ARTEMIS at TRECVID 2013: Instance Search Task

Andrei Bursuc, Titus Zaharia

Institut Mines-Télécom ; Télécom SudParis, ARTEMIS Department, UMR CNRS 8145 MAP5,
9 rue Charles Fourier, 91011 Evry Cedex, France
{Andrei.Bursuc, Titus.Zaharia}@telecom-sudparis.eu

Abstract. This paper describes the approach proposed by the ARTEMIS team at TRECVID 2013, Instance Search (INS) task. The method is based on the Bag-of-Words representation obtained from uniform sampling of the frames of the videos. We propose two types of shot descriptors: one relying on a single representative frame for each video shot and another one collecting visual descriptors from multiple frames of the video clips.

1 Structured Abstract

Briefly, what approach or combination of approaches did you test in each of your submitted runs? (please use the run id from the overall results table NIST returns)

- all runs: Hessian Affine detectors and RootSIFT descriptor quantized into BoW feature vectors using a vocabulary of size 1M.
- **F_NO_ARTEMIS_1_1:** BoW vectors generated at shot level on resized frames (384x288 surface). Single query BoW vector generated from the multiple example images.
- **F_NO_ARTEMIS_2_2:** BoW vectors generated at shot level on resized frames (384x288 surface). A BoW vector is generated for each query image and queried on the entire dataset. The best score of the video clip among the different query runs is selected.
- **F_NO_ARTEMIS_3_3:** BoW vectors generated from a single representative keyframe for each video shot. Single query BoW vector generated from the multiple example images.
- **F_NO_ARTEMIS_4_4:** BoW vectors generated from a single representative keyframe for each video shot. A BoW vector is generated for each query image and queried on the entire dataset.

What if any significant differences (in terms of what measures) did you find among the runs?
Although the shot based descriptor has a richer representation, the resizing of the video resolutions has strongly affected its performances. Using one single frame for every video clip yielded better results. The use of a single query descriptor computed from all provided query images improves significantly the results and the retrieval time.

Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?
The large size of the vocabulary has compensated partially the reduced number of detected interest points from the resized video frames.
The aggregations of the results from multiple runs into a single result list can affect negatively the overall performance. The unified query descriptor gave better results, but makes it impossible to perform a spatial consistency check for re-ranking top results.

Overall, what did you learn about runs/approaches and the research question(s) that motivated them?
We have learned that the shot level BoW vectors provide a rich representation of the video shot and capture information that can be lost when selecting only a few of the frames of the video shot. However, the resizing of the video frames makes it difficult to retrieve small and tiny objects.

2 Instance Search Task description

The Instance Search Task (INS) [1] tackles the problem of retrieving an object in a video dataset as quickly as possible by using a limited amount of data about the object (a couple of images). Given a collection of test video clips and a collection of queries that delimit a person, object or place entity in some example video, participant applications have to locate for each query up to the 1000 clips most likely to contain a recognizable instance of the entity.

The video dataset proposed for the 2013 edition consisted of a 189 hours of video content from the BBC Eastenders TV series (programme material ©BBC) [2]. The videos feature a high quality and resolution (768 x 576) and are encoded in the MPEG-4 format. The corpus consists of approx. 470,000 short video clips of a couple of seconds for which the mastershot references are given. A set of frames from the BBC Eastenders dataset is illustrated in Figure 1.

For testing, 30 query topics were considered, each consisting of a set of 4 example images drawn from test videos containing the item of interest. For each example frame, a binary mask of the region of interest was provided. In addition, each topic included information about the category of the current topic (*e.g.*, person, object, location). A subset of topics and example images are depicted in Figure 2.

Figure 1. Sample from the BBC Eastenders dataset.

3 Approach overview

For our approach, we have considered a large scale adapted Bag-of-Words representation [3] built on a vocabulary of 1M descriptors [4]. We identify the regions of interest with the Hessian Affine covariant region detector [5, 6] and describe each region with the RootSIFT [7] descriptor; which is a SIFT [8] variant using a square root Hellinger kernel for the similarity measure. RootSIFT has yielded superior performances on the Oxford 5k and 105k, Paris 6k and Holidays datasets [7, 11].

