ARTEMIS-UBIMEDIA at TRECVid 2012: Instance Search Task

Andrei Bursuc[1,2], Titus Zaharia[1], Olivier Martinot[2], Françoise Prêteux[3]

[1]Institut Télécom ; Télécom SudParis, ARTEMIS Department, UMR CNRS 8145 MAP5,
9 rue Charles Fourier, 91011 Evry Cedex, France
{Andrei.Bursuc, Titus.Zaharia}@it-sudparis.eu
[2] Alcatel-Lucent Bell Labs France, route de Villejust 91620 Nozay, France
Olivier.Martinot@alcatel-lucent.com
[3] Mines ParisTech, 60, Boulevard Saint Michel, 75272, Paris Cedex, France
Francoise.Preteux@mines-paristech.fr

Abstract. This paper describes the approach proposed by ARTEMIS-UBIMEDIA team at TRECVID 2012, Instance Search (INS) task [1]. The method is based on the Bag-of-Words representation obtained from uniform sampling of the frames of the videos. We propose a query expansion technique that employs the textual description of the queries to identify new instances of the query objects on Flickr in order to enrich the query descriptor with additional representative instances.

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: 1 frame per second sampling from the videos, frames resized to 384x288 surface, Hessian Affine detectors and RootSIFT descriptor.
- **F_X_NO_UbiBWVTR_1:** BoW vectors generated at shot level. Single query BoW vector generated from the multiple example images.
- **F_X_NO_UbiBWVHF_2:** BoW vectors generated at shot level. Query BoW vectors generated from images fetched from Flickr using the provided query textual description.
- **F_X_NO_UbiBWFFM_3:** BoW vectors generated for each frame and for each query image. The score of the frame yielding the best score among the frames of a shot for a query is selected. After the querying the top 500 results are re-ranked with a color consistency check using MPEG-7 Color Structure the descriptor for the whole image queries and a region based object detection method for the partial image/object queries.
- **F_X_NO_UbiBWFFR_4:** BoW vectors generated for each frame and for each query image. The score of the frame yielding the best score among the frames of a shot for a query is selected. For multiple queries of the same topic, the best score of the video clip among the different query runs is selected.

What if any significant differences (in terms of what measures) did you find among the runs?

Overall, the grouping of the video frames in a single video clip descriptor has yielded the best results while reducing significantly the number of BoW vectors to be compared.
The Flickr-based expanded queries have provided satisfying results by employing images crawled from the internet and the query images (without using the binary mask).

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 the reduced number of detected interest points from the resized video frames.
The images collected from Flickr have improved the results for a number of topics which had less representative query images or reduced sizes of the query entities.

The aggregations of the results from multiple runs into a single result list can affect negatively the overall performance.

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 outperform the frame level BoW vectors, while reducing the number of vectors to consider for search.
This task has inspired us on the possibility of using textual descriptors to define a visual query to be then used to search objects in video content.

2 Approach Overview

For our approach, we have considered a large scale adapted Bag-of-Words representation [2] built on a vocabulary of 1M descriptors [3]. We identify the regions of interest with the Hessian Affine covariant region detector [4, 5] and describe each region with the recently introduced RootSIFT [6] descriptor; which is a SIFT [7] variant using a square root Hellinger kernel for the similarity measure. RootSIFT has yielded superior performances on the Oxford 5k and 105k and Paris 6k datasets [3,8].

