JOANNEUM RESEARCH at TRECVID 2014 Instance Search Task

Harald Stiegler, Werner Bailer
JOANNEUM RESEARCH, DIGITAL – Institute for Information and Communication Technologies
8010 Graz, Austria
Email: {firstName.lastName@joanneum.at}

ABSTRACT

We participated in the instance search (INS) task. We submitted one run, using VLAT features created from SIFT descriptors of DoG interest points. Each key frame of the query is used, and the results are combined into one result list. The scores are low, as VLAT is not discriminative enough for many of the queries.

I. APPROACH

A. Approach for submitted runs

For TRECVID 2014 instance search (INS), we investigated the use of compact descriptors based on SIFT descriptors. In order to handle large sets of images and videos in instance search, representations such as Fisher Vectors [1], VLAD [2] and VLAT [3] have been proposed. For our experiments, we have selected VLAT using the C++ implementation from [4].

The approach is based on extraction and matching of image areas around salient key points, using the SIFT (Scale Invariant Feature Transform) algorithm [5]. The SIFT algorithm has become very popular due to its powerful performance and is still used as a basic tool in the area of object recognition, near duplicate detection and other various related tasks. The SIFT algorithm is described in detail in [5], which describes the localization, extraction and matching of key points and their descriptors. Extraction of the descriptors has been implemented on GPU using NVIDIA CUDA[1] in order to speed up processing.

If SIFT descriptors are extracted sparsely around key points, storing the raw SIFT descriptors is still competitive in terms of memory consumption. A single SIFT key point description (eight bytes each for x and y coordinates, orientation and scale) and its descriptor (128 bytes) requires 160 bytes. For a whole frame with 2,000 key points, the amount of required memory is still 320 KB. The default VLAT signature is independent of image data and consists of about 2.1 MB, thus only advantageous in terms of memory consumption when using a large number of key points, e.g., for dense SIFT. More sophisticated VLAT flavors can compress signature down to 4KB (VLAT Compact), but this option was not considered suitable due to the long duration of the signature computation (Gram matrix eigendecomposition). Another VLAT derivative (VLAT packed) could not proven to be working correctly and has not been investigated in detail. VLAT Wise has been found to work well and computes signatures with a size of about 1.2 MB to 1.4 MB per image, depending on parametrization (dictionary size set to 128 in our experiments).

In terms of matching speed VLAT Wise achieves up to 29,000 frame comparisons per second if the signatures are already in memory (multithreaded on an Intel Core i7-2600, 3.4 GHz). If signatures have to be loaded from disk the average number of comparisons per second drops down to 2,700. In comparison, brute force SIFT descriptor matching with GPU support (NVidia GeForce GTX 560 Ti) achieves in the range of 1,200 frame comparisons per second.

The original VLAT Wise implementation loads all signatures into memory and compares them, which is clearly not feasible for large scale data sets. In order to prevent running out of memory when matching against large data sets, the matching algorithm has been modified to loading of signatures piecewise which reduces matching throughput.

In contrast to the original VLAT implementation, we use a precalculated vocabulary for the entire data set. We use a subset of the SIN 2014 videos set to generate a dictionary in advance, which speeds up signature calculation later, and is also the prerequisite for processing data incrementally.

In our experiments, we use every 5^{th} frame of a video, and if interlacing artifacts are present, we discard one field. The frame width is rescaled to 640 pixels, and the height is scaled accordingly to maintain the aspect ratio. The number of SIFT descriptors per frame has been limited to 500, and the VLAT dictionary size has been set to 128.

We used all query sample images, and the results of all samples are combined into one list, which is then ranked.

B. Additional experiments

After the submission of the official runs, we performed the following experiments.

Based on the VLAT results described above, we reranked the 2,000 top results using brute force SIFT matching. Brute force SIFT descriptor matching is performed by assigning the nearest SIFT descriptor neighbor from the target frame to each SIFT descriptor in the source frame. From each matching descriptor pair a mapping from the source to the target frame can be computed and logged in a 4D-histogram (x,y,orientation,scale). In order to speed up SIFT descriptor

[1] http://www.nvidia.com/object/cuda_home_new.html

matching, it has been implemented on a GPU using NVIDIA CUDA.

In addition, we tested matching using only GPU accelerated SIFT over the entire database. This is the approach we also used for INS 2013 [6].

II. Results

A. Submitted runs

The MAP across all queries is rather poor (0.0095). Clearly, the VLAT signature is in most cases not discriminative enough for this task. 9099 (checkerboard band of police cap) is an exception to that, where the aggregation of features of the band obviously allows separating images containing the band from those that do not. Otherwise, the objects contribute too little to the VLAT signature to be sufficiently discriminative for retrieving the right segments at the top of the list.

