Indexing Vectors of Locally Aggregated Descriptors Using Inverted Files

Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro
ISTI-CNR
via G. Moruzzi, 1 - 56124 Pisa Italy
<name>.<last name>@isti.cnr.it

ABSTRACT

Vector of locally aggregated descriptors (VLAD) is a promising approach for addressing the problem of image search on a very large scale. This representation is proposed to overcome the quantization error problem faced in Bag-of-Words (BoW) representation. In this paper, we propose to enable inverted files of standard text search engines to exploit VLAD representation to deal with large-scale image search scenarios. We show that the use of inverted files with VLAD significantly outperforms BoW in terms of efficiency and effectiveness on the same hardware and software infrastructure.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrievals**]: Content Analysis and Indexing; H3.5 [**Information Systems**]: Online Information Services—*Commercial services*

General Terms

Experimentation, Algorithms

Keywords

landmarks recognition, image classification, local features

1. INTRODUCTION

In the last few years, local features [10] extracted from selected regions [14] have emerged as a promising method of representing image content in such a way that tasks of object recognition, can be effectively executed. A drawback of the use of local features is that a single image is represented by a large set of local descriptors that should be individually matched and processed in order to compare the visual content of two images. In principle, a query image should be compared with each dataset object independently. A very popular method to achieve scalability is the Bag-of-Words (BoW) [13] (or bag-of-features) approach that consists in replacing original local descriptors with the *id* of the most similar descriptor in a predefined vocabulary. Following the BoW approach, an image is described as a histogram of occurrence of visual words over the global vocabulary. Thus, the BoW approach used in computer vision is very similar to the traditional BoW approach in natural language processing and information retrieval. However, as mentioned in [16], "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 BoW [9]. In this work we will focus on VLAD which has been proved to be a simplified non-probabilistic version of FV [9]. Despite its simplicity, VLAD performance is comparable to that of FV [9].

Euclidean Locality-Sensitive Hashing [4] is, as far as we know, the only indexing technique tested with VLAD. While may other similarity search indexing techniques [15] could be applied to VLAD, in this work we decide to investigate the use of inverted files for allowing comparison of the VLAD and BoW approach on the same index. Permutation-Based Indexing [3, 2, 5] allows using inverted files to perform similarity search with an arbitrary similarity function. Moreover, in [6, 1] a Surrogate Text Representation (STR) derivated from the MI-File [2] has been proposed. The conversion of the image description in a textual form allows us to employ the search engine off-the-shelf indexing and searching abilities with a little implementation effort.

2. PROPOSED APPROACH

Conventional search engines use inverted index file indexing to speed up the solution of user queries. We are studying a methodology which will enable inverted files of standard text search engines to index vectors of locally aggregated descriptors (VLAD) to deal with large-scale image search scenarios. To this end, we first encode VLAD features by means of the perspective-based space transformation developed in [2]. The idea underlying this technique is that when two descriptors are very similar, with respect to a given similarity function, they "see" the "world around" them in the same way. In a next step, the "world around" can be encoded as a surrogate text representation (STR), which can be managed with an inverted index using a standard text-based search. The conversion of visual descriptors into a textual form allows us to employ off-the-shelf indexing and

ICMR '14 April 01 - 04 2014, Glasgow, United Kingdom

ACM 978-1-4503-2782-4/14/04 ...$15.00.
http://dx.doi.org/10.1145/2578726.2578788.