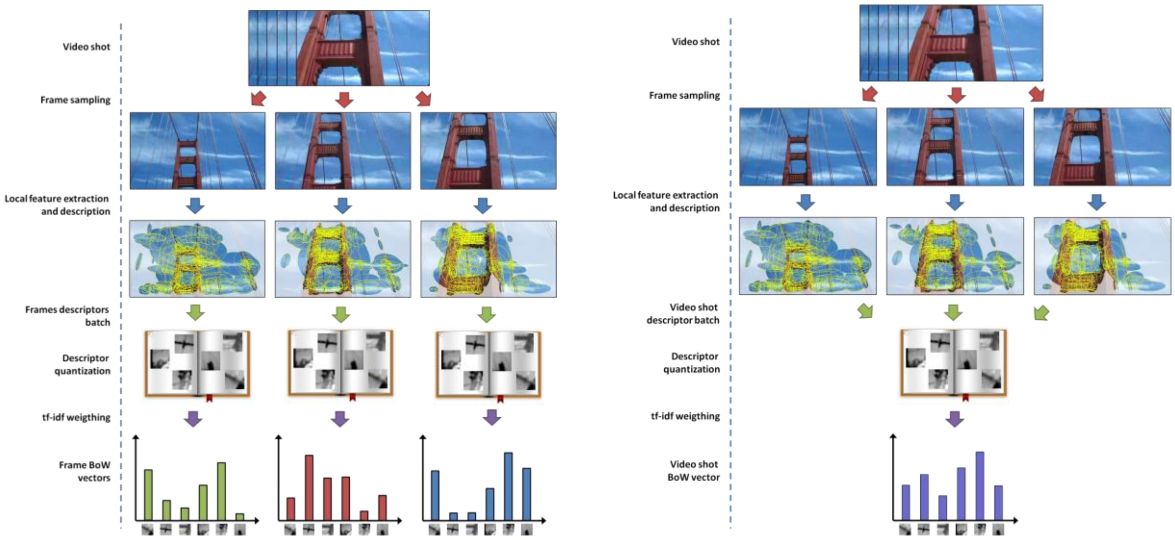
We consider an increased number of frames for the description of the video clips. We extract uniformly 1 frame per second and obtain approximately 683,433 key-frames. In order to reduce the number of descriptors we resize the frames to a surface area similar with 384x288 which has been used in the tasks of the previous years. We then detect the Hessian Affine regions and extract the RootSIFT descriptors. We obtain a 245,575,000 regions and corresponding descriptors. 10% of the descriptors are randomly sampled from each frame and then clustered in a vocabulary of 1M visual words with the approximate k-means method [3]. The largest and the smallest 5% of clusters are added to a stop list in order to discard the most frequent visual words which lack distinctiveness and the less significant ones.

We propose for testing two quantization strategies: frame-based and video clip based. For the former, the frames are quantized individually into BoW vectors. The principle is illustrated in Figure 1a. A video clip is described by multiple BoW descriptors corresponding to the frames sampled from the video. For each query, the frame yielding the best performance among the frames of the same video clip is selected and its similarity score becomes the score of video clip.

The second quantization method is similar with the one proposed in [9] and computes a single BoW vector for a given video shot/short clip. The approach has proven to be very effective in last year's campaign. The regions and descriptors detected in the sampled frames are then quantized together into a single BoW vector, associated to the entire video shot. This step reduces drastically the number of BoW vectors to be considered and hence boosts the computational speed in the search stage. Thus; the number of BoW vectors has been reduced from 683,433 (the total number of extracted frames) to 74,958 (the total number of video shots). The principle is illustrated in Figure 1b. The descriptors from all video frames are projected into a pool of descriptors along with the rest of the descriptors and a single BoW vector is computed after the quantization. Note that in this case, three views of the same object (*i.e.*, Golden Gate Bridge) are integrated in the same representation. This provides additional robustness to viewpoint changes.

Let us not that the sampled frames can be refined before quantization by rejecting the near-duplicate frames. This is highly useful for static video sequence, where successive frames are very similar and do not bring additional information. The near-duplicate frames can be computed quickly with global descriptors such as MPEG-7 ColorStructure descriptor [10] and color histograms [11] or with more advanced techniques employed for shot boundary detectors [12].

The chosen approach for this edition of the INS task [1] differs among the runs and we describe them separately.



a) b)

Figure 1 BoW quantization: a)frame based; b)shot based.

2.1 F_X_NO_UbiBWVTR_1

In this run the BoW vectors generated at shot level and similarly a single query BoW vector is generated from all descriptors collected from the query images under the provided object mask. This approach provided results which are around the median for most of the topics and outperforming it for topics 9051 and 9063.

2.2 F_X_NO_UbiBWVHF_2

In this run, we employ the same BoW vectors generated at shot level. The difference is in the query description.

We noticed that some queries defined object instances that contained little visual information about the objects, making them difficult to distinguish (*e.g.*,9048, 9064, 9068, 9055). While in such cases the, one can use the entire image for querying, the contextual information can prove be noisy for objects that typically occur in different environments (*e.g.*, logos).