For the computation of the visual vocabulary we employ an approximate k-means clustering algorithm [4] which makes it possible to cluster a large amount of high dimensional visual descriptors quickly (e.g., millions of RootSIFT descriptors) while preserving a relatively good quality of the results. Unlike the classic k-means, where most of the computation time is spent on retrieving all nearest neighbors between feature points and cluster centers, the approximate k-means optimizes this step by using approximate nearest neighbor techniques such as FLANN [12]. In this respect, we employ a forest of multiple randomized kd-trees built over the cluster centers at the beginning of every iteration, in order to improve the processing speed. For randomized kd-trees the splitting dimension is chosen randomly from the dimensions with the highest variance. The union of such trees creates then an overlapping partition of the feature space and reduces the possible quantization errors by selecting the top nearest neighbor across all trees. The size of the random forest was set to 8 kd-trees. For a set of 20M SIFT descriptors, the clustering into a 1M codebook on a 4 core machine took approx. 20 hours.

We propose for testing two video description strategies: shot-based (run 1 and 2) and frame-based (runs 3 and 4). For the former, we sample uniformly every 12 frames (approx. 2 frames per second), while for the latter we select one single representative keyframe per shot.

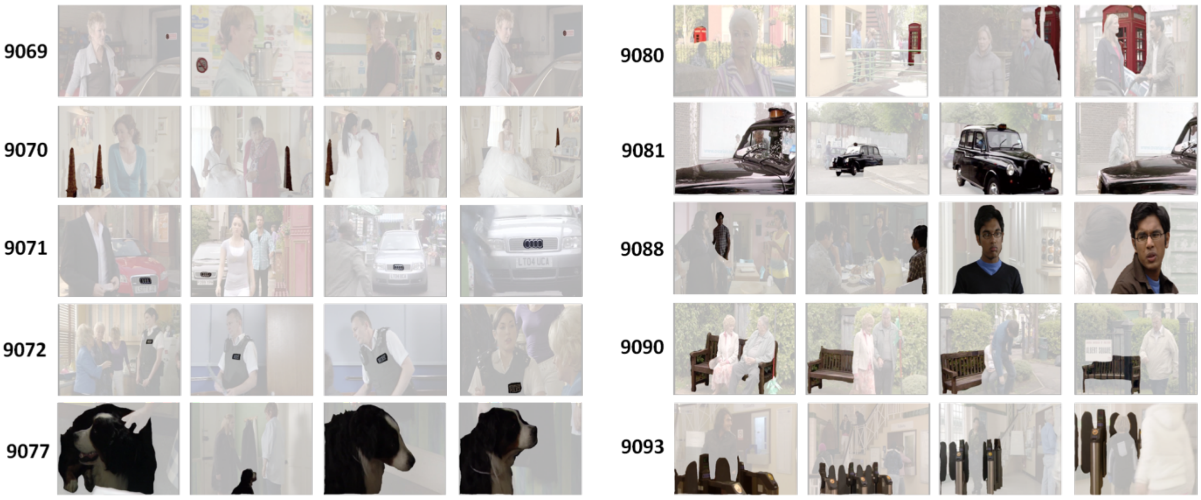

Figure 2. Example of query topics for the BBC Eastenders dataset. The regions defined by the binary mask are illustrated in full color, while the background is whitened. The ID of every topic is given in the first column.

4 Shot-based description (runs 1 and 2)

Our goal is to capture as much as possible the details from the video clips (objects appearing only in a part of the video shot), while reducing the bias introduced by the repetitive structures which might occur in static scenes or in scenes depicting buildings or tiled structures. In this respect we track interest points across consecutive frames and we consider for description only the new points/outliers. The principle is illustrated in Figure 3. Here, we depict the frames that we sample from a typical video shot and the interest points and corresponding descriptors that we consider for the next stages. The tracking of the interest points is performed by matching the RootSIFT descriptors from the frames, using Lowe's ratio test, followed by a fast spatial consistency check generating deformation hypothesis from the elliptical representations of every matched pair of points [4]. For the first frame we store all the points and for the following frames we consider only the non-matched points or outliers (marked with green circles), which are considered as new. The matched points /inliers (marked with blue circles) are rejected as they do not bring new information to our descriptor.

Let us notice that this approach also reduces the burstiness effects of repetitive structures in the BoW feature vectors, where a few artificially large components can dominate the similarity computed between BOW vectors, since the contribution of other smaller but important components is decreased significantly. This problem has been addressed first time in [13] and it is typically solved by a signed square rooting (SSR) normalization [14]. However, in recent work [11] it has been argued that the SSR normalization does not fully suppress the bursts, but only reduces them. We complement this drawback by discarding the inliers/repetitive points in the shot description followed by the SSR normalization of the feature vectors.