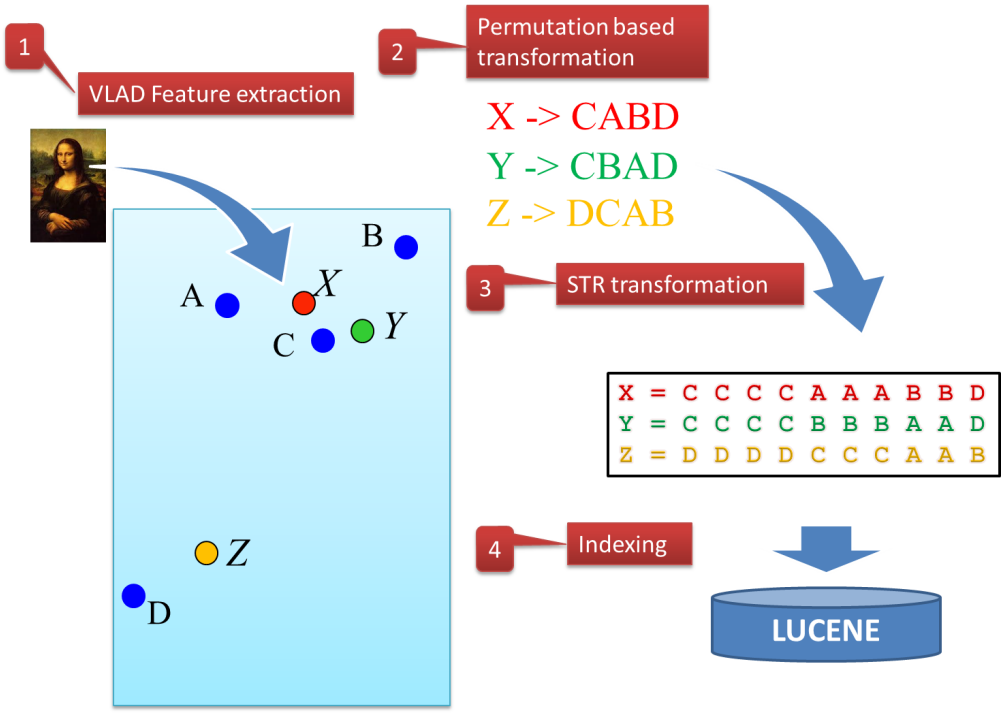

Figure 1: Example of perspective-based space transformation and surrogate text representation: 1) From the images we extract the VLAD features represented by points in a metric space. Blue points are reference features and colored points are data features, 2) The points are transformed into permutations of the references, 3) The permutations are transformed into text documents, 4) The text documents associated with the images are indexed.

searching functions with little implementation effort.

Our transformation process is shown in Figure 1: the blue points represent reference VLAD features; the other colours represent dataset VLAD features. The figure also shows the encoding of the data features in the transformed space and their representation in textual form (SRT). As can be seen intuitively, strings corresponding to VLAD features X and Y are more similar to those corresponding to X and Z. Therefore, the distance between strings can be interpreted as an approximation of the actual VLAD d. Without going into the math, we leverage on the fact that a text-based search engine will generate a vector representation of STRs, containing the number of occurrences of words in texts. With simple mathematical manipulations, it is easy to see how applying the cosine similarity on the query vector and a vector in the database corresponding to the string representations will give us a degree of similarity that reflects the similarity order of reference descriptors around descriptors in the original space. Mathematical details of the technique are outlined in [6].

The idea described so far uses a textual representation of the descriptors and a matching measure based on a similarity offered by standard text search engines to order the descriptors in the dataset in decreasing similarity with respect to the query. The result set will increase in precision if we order it using the original distance function used for comparing features. Suppose we are searching for the most similar (nearest neighbours) descriptors to the query. We can improve the quality of the approximation by re-ranking, using the original distance function d and the first c ($c \geq k$) descriptors from the approximate result set at the cost of more c distance computations. This technique significantly improves accuracy at a very low search cost.

We applied the STR technique to the VLAD method comparing both effectiveness and efficiency with the state-of-the-art BoW approach on the same hardware and software infrastructure using the publicly available and widely adopted 1M photos dataset. Given that the STR combination gives approximate results with respect to a complete sequential scan, we also compare the effectiveness of VLAD-STR with standard VLAD. Moreover, we considered balancing efficiency and effectiveness with both BoW and VLAD-STR approaches. For the VLAD-STR, a similar trade-off is obtained varying the number of results used for re-ordering. Thus, we do not only compare VLAD-STR and BoW on specific settings but we show efficiency vs effectiveness graphs for both. For the VLAD-STR, a trade-off is obtained varying the number of results used for re-ordering.

