Search of objects of interest in videos

Boris Mansencal, Jenny Benois-Pineau, Remi Vieux, Jean-Philippe Domenger

Search of objects of interest in videos

Boris Mansencal Jenny Benois-Pineau Remi Vieux
Jean-Philippe Domenger
LaBRI
CNRS UMR5800 - Université Bordeaux - IPB/ENSEIRB-Matmeca
351 cours de la Libération
33405 Talence cedex
France
{mansenca, benois-p, vieux, domenger}@labri.fr

July 16, 2012

Abstract

The paper addresses the problem of object search in video content. Both Query-By-Example paradigm and context search are explored. In QBE paradigm the object of interest is searched by matching of object signatures built from SURF descriptors with on-the-fly computed signatures in frames. The "context" search is understood as a query on the whole frame with features extracted after a region-based segmentation. Both kinds of features are transcribed in Bag-Of-Words framework. The combination of Bag-of-Visual-Words and Bag-of-Region-Words gives promising results in TRECVID'2011 Instance Search Task.

1 Introduction

Object retrieval in collections of images and videos is probably one of the key tasks in focus of attention of the community. The object search can be formulated as the search for a specific object - prototype or on the contrary, of an object category. In TRECVID international challenge an experimental task of Instance Search has been introduced since 2010 [13]. Here it is necessary to find the example object of interest presented in an example frame, in the set of available video clips. The difficulty consists in a very weak number of example objects generally very few frames - examples are given, excluding possibility of statistical learning of object models. Hence in this sense, the problem of object instance search is very close to the Query-By-Example paradigm in a general Content -Based Image Retrieval Task. Image retrieval is a topic that has been a research challenge for several years. The task is difficult for various reasons. Smeulders *et. al.* [19] introduced the concept of *semantic gap*, that is the discrepancy between the low level descriptors that can be computed from the images and the interpretation of the image done by humans. A query to an image retrieval system is ill-defined by nature. Such a query could take several forms. One of the earliest successful systems, QBIC[6], accepted queries as a user defined color palette, that images should match. A query can be formulated using an example image (Query-By-Example – QBE– paradigm). The system must retrieve the most similar images to the query. In this case, the notion of similarity is implicit for the user, and the system must "translate" this notion into a computable quantity. In the best case, it can be related to several measurements in terms of low level descriptors. In case of search of objects of a given category the problem is much more complex. Here the statistical variety of objects from a given category is used to train a global model and then to perform a classification task. In Instance search task, the Query-by-Example paradigm cannot be applied as is due to the strong variability of object appearance in the searched video content. Furthermore, the small amount of query instances for each object does not reasonably allow to perform a statistical learning of object category models. In the last decade, a breakthrough in image retrieval and object recognition have been achieved using the Bag-Of-Visual-Words (BOVW) model based on interest-point descriptors such as SIFT[10]. In the mean time, methods based on region-based properties of the image have known a decreasing of popularity for CBIR and classification tasks, since the fundamental work of Duygulu *et. al.* [4]. Few examples include Souvannavong *et. al.* [20] for video content indexing and retrieval and Gokalp and Askoy [7] for scene classification. However, current state-of-the-art for accurate object class image segmentation rehabilitates image segmentation and region-based visual description of the image content [8, 21, 23].

The present research is an attempt to combine both local features issued from keypoint detection and contextual features issued from segmentation of frames into regions. We restrict ourselves to an unsupervised context. The keyframes of videos in the database are only processed for keypoint or region extraction and quantization of relative descriptors. This quantization is done with a generic dictionaries computed on the whole database consisting of a large number of classes. We believe that an unsupervised context is closer to a majority of usage scenarii in object based video image retrieval. The only user intervention is relative to the query. He has to select the region of interest on the query frame, to delimit the searched object.

In the classical BOVW model [18], as signatures are histograms of words, spatial relationships between features are lost. Thus the location of the object of interest in the response to query images is not known. To improve the retrieval performance, Philbin et al [14] complement a BOVW model with a spatial verification stage. Despite their approach is applicable to the frame based queries or object based queries, the spatial verification step implicitly means that they try to find objects in a displaced position in the database images. Indeed they use a RANSAC algorithm to approximate a transformation model between the query region and the database image. With this spatial re-ranking, they consistently improve their results (5% of MAP).

On the contrary to their approach, we propose a method based on the BOVW as a sort of spatial correlation at the stage of BOVW comparison between query object and database frames. Hence object recognition and localization are simultaneously addressed. In real life applications, simulated by instance search task of TRECVID competition, the object size in the query can be insufficient for application of BOVW approaches and the variability of object appearance in the database can be extremely strong. In this case, the context is interesting to use to get a wider object appearance. Here we apply BORW model we developed in [22].

