Large Scale Image Retrieval Using Vectors of Locally Aggregated Descriptors

by Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi and Claudio Gennaro

We propose using vectors of locally aggregated descriptors (VLAD) to address the problem of image search on a very large scale. We expect that this technique will overcome the quantization error problem faced in Bag-of-Words (BoW) representations.

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 aggregrated 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 [1]. 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 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 original VLAD distance *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 [2].

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 *d* used for comparing features. Suppose we are searching for the *k* most similar (nearest

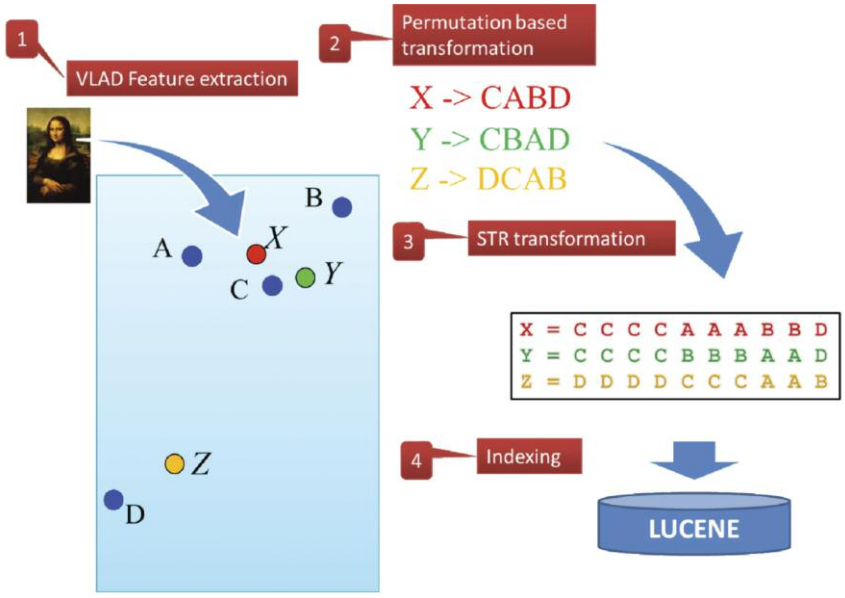

Figure 1: Example of perspective-based space transformation and surrogate text representation: 1) From the images we extract the VLAD features represented by points in a metric space. Blue points are reference features and colored points are data features, 2) The points are transformed into permutations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

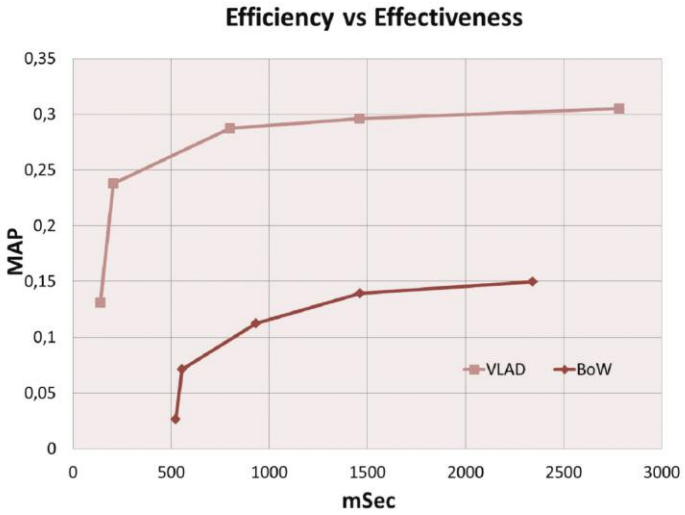

Figure 2: Effectiveness (MAP) with respect to efficiency (mSec per query) obtained by VLAD and BoW for various settings.

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 $c \cdot (c \geq k)$ 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.

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. In Figure 2, we plot mean average precision (MAP) with respect to the average query execution time for both BoW and VLAD. 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.

References:

[1] G. Amato, P. Savino: "Approximate similarity search in metric spaces using inverted files", in proc. of InfoScale '08, pages 28:1-28:10, ICST, Brussels, Belgium

[2] C. Gennaro, G. Amato, P. Bolettieri, P. Savino: "An approach to content-based image retrieval based on the Lucene search engine library", in proc. of ECDL 2010, Springer LNCS.

Please contact:

Giuseppe Amato
ISTI-CNR, Italy
Tel: +39 050 3152810
E-mail: giuseppe.amato@isti.cnr.it