B. Additional experiments

Reranking nearly doubles the MAP across all queries to 0.0181 from the original VLAT results, with only a very moderate additional computation cost. Except for query 9099, the results for all queries improve by reranking, further supporting the assumption that the relatively high score of 9099 from the original VLAT matching is an outlier. However, for some queries only a small fraction of relevant segments made it into the top 2,000 after VLAT matching, so that they are out of reach for reranking.

Brute force SIFT matching yields significantly better results, for example, MAPs of 0.1124 for query 9102 and 0.6524 for 9112. Also queries that returned no relevant results after the initial VLAT matching return some of the relevant shots (e.g., MAP 0.0126 for 9106).

III. Conclusion

While VLAT allows for fast matching, it clearly has limitations for the type of queries used in INS. Reranking with a more powerful method can improve the results consistently for all but one of the queries. However, many relevant shots are already lost after the initial VLAT matching, so that reranking cannot provide significant improvements for some of the queries.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme under grant agreement n° 610370, "ICoSOLE – Immersive Coverage of Spatially Outspread Live Events" (http://www.icosole.eu/), and by the project grant "QuOIMA" under Austrian National Security Research Development Programme KIRAS.

References

[1] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2007.383266

[2] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[3] R. Negrel, D. Picard, and P. H. Gosselin, "Web-scale image retrieval using compact tensor aggregation of visual descriptors," *IEEE MultiMedia*, vol. 20, no. 3, p. 2433, 2013. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/MMUL.2013.14

[4] "VLAT Software," http://vlat.fr/software.html.

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] W. Bailer, H. Stiegler, and R. Mrzinger, "Joanneum research at trecvid 2013: Semantic indexing and instance search," in *Proceedings of TRECVID Workshop*, Gaithersburg, MD, USA, Nov. 2013.

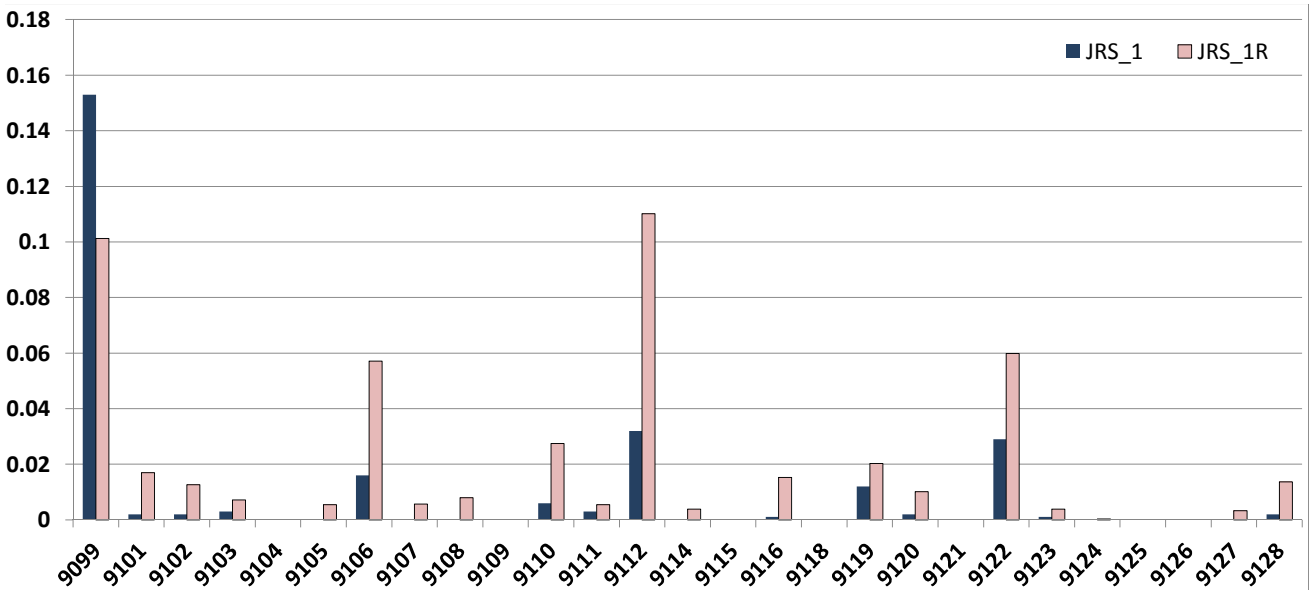

Fig. 1. Scores of the submitted run (JRS_1, using VLAT) and of the run using SIFT reranking (JRS_1R).