The paper is organized as follows. In section 2, we will present both object based and frame based visual signatures. In section 3, we will describe our method for similarity search. In section 4, we will present the results obtained on public dataset and in the framework of TRECVID Instance Search task 2011. In section 5, discussion and conclusion will be given.

2 Object based and frame based visual signatures

In the real life, an object of interest may be present in the database in the same context or background as it is in the query image. Hence to better address recognition of unknown visual scenes, we use both object signatures and frame signatures.

2.1 Object and frame BOFs

Let us consider a domain $D \subset I$ where I is the image support, and $X = \{x_1, ..., x_k\}$ a set of features computed in I. Note that in our application D is a pixel-wise object mask supplied with a query frame. If $\forall x_i, x_i \in D$ we call X object features set, and we call X frame features set otherwise. In the paper of Lowe [9] such feature sets are called "Bag-Of-Features" (BOF).

To compute feature sets we use two approaches. The first one is the SURF detector[2]. The second approach is a Bag-Of-Regions (BOR) model, as proposed by Vieux et al. in [22]. In this second approach regions in image plane are obtained by segmenting images by Felzenszwalb and Huttenlocher method [5].

Both SURF features and BOR approach may be used for object BOF and frame BOF computation.

Figure 1 illustrates these approaches. We present the information available for one query for instance 9026 of

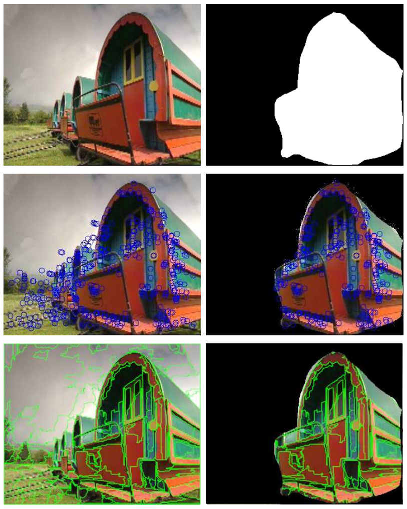

Figure 1: Example of instance from TRECVID test set: original image, mask, interest points on whole image, interests points on masked image, segmented regions on whole image, segmented regions on masked image.

TRECVID 2011: the original frame and the object of interest mask. The feature points or regions are extracted for the whole image or only on the object mask.

Note that in the following of the work, we will not use BOR for object BOF as the amount of regions on an object mask can be much smaller than the quantity of SURF points and hence not statically representative. Similarly if the number of features in an object BOF is sufficiently small, we will not use object BOF but the frame BOF.

2.2 Content descriptors

In case of SURF features, it is natural to use SURF descriptors which proved to be efficient in the state of the art literature [2].

For BOR approach, the global descriptor such as HSV histogram was computed, expressing color distribution. For this histogram, we set a uniform quantizing parameters in order to limit the descriptor size to approximately 100 bins and to privilege the finest encoding of the Hue component. This led us to 45+32+32 bins in the descriptor representing concatenated normalized marginal distributions. We note that HSV histograms of frames proved to be an efficient descriptor for video similarity search [3].

2.3 Descriptor quantization

For the huge databases the BOFs comparison between query object and frames is computationally heavy and unstable. This is why the Bag-Of-Visual-Words (BOVW) approaches [18] have received growing popularity. All the descriptors in the database are quantized with regard to the visual dictionary built by clustering of all the descriptors of all features in the database. Usually the K-means clustering approach [12] is applied [18, 17]. We used K-means++ [1] for SURF descriptors with 16K clusters, and incremental algorithm [11, 22] for region descriptors with 2K clusters, because of a smaller number of region features than those of SURF. For both clustering approaches, the L2 distance was used, as yielding smoother clusters shapes.

As result of quantization, for each object from query example images and for each query example image we compute the histogram of visual words, that is Bag Of Visual Words (BOVW) or Bag Of Region Words (BORW). In the following we will denote them by h.

3 Similarity search

In this section, we will present both object based similarity search and context based similarity search for object instance retrieval.