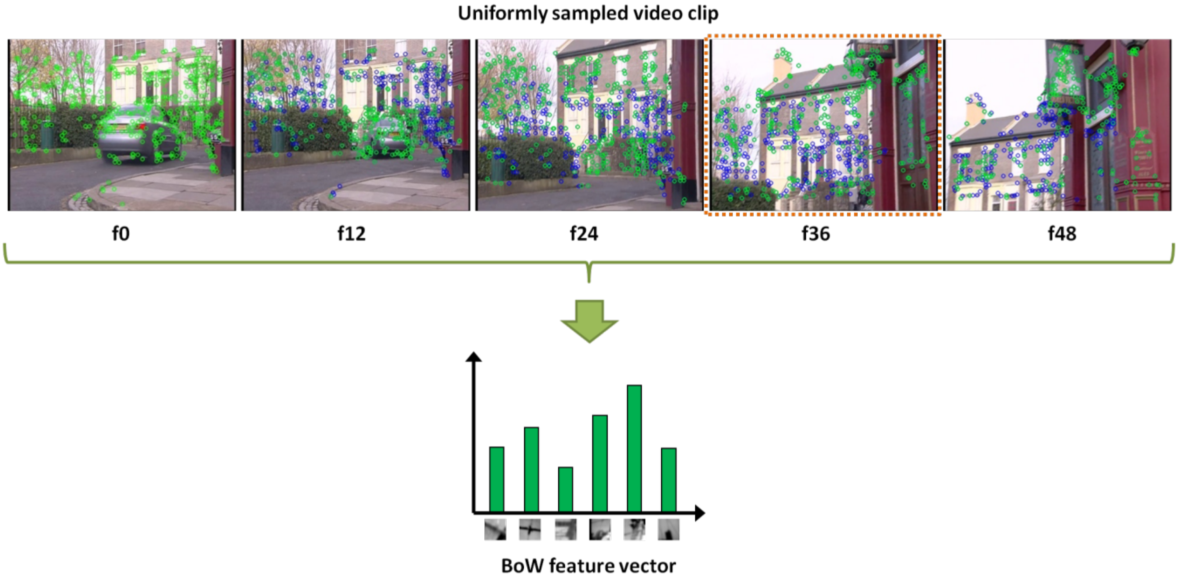

Figure 3. Shot based BoW quantization. Every 12th frame is considered for processing. The repetitive interest points (blue) are neglected, while the unique points (green) are kept and quantized into a single BoW feature vector. A representative frame is selected as the one having the highest number of interest points (dashed orange bounding box).

In order to obtain a rich description of the video clips we sample uniformly every 12th frame. The uniform sampling leads to approx. 1.4M frames to describe. In order to reduce the number of descriptors we resize the frames to a surface area close to 110592 (384 x 288) which has been used in the tasks of the previous years [15]. We then detect the Hessian Affine regions and extract the RootSIFT descriptors. We obtain a 365M regions and corresponding descriptors. 5% of the descriptors are randomly sampled from each video clip and then clustered in a vocabulary of 1M visual words with the approximate k-means method [4].

The difference between run 1 and 2 is that for run 1 we generate a single BoW feature vector for the query topics by quantizing descriptors from all query images belonging to the same topic. For run 2 we compute a BoW vector for every query image and then chose for every video clip the best score among the different query runs [15].

5 Frame-based description (runs 3 and 4)

The second quantization method considers a single representative keyframe from every video clip. This makes it possible to reduce the number of descriptors to compute and suitable for cases when the computational resources are limited.

One of the main drawbacks of such an approach is related to the artifacts of the video encoding, especially the motion blur. If the representative keyframe for a video clips contains motion blur, it would make the video clips practically non-retrievable. In order to alleviate this problem, we rely on the approach from Section 4 which extracts interest points from every 12th frame. We select the representative keyframe as the one having the highest number of local features which would correspond to the highest degrees of sharpness among the frames considered for the video clip. For the video clip from Figure 3, the representative keyframe is highlighted by a dashed orange rectangle.

For runs 3 and 4 we use the same description framework as in the previous runs (Hessian Affine and RootSIFT descriptors quantized into 1M words BoW feature vectors). Here we use the original size of the frames instead of resizing them. For all runs we employ the same visual vocabulary.

The difference between runs 3 and 4 is that for run 3 we generate a single BoW vector per query topic, while for run 4 we generate for each query 4 BoW vectors corresponding to the example images given for each topic.

