JOANNEUM RESEARCH at TRECVID 2016 Instance Search Task

Martin Höffernig and Werner Bailer
JOANNEUM RESEARCH, DIGITAL – Institute for Information and Communication Technologies
8010 Graz, Austria
Email: {martin.hoeffernig, werner.bailer}@joanneum.at

ABSTRACT

We participated in the instance search (INS) task. We submitted two runs, both using a compact video descriptor implemented using MPEG CDVS, without specific tools for person identification. A database is built for every location (mined from the samples), against which the queries are run. The two runs differ in terms of the fusion strategy.

I. APPROACH

For TRECVID 2016 instance search (INS), we implemented a system for retrieving relevant shots depicting specific persons in specific locations. As a preprocessing step, for each shot of the test videos, location information is computed. For this task, the set of known location example videos, the test videos, and the related master shot boundary information are considered. Firstly, for each location, the relevant location example videos are concatenated to one separate location video (using FFmpeg). These location videos are used to set up a ground truth database for finding relevant locations of interest in the test videos. This step is done by implementing the visual matching approach described below. As a result, for each test video, time span and matching score information about possible detected locations are returned. This information is annotated to the corresponding shots in the test videos.

The available person sample images are then run against the identified video subsets corresponding to the different locations. The results from each of the samples and each of the locations are then fused to obtain the final result list.

A. Visual matching

In our approach we use a generic visual instance search method. Instance search, i.e., finding video clips containing a similar foreground object, background or scene as in the query, is still a challenging problem in large-scale video collections. In contrast to video copy detection, the problem cannot be addressed only by global visual descriptors, due to the variability with which the object of interest may be depicted. In recent years, there has been significant progress in defining more compact visual descriptors, typically by aggregating local descriptors (either sampled from interest points or densely) and applying means such as dimensionality reductions and binarisation. Examples of such methods are Fisher Vectors [1], VLAD [2] and its improvements [3], VLAT [4] and CDVS [5]. While these descriptors achieve good matching performance even at small descriptor sizes, they are all descriptors for still images that need to be applied independently to individual frames of the video. Thus, they do not make use of the temporal redundancy of the video. This is not only an issue of the size of the extracted descriptor, but also of the matching complexity, as pairwise matching of the frame descriptors has to be performed.

In order to better support the nature of video we use a descriptor for image sequences, which encodes a set of consecutive and related frames (i.e., a segment such as a shot) as a single descriptor. The descriptor is created from an aggregation of sets of local descriptors from each of the images, and contains an aggregation of global descriptors and a time and location indexed set of the extracted local descriptors. The descriptor extraction is based on a method for local descriptor extraction from interest points and a method for aggregation of such descriptors to a global descriptor, but is agnostic of the specific type of descriptor and aggregation method (as long as they fulfill certain properties). Depending on the bitrate, temporal subsampling and (possibly lossy) compression of local descriptors can be applied (we used only a lossless mode for the INS experiments). The matching process is hierarchical, in the sense that matching of details is only performed if some level of similarity is found on the coarser level.

While the proposed descriptor could be implemented using different local descriptors and aggregation methods, we base the compact image sequence descriptor on the MPEG CDVS descriptor, making use of the global and local parts of the descriptor. A CDVS descriptor contains a set of local SIFT descriptors [6] sampled around ALP interest points [7], which are quantised to a ternary representation. In addition, it contains an aggregated global descriptor, represented as Scalable Compressed Fisher Vector (SCFV) [8] as a binary vector. Retrieval is performed by using an index for the global descriptors, and then performing pairwise matching of the query with the top k results returned from the global index.

By comparing the resulting time span information and the available master shot boundary information, corresponding shots in the test videos are identified. Then the location information is assigned to these shots based on the matching scores. In case multiple location assignments for one shot are possible, the location having the highest matching score value is selected only.

B. Person queries

After the location information is assigned to each shot of the test videos, person-related information is also annotated to the test videos. Therefore the the available person sample images are considered. The background of the these sample images for each person of interest is blurred in order to eliminate the influence of the background. This step is done by considering the corresponding mask information of these frame images. Then the occurrence of these frame images in the test videos with respect to the assigned locations are computed. We use the same visual matching approach as above, treating the problem as an instance search problem without specific means for person identification. As a result, for each location, time span and matching score information about detected person frame images are returned. Again, this information is annotated to the corresponding shots in the test videos.

C. Result fusion

Finally, the topic results are composed based on the annotated location and person-related information of the shots in the test videos. For each topic, the shots depicting relevant person sample images are selected. Then these shots are ranked. Therefore two different ranking methods are available. The first method is based on a separate shot ranking for each relevant person sample image followed by the fusion of these rankings. In detail, the detected shots of the person sample images of the person of interest are sorted separately in descend score value order and fused together to one result list. The underlying fusion method is based on the iterative ranking of the top ranked shot of each person sample shot list. Therefore the respective top ranked shot of each of the four person sample shot list is selected and sorted by the score value. Then these sorted shots are added to the result list. The second ranking method is based on the global ranking of the score values. Here all the shots from the relevant person sample shot lists are merged together and then sorted by the score value.