Since we disposed of a textual description for each topic (e.g, 9048 – “Mercedes star”) we have tested the possibility of defining a visual query descriptor starting from these textual descriptions. We have thus launched 2 sets of queries on Flickr using the textual descriptions to search at the textual description level of the images and at the tag level. For each topic we have downloaded up to 50 images containing different instances of the respective object topic, but also multiple noisy images.

Figure 2 illustrates the results of a query performed on Flickr for retrieving the “Eiffel tower”. Notice that different instances of the object of interest are retrieved. Such results contribute to building a rich model of the Eiffel Tower visual query.

2.2.1 Interest point matching

In order to discard the false positive images and to identify the most representative images for a given topic, we first detect the local features and extract their descriptors by employing the Hessian-affine covariant region detector [5] and the RootSIFT descriptor [6]. All images from this set are matched one by one among themselves using the RootSIFT descriptor and Lowe’s ratio test [7] for selecting the reliable matches. The matched points are checked for geometric consistency using the fast spatial consistency check proposed by Philbin *et al.* [3]. We consider that two images contain the same object if they have at least 5 geometrically verified interest points successfully matched.

The role of this exhaustive matching procedure of all retrieved images is twofold. First, it makes it possible to reject the false positive images that have been retrieved (Figure 2). Usually such false positives are quite different from the rest of the true positive matches and they will be cleared out when matched with the rest of true positives. Second, as we can observe in Figure 2, the web search has retrieved different instances of the object of interest which are less likely to be matched using interest point matching. In this case, the role of the one to one matching is to identify groups of similar instances of the same object (e.g., Eiffel Tower seen from distance, Eiffel Tower photographed from one of it pillars.) in order to construct multiple query examples and descriptors.

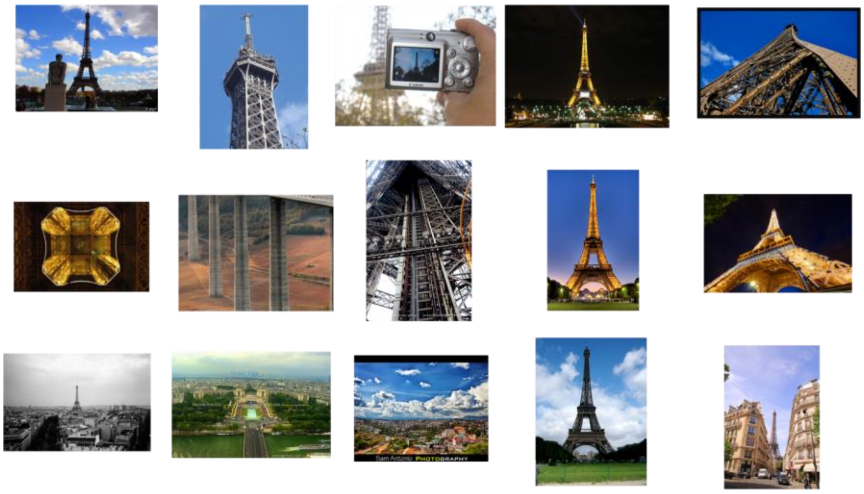


Figure 2 Flickr search results for "Eiffer tower". Notice that different instances of the Eiffel tower are retrieved along with a number of false positives.

2.2.2 Construction of the query graph

In order to identify the different instances of the query object, we employ the matching results from the previous step and construct a query graph similarly with the image graph introduced in [13]. The nodes of this graph are images and the edges connect images which have at least 5 geometrically verified matches.

An example of such a graph built over a set of 50 images is illustrated in Figure 3. The graph is computed from the first 50 images returned by Flickr's search engine. Note how the false positives from Figure 2 have been discarded, as such images have been identified as isolated nodes, with no edge to any other images in the data set. In addition, the number of images to be considered has been significantly reduced (half of the initial number of retrieved images).

In Figure 3 we can notice that images containing similar instances of the object of interest are strongly inter-connected. In addition, the clusters of inter-connected regions can be easily identified as connected components in the query graph. In the case of the query graph in Figure 3 we can extract 6 connected components, each consisting of images with similar instances of the Eiffel tower (*e.g.*, tower viewed from distance, pillar view, night view,…). The less representative instances are either completely rejected in the matching sequence or compose small connected components with poor interconnectivity.