3.1 Object based similarity search

In the problem of Object retrieval in a video database, we can reasonably suppose that the example object can be found int database frames as it preserves its structural and contrast properties. That is its visual signature (BOVW) does not change a lot. Nevertheless it can appear in database frames in a transformed position. We will suppose here an affine transformation such as Pan/Tilt/Zoom (PTZ). The locus of object in the database frame is unknown. Therefore the method we propose consists in computation of a sort of correlation that is similarity of query object and potential object candidates in affine transformed image space. In practice it means that we transform the query object mask in PTZ parameters space and compute multiple visual signatures of database frame by scanning it with the transformed object mask. To measure the similarity between BOVW of query objects BOVW(q) and BOVW of database frame BOVW(b), we used the histogram intersection kernel [16]:

$$k_{BOVW}(h_q, h_b) = \sum_{i=1}^{N} min(h_q(i), h_b(i)) \quad (1)$$

and induced the distance $d(h_q, h_b) = 1 - k_{BOVW}(h_q, h_b)$.

Pan and Tilt parameters were chosen in such a way that query instance mask overlapped the DB frame at least of two thirds of its area. The Zoom factor was chosen from the set 0.25, 0.5, 1, 2, 4. The optimisation was fulfilled by full search in parameter space. This method is obviously more computationally demanding than the traditional BOVW. Indeed, signatures can not be computed in a processing step for all the images of the DB, but have to be computed on-the-fly in image area overlapped by image mask.

3.2 Context based similarity search

As we stated in the introduction the instance search task is more complex than Query-By-Example paradigm as the object can appear under strongly variable viewpoints and with changed appearance. In this case, the object based can be completed by context based search. We perform the latter by comparing visual signatures, BOVW and BORW on the frame basis. Hence we precompute the BOVW and BORW for all the frames in the database. In case of BORW the distance between visual signatures is the L1 norm of difference :

$$d_{BORW}(h_q, h_b) = \sum_{i=1}^{N} |h_q(i) - h_b(i)| \quad (2)$$

Finally to get the ranked list of database frames with regard to query object, we compute the similarity measure S

$$S(h_q, h_b) = \frac{1}{d(h_q, h_b) + \epsilon} \quad (3)$$

and perform fusion of results accordingly to the mean operator

$$S_{qb} = \frac{1}{2IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (S_{ij}(BOVW) + S_{ij}(BORW)) \quad (4)$$

Here I is the number of example images containing query object and J is the number of keyframes representing each content item in the database. If $I = 1$ and $J = 1$ we have a standard Query-By-Example paradigm. In case of instance search task in TRECVID, $I >= 1$ and $J >= 1$ as some frames containing object of interest are supplied for query and each video in the database is represented by several keyframes.

4 Experiments and results

We tested the proposed method on SIVAL dataset[15] and TRECVID 2011 Instance Search database.

The SIVAL database is a perfect dataset for object based Query-By-Example paradigm. This database contains 1500 images, with 25 objects pictured at different locations on the same set of complex backgrounds. Images are of 1024x768 resolution.

We used all the 1500 images for dictionary computation. The size of the dictionary for BOVW was 8K. For BORW, the size was 2K chosen accordingly to analysis in[22]. We then queried all the images against the whole database. We used the bounding box of objects for object based queries.

Table 1 shows results in terms of Mean Average Precision (MAP) of frame-based and object-based queries on this database. Besides, it also presents results for an ideal case, where bounding boxes of objects are used to compute the signature of images of database. This drastically reduces the influence of the background and thus is tailored specifically for this query-by-object experiment. However, this case is completely artificial, as it requires to have the whole database annotated. For this experiment BOVW results are better than BORW results. We can also see that, as expected, object based queries are far better than frame based queries for both methods, 0.0920 vs 0.1480 for BORW and 0.1501 vs 0.3326 for BOVW. Besides, BOVW with object based query with transformation is better than simple object based query. It is even better than ideal case for BORW. It stays inferior to ideal case

Method	MAP
BORW frame based query	0.0920
BORW object based query	0.1480
BORW ideal case	0.4221
BOVW frame based query	0.1501
BOVW object based query	0.3326
BOVW object based query with transformation	0.4512
BOVW ideal case	0.5123
RANSAC	0.3509

Table 1: Results on whole SIVAL database

for BOVW. Note that the Pan/Tilt/Zoom factors used for transformations do not cover the whole parameters space.

We have also computed the results of classical method based on feature matching coupled with RANSAC algorithm [14] to find a transformation between the object based query and the database image. This result is noted RANSAC in table 1. We can see that BOVW with object based query with transformation outperforms this method on SIVAL database.

The Figure 2 shows first results of object based query with transformations for one kind of object, an AJAX bottle, from the SIVAL database. It is noteworthy that this object is challenging as it is partially translucent. We can see that the bottle was matched at different positions, with different scales, on different backgrounds than the query image.

