e., 400k) was used (as expected). It is interesting to note that this approach outperforms VLAD for $r = 10, 100$. Given that recent works as [3] have shown that VLAD can be more efficiently indexed than BoF, still VLAD is preferable.

Table 1. Comparison of results obtained by the overall best approaches ordered by mAP

Approach	*avg* SIFTs	codebook size	Bytes	p r=1	p r=10	p r=100	*mAP*
VLAD	235	256	131,072	**.69**	.74	.84	**.52**
BoF / cos TF-IDF	235	400,000	940	.64	**.76**	.87	.51
VLAD	235	128	65,536	.64	.73	.87	.49
BoF / cos TF-IDF	235	200,000	940	.60	.71	.81	.46
VLAD	1591	256	131,072	.56	.71	**.90**	.42
VLAD	1591	128	65,536	.56	.69	.87	.41
BoF / cos TF-IDF	235	100,000	940	.56	.69	.79	.42
VLAD	235	64	32,768	.53	.70	.86	.40
VLAD	1591	64	32,768	.50	.61	.79	.37
VLAD-PCA (d'=512)	1591	128	2,048	.44	.59	.79	.37

5 Conclusions and Future Work

In this work, we tested state-of-the-art object recognition techniques on an epigraphs dataset consisting of 17,155 photos. The best accuracy was obtained by using the VLAD approach that has been recently proposed for performing object recognition on a large scale.

The obtained accuracy was of .69, which is good considering the difficulties of the task and the few images available for each epigraph in the dataset. In fact, the dataset consists of 17,155 photos related to 14,560 epigraphs. This results in most of the epi-

graphs been represented by only one or two photos. However, we plan to improve this results performing re-ranking of the images obtained using these scalable techniques performing direct local features matching. To this goal, we also reported the probability of having a relevant images between the retrieve images. The results show that it is possible to have a relevant image between 100 retrieved ones with probability .90 using the VLAD approach with a codebook of size 256 and filtering the SIFT. Thus, we plan to try binary local features and other techniques in order to improve the obtained .69 accuracy up to the .90 obtainable, in theory, by re-ranking the 100 obtained using VLAD.

Bibliography

1. G. Amato, P. Bolettieri, F. Falchi and C. Gennaro, "Large Scale Image Retrieval Using Vector of Locally Aggregated Descriptors," in Similarity Search and Applications, vol. 8199, N. Brisaboa, O. Pedreira and P. Zezula, Eds., Springer Berlin Heidelberg, 2013, pp. 245-256.
2. G. Amato, F. Falchi and C. Gennaro, "Geometric consistency checks for kNN based image classification relying on local features," in SISAP '11: Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011, 2011.
3. G. Amato, F. Falchi and C. Gennaro, "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, 2013.
4. S. Dudani, "The Distance-Weighted K-Nearest-Neighbour Rule," IEEE Transactions on Systems, Man and Cybernetics, Vols. SMC-6(4), pp. 325-327, 1975.
5. H. Jégou, M. Douze, C. Schmid and P. Pérez, "Aggregating local descriptors into a compact image representation," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010.
6. H. Jégou, F. Perronnin, M. Douze, J. Sànchez, P. Pérez and C. Schmid, "Aggregating local image descriptors into compact codes," IEEE Transactions on Pattern Analysis and Machine Intelligence, Sep 2012.
7. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
8. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, no. 10, pp. 1615-1630, oct. 2005.
9. G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, New York, NY, USA: McGraw-Hill, Inc., 1986.
10. J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, Washington, DC, USA, 2003.
11. T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," Found. Trends. Comput. Graph. Vis., vol. 3, no. 3, pp. 177-280, 2008.
12. X. Zhang, Z. Li, L. Zhang, W. Y. Ma and H. Y. Shum, "Efficient indexing for large scale visual search," in Computer Vision, 2009 IEEE 12th International Conference on, 2009.

13. Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. S. Chua and H. Neven, "Tour the world: Building a web-scale landmark recognition," in CVPR, 2009.

Acknowledgement

This work was partially supported by the Europeana network of Ancient Greek and Latin Epigraphy (EAGLE, grant agreement number: 325122) co-funded by the European Commission within the ICT Policy Support Programme.