6 Results

In this section we present some of the results of our runs on the INS task. In Figures 4-7 we show the average precision of our four runs versus the medina and best scores by topic.

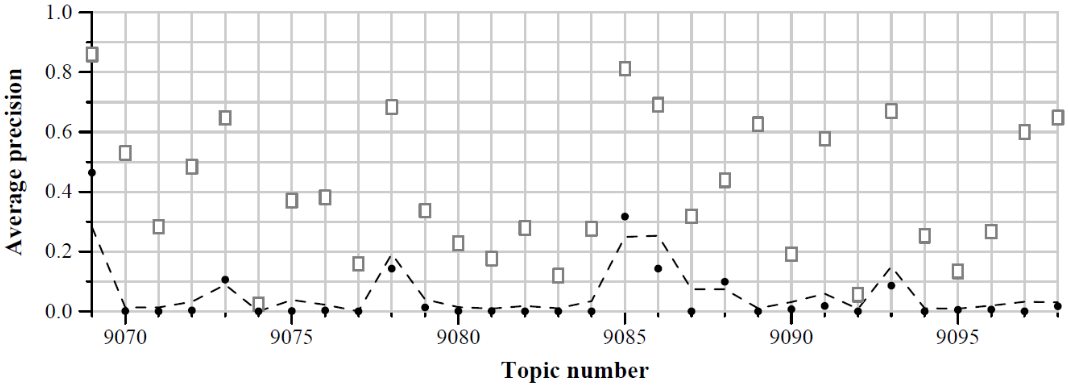

Figure 4. Average precision of run F_NO_ARTEMIS_1_1 (dot) versus median (---) versus best (box) by topic.

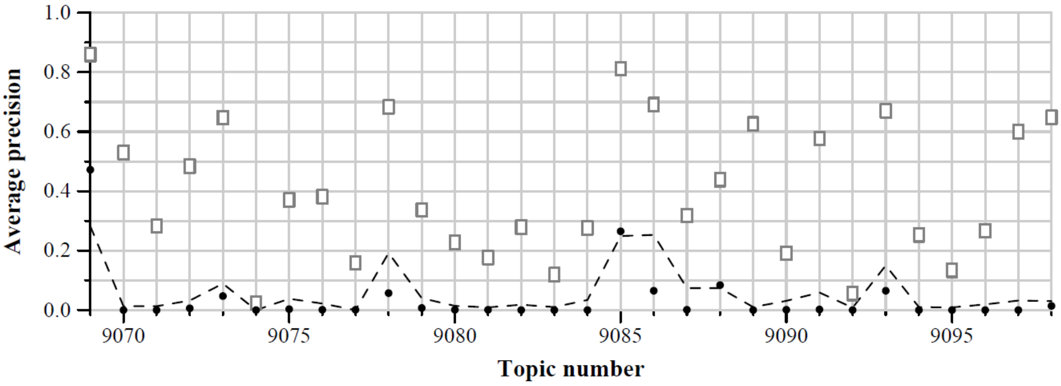

Figure 5. Average precision of run F_NO_ARTEMIS_2_2 (dot) versus median (---) versus best (box) by topic.

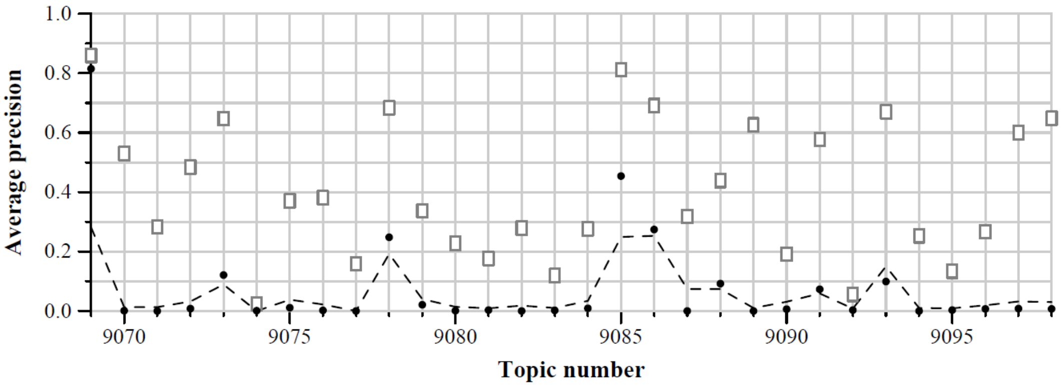

Figure 6. Average precision of run F_NO_ARTEMIS_3_3 (dot) versus median (---) versus best (box) by topic.