In TRECVID 2011 Instance Search task, four types of entities to search for were proposed: person, character, object, location. Each entity was represented by a few example frames with object masks. The task consisted in searching for video in the database containing the instances of the entities. Each video clip was represented by several keyframes. Table 2 shows the distribution of instances by type and the number of examples for each type. CHARACTER was not in the database. Test data set was composed of rushes, that is raw, unedited data, of BBC series or documentaries.

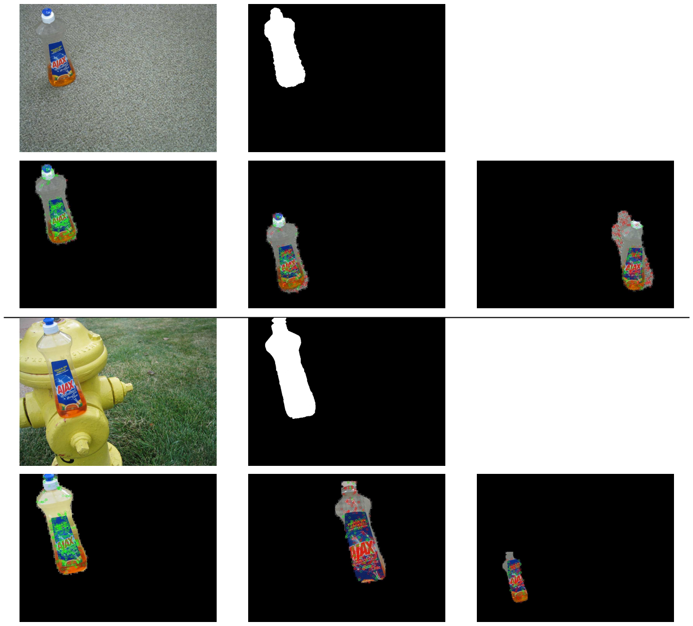

Figure 2: Examples of queries and matched transformed masks

type	overall number of examples	number of different instances	mean number of examples per instance
PERSON	28	6	4.67
OBJECT	62	17	3.65
LOCATION	5	2	2.5
total	95	25	3.8

Table 2: Distribution of instances for test set 2011

This kind of data often contains several takes of the same scene, maybe with a different camera angle. We expected that images between these takes could be quite similar. Hence the use of context such as global BORW signatures for example frames and DB was justified. Furthermore, if object based query is considered, the mask of query could be small. This would entail too few points inside. Therefore, in our runs, we limited the use of the mask for the query only if enough SURF features were detected inside. After studying query images and available object masks, we have decided to use object signatures only if we had at least 8 interest points detected.

We have computed four results: BOVW for the whole frame, BOVW for object based query supposing the object in DB frames was of approximately the same size and at the same position as in query example, BOVW for object based query with affine deformation, BORW for the whole image. These results are computed for all keyframes. The database was then composed of 40K images, at mainly CIF resolution (352x288pixels). Finally, we have submitted four fully automatic runs:

- run1: we merge BORW and BOVW both for the whole frame.
- run2: if we have enough points of interest in query, we merge BORW for the whole frame and BOVW results for object based query with affine deformation. Otherwise, we keep only BORW for the whole frame.
- run3: if we have enough points of interest in the query, we merge BORW for the whole frame and BOVW results for object based query without affine deformation. Otherwise, we keep only BORW for the whole frame.
- run4: pure BORW for the whole frame.

There were 37 fully automatic runs submitted this year. Table 3 presents our results for the different runs, for the various instances for topics 9043 and 9039, and on average. Topics 9043 and 9039 are the topics for which run1 produces respectively the best and the worst results. Features matching coupled with RANSAC method is also present for reference, altough it was not submitted and thus is not ranked.

Generally:

- Our runs sorted from best to worst are : run1, run3, run4 and run2.
- All four runs are better than median.
- run1, run3 and run4 are in the first third of the sorted results.
- The fact that run3, object based query without affine deformation, outperforms run2, object based query with affine deformation is surprising. It seems related to our too coarse exploration of the parameters space for the affine transformation. This has to be further investigated.
- Features matching coupled with RANSAC gives worse results on average than all runs. However on topic 9039, it produces better results than all 4 runs.

It is noteworthy that TRECVID videos resolution and quality of extracted keyframes is strongly inferior to SIVAL images. Thus we get less points per image and they may be impacted by poor image quality. It may explain why object-based query results are not so good on this dataset.