3. EXPERIMENTS

3.1 Setup

INRIA Holidays [8, 9] is a collection of 1,491 holiday images. The authors selected 500 queries and, for each of them, a list of positive results. To evaluate the approaches on a large scale, we merged the Holidays dataset with the Flickr1M[1] collection as in [7, 8, 9]. The ground–truth is the one built on the INRIA Holidays dataset alone, but it

[1] `http://press.liacs.nl/mirflickr/`

BoW		
avg#Words	mAP	avg mSec
7	0.03	525
16	0.07	555
37	0.11	932
90	0.14	1463
233	0.15	2343

Table 1: Effectiveness (mAP) and efficiency (mSec) with respect to the average number of distinct words per query obtained with the BoW approach varying the query size.

VLAD		
#reordered	mAP	avg mSec
0	0.13	139
100	0.24	205
1000	0.29	800
2000	0.30	1461
4000	0.31	2784

Table 2: Effectiveness (mAP) and efficiency (mSec) obtained with the VLAD approach in combination with STR, with respect to the number of results used for reordering.

is largely accepted that no relevant images can be found between the Flickr1M images. SIFT descriptors and various vocabulary were made publicly available by Jegou et al. for both the Holidays and the Flickr 1M datasets[2]. For the BoW approach we used the 20K vocabulary.

For representing the images using the VLAD approach, we selected 64 reference descriptors using *k-means* over a subset of the Flickr1M dataset. As explained Section 2, a drawback of the perspective based space transformation used for indexing the VLAD with a text search engine is that it is an approximate technique. However, to alleviate this problem, we reorder the best results using the actual distance between the VLAD descriptors. For the STR we used 4,000 references (i.e., $m = 4,000$) randomly selected from the Flickr1M dataset.

During the experimentation also 256 references for VLAD and up to 10,000 references for the STR were selected but the results were only slightly better than the ones presented while efficiency significantly reduced.

All experiments were conducted on a Intel Core i7 CPU, 2.67 GHz with 12.0 GB of RAM a 2TB 7200 RPM HD for the Lucene index and a 250 GB SSD for the VLAD reordering. We used Lucene v3.6 running on Java 6 64 bit over Windows 7 Professional.

The quality of the retrieved images is typically evaluated by means of precision and recall measures. As in many other papers [12, 7, 11, 9], we combined this information by means of the mean Average Precision (mAP), which represents the area below the precision and recall curve.

3.2 Results

In Table 1, we report the mAP obtained with the BoW approach varying the size of the query in terms of average number of distinct words. In this case, the query words have been filtered using the *tf*idf* approach. The average number of words per image, as extracted by the INRIA group, is 1,471 and they were all inserted in the index without any filtering. The filtering was used only for the queries and results are reported for average number of distinct words up to 250. In fact, bigger queries result in heavy load of the system. It is worth to mention that we were able to obtain 0.23 mAP performing a sequential scan of the dataset with the unfiltered queries.

The results show that while the BoW approach is in principle very effective (i.e. performing a sequential scan), the high number of query visual words needed for achieve good results significantly reduces his usability.

In Table 2, we report the results obtained using the VLAD approach in combination with the use of the STR. Given that for indexing the images we used a STR, it is useful to reorder the better results obtained from the text search engine using the actual VLAD distance. Thus, we report mAP and avg mSec per query for the non–reordering case and for various values of results used for reordering. The reordering phase dominates the average query time but it significantly improves effectiveness especially if only 100 or 1,000 objects are considered for reordering. As mentioned before, we make use of SSD for speed-up reordering phase but even higher efficiency could be obtained using PCA as proposed in [9]. Please note that even though the reordering phase cost for VLAD can be reduced, the reported results already show that VLAD outperform BoW.

It is worth to mention that we also performed a sequential scan of the entire dataset obtaining a mAP of 0.34 for VLAD. In fact, as depicted in 2, the results obtained with the VLAD-STR approach are an approximation of the results obtained with a complete pair wise comparison between the query and the dataset object. The same is true when LSH indexing is used as in [9]. Results show that the approximation introduced byt STR does not impact significantly the effectiveness of the system when at least 1,000 objects are considered for reordering.