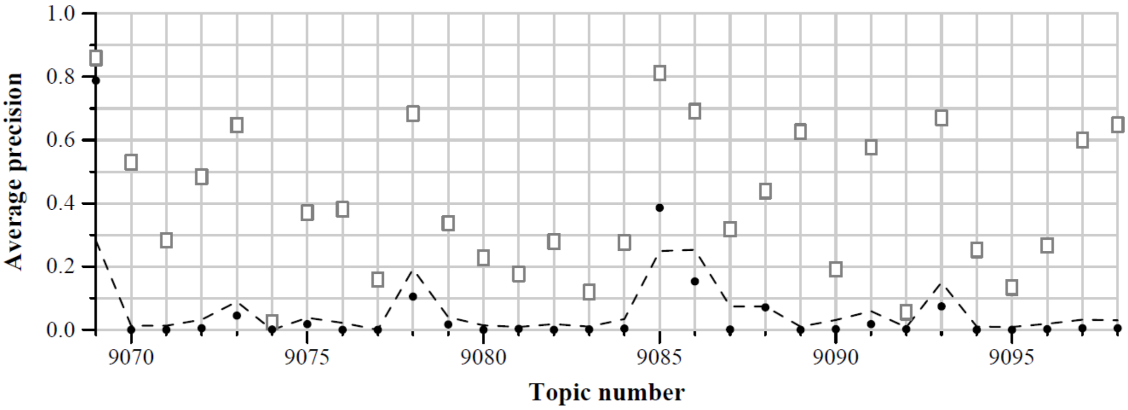

Figure 7. Average precision of run F_NO_ARTEMIS_4_4 (dot) versus median (---) versus best (box) by topic.

From these figures we can notice that the joint quantization of the example images into a single BoW feature vector per topic provides better results, even though there is no spatial consistency check and re-ranking. The frame-based representations perform better than the shot-based ones mainly due to the resizing performed for the latter. However, we suspect that a bug in the code might have affected the performances of runs 1 and 2.

Our system performed better for runs 9069 – *'no smoking' logo*, 9073 – *ceramic cat face*, 9078 – *JENKINS logo*, 9085 – *David refrigerator magnet*, 9086 - *scales*. The objects depicted in these topics are rather small.

7 Conclusion

In this paper we presented our experiments performed in the Instance Search of the TRECVID 2013 campaign. A system using a single representative keyframe per video shot provided good results. A novel approach for video clip description has been proposed and tested. While the results for this technique are less impressive, possibly due to some bugs, the approach promising.

The participation in the TRECVID campaign represented for us a rewarding experience in advancing forward our research and in finding new ideas and research directions in the challenging domain of object-based video retrieval.

8 Conclusion

We are grateful to the BBC for providing the *EastEnders* programme material ©BBC [2] for this task.

References

1. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton, and G. Quénot, "TRECVID 2013 – An overview of the Goals, Tasks, Data, Evaluation, Mechanisms and Metrics"; *Proc. TRECVID 2013*. NIST, USA, 2013.
2. BBC. http://www.bbc.co.uk/programmes/b006m86d.
3. J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos", *Proc. IEEE International Conference on Computer Vision*, 2003.
4. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.
5. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, Nov. 2005.

6. M. Perdoch, O. Chum, and J. Matas, “Efficient Representation of Local Geometry for Large Scale Object Retrieval,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

7. R. Arandjelovic, A. Zisserman, “Three things everyone should know to improve object retrieval,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2911-2918, 2012.

8. D. Lowe, “Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, Nov. 2004.

9. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, ”Lost in quantization: Improving particular object retrieval in large scale image databases,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

10. J. Delhumeau, P.H. Gosselin, H. Jégou, and P. Pérez, “Revisiting the VLAD image representation”, *Proc. ACM Multimedia*, 2013.

11. R. Arandjelovic, A. Zisserman, “All about VLAD,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

12. J M. Muja and D. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” *Proc. International Conference on Computer Vision Theory and Applications*, 2009.

13. H. Jégou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

14. H. Jégou, M. Douze, C. Schmid, P. Pérez, “Aggregating local features into compact image representation,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

15. A. Bursuc, T. Zaharia, O. Martinot, and F. Prêteux, “ARTEMIS-UBIMEDIA at TRECVid 2012: Instance Search,” *Proc. TRECVid Workshop*, 2012.