Besides, the tested size of vocabulary is rather low. In [14], a dictionary of 1M words gave the best results for 16.7M features. Here, there are roughly 18M points in keyframes of this TRECVID dataset, and the dictionary use was only of 16K words. We plan to investigate larger vocabulary size, although clustering for large vocabulary is also challenging. The last, but not least for the performance evaluation is the instability of the feature-point detector with regard to scale transformations and the poor quality of frames when Shannon theorem is not respected (aliasing generating ”false points).

Our BOVW object base query with transformation method is obviously more computationally intensive than the traditional BOVW method. But features matching coupled with RANSAC method can also be rather heavy, even when features are filtered by Hough transform as in [10]. In our experiments, both methods took roughly the same time on SIVAL database, but were around 122 times slower than the BOVW method. On TREC database our BOVW with transformation was around 15 times slower than RANSAC method, and around 200 times slower than the simple BOVW method. A fast approximate search in the parameter space (gradient or bisection) could improve the time complexity.

4.1 Discussion and conclusion

In this paper, we proposed a method for object instance retrieval in image and video databases. Both object based visual signatures and context based visual signatures were designed in the framework of classical BOVW and recently proposed BORW approaches. Fusion of approaches was also elaborated to satisfy the ambiguous nature of instance search task of TRECVID 2011 challenge.

In our opinion, the choice of optimal approach: BOVW for object based query, whole frame based query, BORW, or their combination is very much dependent on data. Indeed for SIVAL dataset, we obtained better results for the object based query with affine deformation than for object

topic	run1		run2		run3		run4		RANSAC
	map	rank	map	rank	map	rank	map	rank	
9043	0.4994	1	0.2769	8	0.4910	3	0.4971	2	0.0627
9039	0.0462	18	0.0444	20	0.0469	17	0.0367	21	0.1093
mean	0.2735	10.36	0.1662	16	0.2588	11.44	0.2511	11.96	0.1151

Table 3: Results for 4 runs on test set 2011, for topics 9043 and 9039 and on average.

based query without affine deformation.

We stress that our approach is totally generic. We do not use the knowledge that some instances represented persons for example. All the queries are considered containing generic objects.

In order to increase the performances of our approach we will have to more finely investigate the application of object or frame based signatures, and take into account the fact that in case of strong affine transformations keypoint detectors which are supposed to be invariant fail. Hence the intelligent combination of global object descriptors with local ones has to investigated.

From the computational cost point of view, we can also improve the object based queries by an optimal traversal of parameters space with a coarse-to-fine strategy.

Besides, theses searches with different transformation parameters can be parallelized and thus this method is a good candidate for heavy parallelisation, on a GPU for example.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *SODA07*, pages 1027–1035, 2007.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *9th European Conference on Computer Vision*, 2006.

[3] E. Dumont and B. Merialdo. Rushes video summarization and evaluation. *Multimedia Tools and Applications, Springer, Vol.48, N1, May 2010*, 2010.

[4] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV'02*, pages 97 – 112, 2002.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004.

[6] M. Flickner, H. Sawhney, W. Niblack, J. Ashely, Q. Huang, B. Dom, M. Gorkani, J. Hafner, denis Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the qbic system. *IEEE Computer*, 28(9):23–32, 1995.

[7] D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *CVPR'07*, pages 1–8, 2007.

[8] L. Ladicky, C. Russel, and P. Kohliwu. Associative hierarchical crfs for object class image segmentation. In *ICCV09*, 2009.

[9] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.

[11] E. Lughofer. Extensions of vector quantization for incremental clustering. *Pattern Recognition*, 41:995–1011, 2008.

[12] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, page 281297, 1967.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR2007*, 2007.

[15] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts. Localized content based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 30:1902–1912, 2008.

[16] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[17] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV 2005*, volume 1, pages 370 – 377 Vol. 1, oct. 2005.

[18] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV'03*, volume 2, pages 1470–1477, 2003.

[19] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[20] F. Souvannavong, B. Merialdo, and B. Huer. Region-based video content indexing and retrieval. In *CBMI05*, 2005.

[21] J. Tighe and S. Lazebnik. Superparsing: Scalable non-parametric image parsing with superpixels. In *European Conference on Computer Vision*, 2010.

[22] R. Vieux, J. Benois-Pineau, and J.-P. Domenger. Content based image retrieval using bag of region. In *MMM 2012 - The 18th International Conference on Multimedia Modeling*, 2012.

[23] R. Vieux, J. Benois-Pineau, J.-P. Domenger, and A. Braquelaire. Segmentation-based multi-class semantic object detection. *Multimedia Tools and Applications*, Available Online Oct. 2010:1–22, 2010.