II. Results

A. Submitted runs

We have submitted two runs using the method described above. The runs differ only by the ranking method. The results of run JRS1 are based on the fusion of separate shot ranking for each frame image, while the results of run JRS2 are ranked based on the global scores only. However, no significant differences between our results could be found.

The MAP of both runs is very low, 0.000333 for JRS1 and 0.000267 for JRS2.

B. Analysis of location identification

Based on the topic results provided by NIST, a ground truth list for locations is created. In case a shot is successfully designated to a topic in the NIST result list, the actual location information of this shot is inferred. This ground truth list about assigned locations serves as the basis for our analysis of our computed location information of the shots of the test videos.

After comparing this ground truth list with our runs, we found out that the recall of detecting correct locations is about 0.15. Furthermore the average precision of our location assignment algorithm is calculated. The basis for this calculation is a shot list containing all shots with an assigned location information sorted by score value. It turned out that the average precision is about 0.12. Assigning all shots without any location information attached, the correct location information from the ground truth list, the average precision would rise to 0.52. This value designates the upper bound achievable with our assignment algorithm. In addition, the inferred average precision (infAP) [9] is also computed for each location. This value is between 0.27 for the location Laundrette and 0.06 for location Kitchen1. Then mean inferred average precision about all locations is 0.14. In Figure 1 the calculated inferred average precision for all locations of interest is depicted.

It has to be noted that increasing the number of retrieval results to be taken into account by our location assignment algorithm, has no significant effect on the accuracy of our runs.

III. Conclusion

As can be expected, the performance of a generic instance search approach is low for person queries. However, the analysis of the location results also shows that improvement in the recall of shots of a specific location is required, in order to provide a good basis for subsequent person queries.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610370, ICoSOLE ("Immersive Coverage of Spatially Outspread Live Events", http://www.icosole.eu), and from the Austrian Research Promotion Agency under the KIRAS grant E.V.A.

References

[1] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[2] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 3304–3311.

[3] R. Arandjelovic and A. Zisserman, "All about vlad," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 1578–1585.

[4] D. Picard and P.-H. Gosselin, "Improving image similarity with vectors of locally aggregated tensors," in *IEEE International Conference on Image Processing*, Brussels, BE, Sept. 2011.

[5] "Information technology – multimedia content description interface – part 13: Compact descriptors for visual search," 2014.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] "ISO/IEC 15938-13, Information technology – Multimedia content description interface – Part 13: Compact descriptors for visual search," 2015.

[8] J. Lin, L.-Y. Duan, Y. Huang, S. Luo, T. Huang, and W. Gao, "Rate-adaptive compact fisher codes for mobile visual search," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 195–198, 2014.

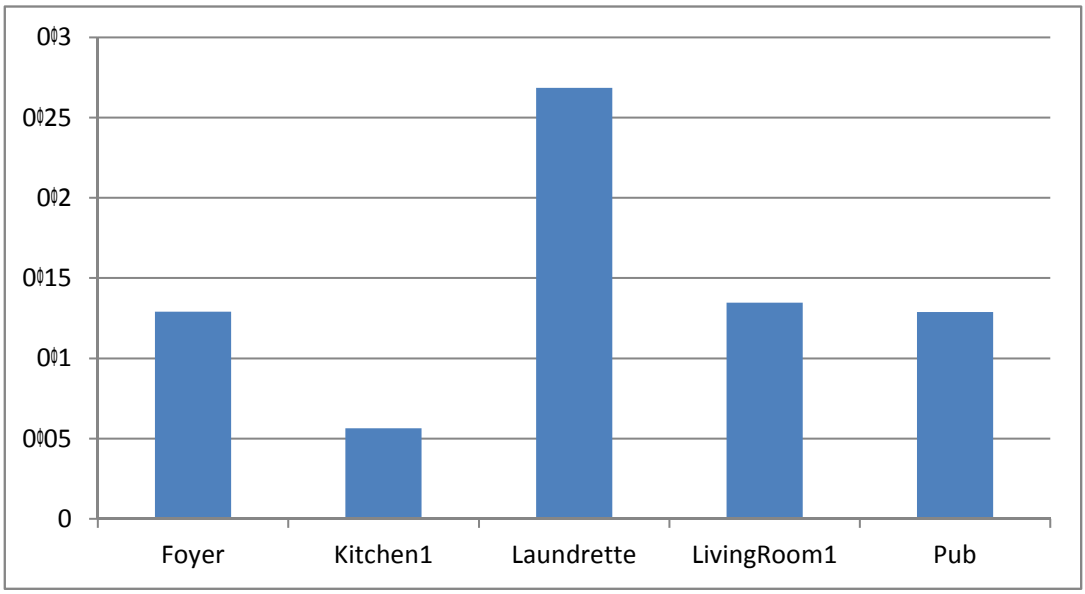

Fig. 1. Calculated inferred average precision (infAP) of the locations of interest.

[9] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, ser. CIKM '06. New York, NY, USA: ACM, 2006, pp. 102–111.