2.2.3 Identification representative images from each connected component

While the number of images to consider for building a query visual descriptor has been reduced in the matching stage, the number of images is still high. In order to further reduce this number, we select only the most representative images for an object for each connected component determined.

The images from each connected component are ranked according to the number of matched images (i.e., edges in the graph). This measure corresponds to the degree (or valence) of each node within the graph.

For images having the same degree, the ranking order is made according to the total number of geometrically verified matches. This is the case of the purple, green and orange components in Figure 3, where several images have the same degree and the iconic image is selected by comparing the number of spatially verified matches cumulated over the all the matchings of the images. Thus, the retained representative image is the one yielding the highest number of verified matches.

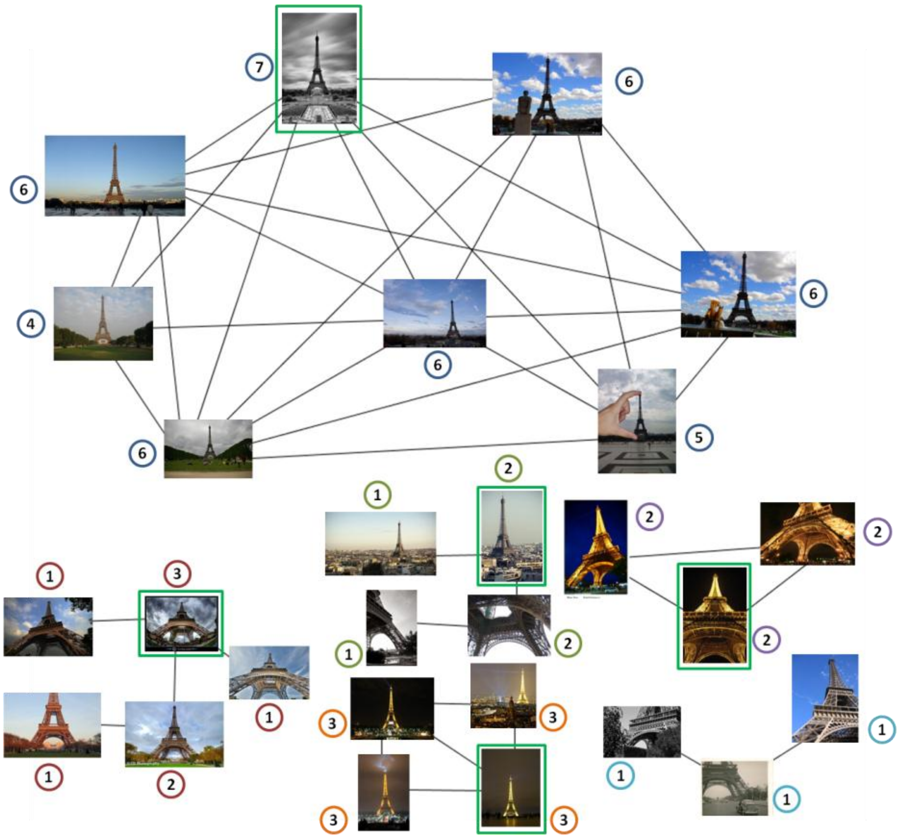


Figure 3 Query graph obtained for "Eiffel tower". Each node has marked its degree in a circle. The colors of the circle indicate the connected component which the current node is a part of. The iconic images of from each component are highlighted with a green bounding box.

In Figure 3 the iconic images are highlighted with a green bounding box. We can notice that the iconic images contain the most common views for the given object of interest. In addition, in order to avoid less representative images, we constrain them to have at least two verified matched images. The iconic images obtained for our example are illustrated in Figure 4.

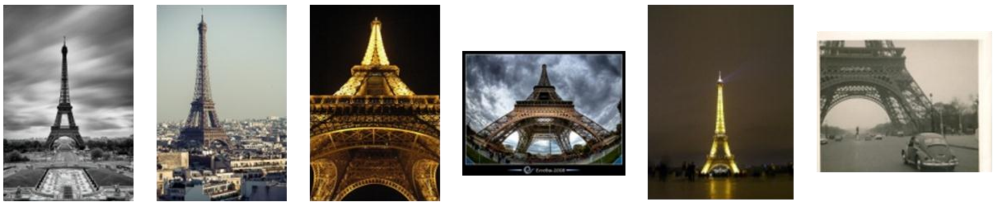


Figure 4 Iconic images for the "Eiffel tower".