In Figure 2, we plot mAP with respect to the average query execution time for both BoW and VLAD as reported in Table 1 and Table 2. The graph underlines both the efficiency and effectiveness advantages of the VLAD technique with respect to the BoW approach. The efficiency vs effectiveness graph reveals that VLAD-STR obtains the same mAP values as BoW, for an order of magnitude less in response time. Moreover, for the same response time, VLAD-STR is able to obtain twice the mAP of BoW.

4. CONCLUSIONS

In this work, we proposed the usage of STR in combination with VLAD descriptions in order to index VLAD with off-the-shelf text search engines. Using the very same hardware and text search engine (i.e., Lucene), we were able to compare with the state-of-the-art BoW. Results obtained for BoW confirm that the high number of visual terms in the query significantly reduces efficiency of inverted lists. Even

[2] `http://lear.inrialpes.fr/~jegou/data.php`

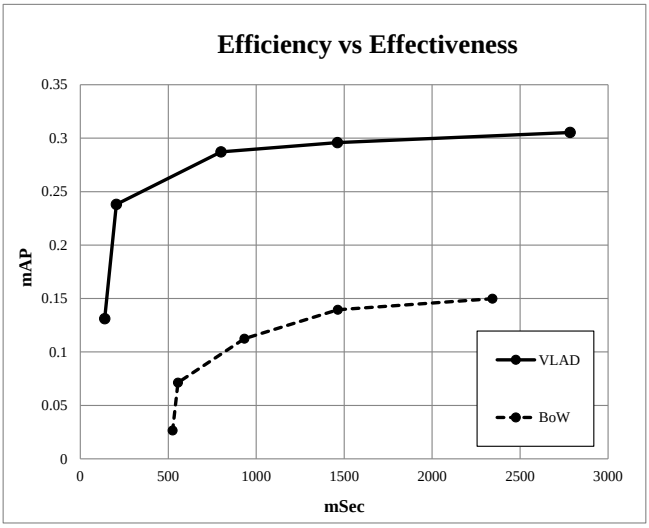

Figure 2: Effectiveness (mAP) with respect to efficiency (mSec per query) obtained by VLAD and BoW for various settings.

though results showed that this can be mitigated reducing the number of visual terms in the query with a *tf*idf* weighting scheme, the VLAD-STR significantly outperforms BoW in terms of both efficiency and effectiveness. The efficiency vs effectiveness graph reveals that VLAD-STR is able to obtain the same values of mAP obtained with BoW for an order of magnitude less in response time. Moreover, for the same response time, VLAD-STR is able to obtain twice the mAP of BoW.

Acknowledgments

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.

5. REFERENCES

[1] G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, and F. Rabitti. Combining local and global visual feature similarity using a text search engine. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 49 –54, june 2011.

[2] G. Amato and P. Savino. Approximate similarity search in metric spaces using inverted files. In *Proceedings of the 3rd international conference on Scalable information systems*, InfoScale '08, pages 28:1–28:10, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[3] G. Chavez, K. Figueroa, and G. Navarro. Effective proximity retrieval by ordering permutations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1647 –1658, sept. 2008.

[4] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, SCG '04, pages 253–262, New York, NY, USA, 2004. ACM.

[5] A. Esuli. Mipai: Using the pp-index to build an efficient and scalable similarity search system. In *Proceedings of the 2009 Second International Workshop on Similarity Search and Applications*, SISAP '09, pages 146–148, Washington, DC, USA, 2009. IEEE Computer Society.

[6] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino. An approach to content-based image retrieval based on the lucene search engine library. In *Proceeding of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2010)*, LNCS.

[7] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2357 –2364, 29 2009-oct. 2 2009.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3304–3311, jun 2010.

[9] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sept. 2012. QUAERO.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615 –1630, oct. 2005.

[11] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3384 –3391, june 2010.

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

[13] 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*, ICCV '03, pages 1470–, Washington, DC, USA, 2003. IEEE Computer Society.

[14] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008.

[15] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search - The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Kluwer, 2006.

[16] 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*, pages 1103 –1110, 29 2009-oct. 2 2009.