Visual Recognition in the EAGLE Project

Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi,
Fausto Rabitti, and Lucia Vadicamo

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche, Pisa, Italy
<name>.<surname>@isti.cnr.it

Abstract. In this paper, we present a system for visually retrieving ancient inscriptions, developed in the context of the ongoing Europeana network of Ancient Greek and Latin Epigraphy (EAGLE) EU Project. The system allows the user in front of an inscription (e.g, in a museum, street, archaeological site) or watching a reproduction (e.g., in a book, from a monitor), to automatically recognize the inscription and obtain information about it just using a smart-phone or a tablet. The experimental results show that the Vector of Locally Aggregated Descriptors is a promising encoding strategy for performing visual recognition in this specific context.

1 Introduction and Related Work

The large availability of digital cameras, especially embedded in smartphones and tablets, allows final users to make photos of their objects of interest at almost no cost. On one side, there are users making thousands of photos. On the other side, cultural heritage institutions typically have photos and metadata, both in digital form, related to the objects they preserve. In this context, there is a growing demand of technologies for content-based multimedia information retrieval.

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 that are automatically selected. Then, images are compared by matching their local features and searching for a geometric transformation that can associate the regions of both images. To deal with large dataset, compact images signatures based on the aggregation of local features have been proposed [10, 6].

Visual objects recognition has also been studied in the context of cultural heritage and computing. As an example, the VISITO Tuscany[1] project has investigated the visual recognition of cultural heritage objects (such as monuments, landmarks, etc.) [3]. However, to the best of our knowledge, there are no results in the literature regarding experiments conducted on ancient inscriptions.

The research reported in this paper summarizes the results presented in [4]. The focus is on searching for the most similar inscriptions in an archive with

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

respect to the one represented in a photo. This functionality will be integrated on an official EAGLE mobile application in order to allow the user to take a picture of an inscription (e.g., in a museum, in an archaeological site, in a book, etc.), send it to the central repository and receive back the information associated with that inscription.

2 Experiments

The dataset we used consists of 17,155 photos related to 14,560 ancient inscriptions that were made available by Sapienza University of Rome, within the EAGLE project. In order to visual recognizing inscriptions, we selected and tested the most promising approaches from the recent literature. As local feature we used the well known Scale Invariant Feature Transform (SIFT) [7]. 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 by selecting only the most important ones. With the goal of efficiently searching in the archive, we tested the most famous local features aggregation techniques: the Bag-of-Features (BoF) [10] and the Vector of Locally Aggregated Descriptors (VLAD) [6]. Both approaches use a codebook of visual words and the cosine similarity. For BoF, we applied the TF-IDF weighting [9].

To recognize the actual object in a query image, we perform a visual similarity search between all the images in the dataset. The optimum would be to have an image of the same inscription 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, traditional computer vision techniques could be applied on the results in order to achieve better effectiveness. Thus, we use the probability p of finding an image of the same object within the first r results, as quality measure. For $r = 1$, p equals the accuracy of a classifier that classify the query inscription as the most similar that has been found (i.e., a 1-NN classifier). A common measure of effectiveness in similarity search applications is the mean-Average Precision (mAP) which effectively summarizes the average and precision curves.

In Table 1, we report the best results obtained ordered with respect to the mAP. 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 selection was applied and 1,591 otherwise). The third column reports the number of words used in the aggregation phase. 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 relevant 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

Table 1. Experimental Results

Approach	**avg SIFTs**	**coodebook size**	**Bytes**	p **r=1**	p **r=10**	p **r=100**	**mAP**
VLAD	235	256	131,07	**.69**	.74	.84	**.52**
BoF / RANSAC	1591	200,000	19,092	.66	.70	.74	**.52**
BoF / cos TF-IDF	235	400,000	940	.64	**.76**	.87	.51
VLAD	235	128	65,536	.64	.73	.87	.49
BoF / RANSAC	1591	100,000	19,092	.64	.71	.77	.50
BoF / RANSAC	1591	400,000	19,092	.64	.66	.67	.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

on the nearest image in the dataset, the best approach is the VLAD that obtained an accuracy of 0.69 for a codebook size of 256 and selecting the 235 most relevant local features. The second best is the BoF in conjunction with geometry consistency checks performed using RANSAC [5]. However, this approach is not indexable and was only used as an effective but not efficient baseline. The more traditional cosine TF-IDF similarity applied to BoF obtained good results only in conjunction with a very large codebook (i.e., 400k). It is worth to note that this approach outperforms VLAD for r =10,100. We believe that VLAD is still preferable, since recent works as [1] have shown that VLAD can be more efficiently indexed than BoF.

3 Conclusions

In this work, we tested state-of-the-art object recognition techniques on a dataset of 17,155 photos related to 14,560 inscriptions. The best accuracy was obtained by using the VLAD approach that has been recently proposed for performing object recognition on a large scale. Surprisingly, even the BoF approach in conjunction with geometry consistency checks was not able to outperform the VLAD representation, that can be also more efficiently indexed than BoF. The obtained accuracy was of 0.69, which is good considering the difficulties of the task and the few images available for each inscription in the dataset. However, we plan to improve this results by performing re-ranking and direct local features matching. To this goal, we also reported the probability of having a relevant images between the retrieved images. The results show that it is possible to have a relevant image between 100 retrieved ones with probability 0.90 using the VLAD approach with a visual vocabulary of size 256 and filtering the SIFT. Thus, we plan to try binary local features and other techniques in order to improve the

obtained 0.69 accuracy up to the 0.90 obtainable, in theory, by re-ranking the first 100 image retrieved using VLAD.

Acknoledgments

This work was partially supported by EAGLE (Europeana network of Ancient Greek and Latin Epigraphy, co-founded by the European Commision, CIP-ICT-PSP.2012.2.1 - Europeana and creativity, Project Reference 325122).

References

1. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C.: Large scale image retrieval using vector of locally aggregated descriptors. In: Similarity Search and Applications, Lecture Notes in Computer Science, vol. 8199, pp. 245–256. Springer Berlin Heidelberg (2013)
2. Amato, G., Falchi, F., Gennaro, C.: On reducing the number of visual words in the bag-of-features representation. In: VISAPP 2013 - Proceedings of the International Conference on Computer Vision Theory and Applications, Volume 1, Barcelona, Spain, 21-24 February, 2013. pp. 657–662. SciTePress (2013)
3. Amato, G., Falchi, F., Rabitti, F.: Landmark recognition in VISITO: VIsual Support to Interactive TOurism in Tuscany. In: Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR2011) (2011)
4. Amato, G., Falchi, F., Rabitti, F., Vadicamo, L.: Inscriptions visual recognition. A comparison of state-of-the-art object recognition approaches. In: Proceedings of the First EAGLE International Conference. vol. 26, pp. 117 –131. Sapienza Universitá Editrice (2014)
5. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (Jun 1981)
6. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(9), 1704–1716 (2012)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(10), 1615–1630 (2005)
9. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
10. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 1470–1477. IEEE (2003)
11. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Foundations and Trends® in Computer Graphics and Vision 3(3), 177–280 (2008)