2.2.4 Computation of query descriptors

We propose instead to exploit the information from the images that the representative images have been matched with. An iconic image can be thus described by its own features and by the features that have been matched with other similar images. The matched features complement the existing features and are used to enrich the current image. Practically, for each matched feature, when computing the BoW vectors, we assign a weight proportional to the number of verified matches. For example, for a feature that has been matched three times, its non-normalized *tf* weight is updated from 1 to 4. Alternatively, the representative image descriptor can be computed by considering the descriptors from the images matched with the representative image. These descriptors are collected in a pool of descriptors and quantized in a single BoW vector. We refer to this as the distributed representative query descriptor.

In Figure 5, we illustrate the feature matching between a representative image and its similar images from the same connected component of the query graph. Figure 6c illustrates the weighted centered representative image. The thickness of the elliptical shapes is proportional with the number of verified matches of the respective region. Complementary, the distributed representative image descriptor is composed from the point descriptors from all the images matched with the representative image. This weighting mechanism enriches the query descriptors and emphasizes the most representative features of the current object, increasing the chances of retrieving it accurately.

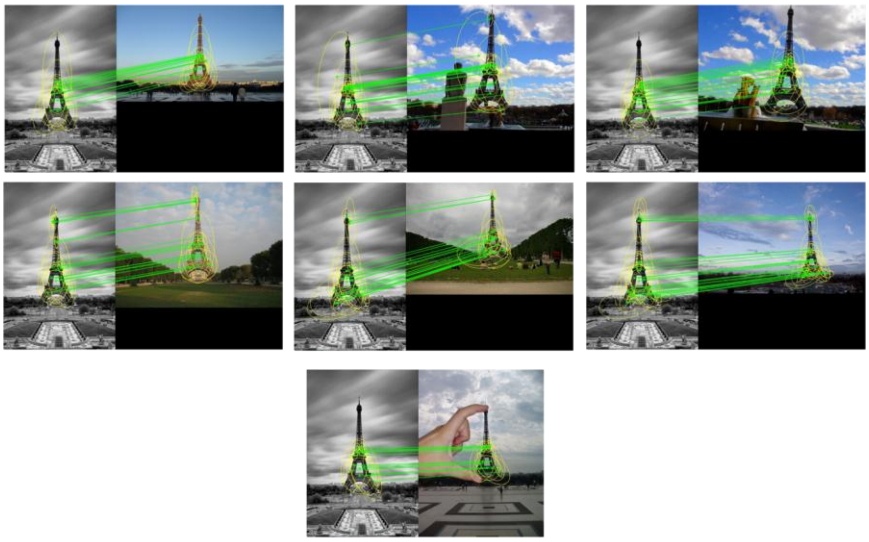


Figure 5 Representative image and its geometrically verified matches.

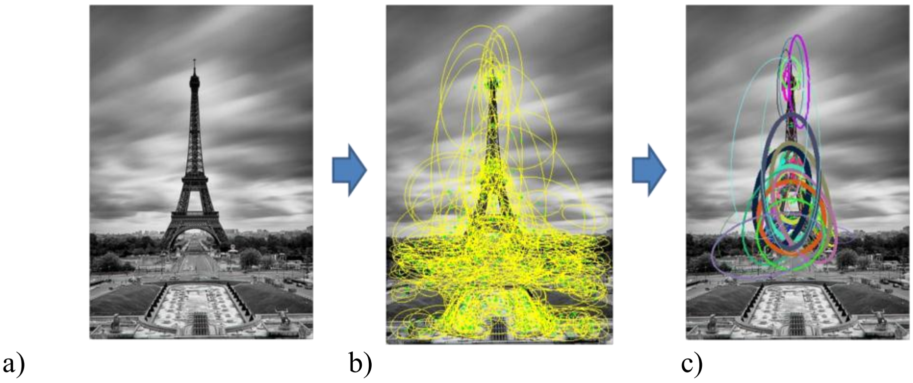


Figure 6 Enriched representative image with geometrically verified regions weighted proportionally with the number of matches from Figure 5. a) Original image, b) Detected Hessian-affine regions, c) Verified regions and their weights (the ellipse thickness is proportional with the weighting).

Once the descriptors of the representative images are computed we use them to launch dedicated queries among the shot based BoW vectors. For each video shot we select its best score among the all queries of the same topic.

2.2.5 Submitted run

We propose for testing, representative images descriptors identified from sets of up to 100 downloaded Flickr images. Let us note that since we have used the exact textual information provided, some queries could not be retrieved in more than a couple of instances (e.g., "puma logo animal") making the topic not-searchable in the dataset as not enough visual information was available. In other cases the general definition of the query has introduced multiple false positives in the query graph and affected negatively the score. For example, in the case of the topic 9056-"Pantheon interior" multiple images containing the "Paris Pantheon" have been retrieved and integrated in the query descriptor.

The query examples are considered in the computation of the query graph without the specification of the object mask. The identified representative images are then described in a distributed manner, by considering the descriptors of the images that with which it had successful matches.

Most of the results were good. Some query descriptors could not be defined as few images could be retrieved from Flickr using the textual description (e.g., "Puma logo animal", "Pepsi logo circle").

2.3 F_X_NO_UbiBWFFR_4

For this run, BoW vectors were generated for each frame and for each query image.
For each query image a dedicated search was issued. The score of the frame yielding the best score among the frames of the shot for a given query is selected. For multiple queries of the same topic, the best score of the video clip among the different query runs is selected.

The performances were lower in this case. This might me due to the different aggregation steps considered for computing a single list of results: selecting a best frame for each video, selecting the best score for each video among different runs. In addition, since some topics display rather different instances of the same object, the top results and the similarity scores for the different queries might be different. A top score for one of the queries can be a weak score for another one. The aggregation can then discard positive candidates that yield a weaker similarity score with one of the queries

2.4 F_X_NO_UbiBWFFM_3

This run has the same setting of the previous one (**F_X_NO_UBiBWFFR_4**). Before the results aggregation, the first 2000 retrieved frames are re-ranked using color-based descriptors. The query examples having full frame masks are re-ranked with the MPEG-7 Color Structure descriptor [10], while the other queries depicting objects as parts of the image are queried using a variant of the region-based object retrieval method proposed in [14].

The submitted results were subject to a bug that we identified after the submission. The correct results were slightly lower than the run **F_X_NO_UBiBWFFR_4.** However, for some of the poorly textured queries (9057, 9063, 9068) the color based re-ranking has improved the performance with up to 0.15 MAP.

3 Conclusion

In this paper we presented our experiments performed in the Instance Search of the TRECVid 2012 campaign. 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 retrie0val.

Acknowledgments. The current work has been developed within the framework of the UBIMEDIA Common Laboratory established between Institut TELECOM and Alcatel-Lucent Bell Labs France.

References

1. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. Smeaton, G. Quénot, "TRECVID 2012 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics", *Proceedings of TRECVID 2012*, 2012, NIST, USA.
2. J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos", *Proc. IEEE International Conference on Computer Vision*, 2003.
3. 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.
4. 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.
5. 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.
6. 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.
7. D. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, Nov. 2004.
8. 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.
9. C.Z. Zhu and S. Satoh, "Large vocabulary quantization for searching instances from videos," *Proc. 2nd ACM International Conference on Multimedia Retrieval, 2012.*
10. International standard ISO/IEC 15938-3:2002, Information technology - Multimedia Content Description. Interface - Part 3: Visual. 2002.
11. A. Bursuc, T. Zaharia, and O. Martinot, "ARTEMIS-UBIMEDIA at TRECVid 2011: Instance Search," *Proc. TRECVid Workshop*, 2011.
12. R. Tapu, T. Zaharia, "A complete framework for temporal video segmentation," *Proc. IEEE International Conference on Consumer Electronics*, 2011.
13. J. Philbin, J. Sivic, A. Zisserman, "Geometric Latent Dirichlet Allocation on a Matching Graph for Large-scale Image Datasets," *International Journal of Computer Vision*, vol. 95, no. 2, pp. 138-153, Nov. 2011.
14. A. Bursuc, T. Zaharia, and F. Prêteux, "Detection of Multiple Instances of Video Objects," *Proc. IEEE/ACM .International Conference on Signal Image Technology and Internet-based systems*, 2011.