Multi-modal query expansion for video object instances retrieval

Andrei Bursuc and Titus Zaharia
Institut Mines-Télécom ; Télécom SudParis, ARTEMIS Department, UMR CNRS 8145 MAP5
{Andrei.Bursuc, Titus.Zaharia}@telecom-sudparis.eu

Abstract

In this paper we tackle the issue of object instances retrieval in video repositories using minimum information from the user (e.g., textual description/tags). Starting for a set of tags, images containing the object of interest are crawled from popular image search engines and repositories (e.g., Bing[1], Fickr[2], Google[3]) and the positive and most representative instances of the object are automatically identified. These positive images are then used to generate a visual query descriptor and to retrieve videos containing the object of the interest. This multi-modal approach makes it possible to retrieve video content through images obtained from textual queries, without the use of any advanced learning technique. We test out method on the Flickr corpus of the TRECVID 2012 Instance Search Task.

1. Introduction

The retrieval of different instances of the same object in video repositories is among the most challenging tasks in the field of computer vision applications. Variations in visual appearance and object's pose have to be taken into account appropriately. Typically, such variations are covered within a complete and well annotated set of positive images that it is used for training a classifier over multiple descriptors and search strategies. Since, the amount of positive sets is limited, such approaches usually tackle on recognizing general object categories and classes which are pre-defined [1]. In the case of object instances retrieval, the number of possible object entities that users might want to look for is enormous and the development of dedicated learning sets for each of them is impossible. Different strategies need to be identified for such problems. This relatively recent topic of research has been considered in the TRECVID evaluation campaign, under the so-called Instance Search Task (INS) [2].

Meanwhile, the popular image search engine are becoming increasingly powerful and provide multiple and various positive results for the majority of textual queries performed. In addition, users are increasingly accustomed to perform textual queries naturally for all their search operations (*e.g.*, documents, shopping items, images, videos ...). In our work, we want to leverage on the textual information from the user in order to identify a large variety of objects without the use of a learning stage. In this respect we employ the image search engines to identify a small set of possibly positive images and use them to generate a new visual query to be used to retrieve relevant video content.

2. Related work

The underlying principle of query expansion methods can be stated as follows. An initial user-selected query is formulated, returning a list of results. Then, the first ranked documents are selected from this list and used to construct a new, richer query, which contains additional terms, relevant to the intended query. The enriched query is then re-issued and the results returned to the user. If the initial query is sufficiently successful (*i.e.*, if noisy inliers and false positives are not added to the new query), the technique can boost the recall of the retrieval, and as a collateral effect, its precision.

In [3] the query expansion paradigm has been for the first time exploited within the context of object-based image retrieval using the Bag-of-Words representation (BoW) [4]. Here, the query is enriched with geometrically verified points from the top ranked results obtained by performing a RANSAC consistency check [5]. A set of new BoW vectors are generated from the verified matches and the new query vector is obtained by averaging the initial query vector with the verified BOW vectors. A new enriched query, embedding information that might have been missed in the initial query (*e.g.*, different object pose), is then presented to the system. The obtained results display a significantly higher recall rate. Up to 200 top results are considered during the expansion and the technique can be iterated several times.

An improved version of this method has been recently proposed in [6]. Here, the geometric verification and the re-ranking build a statistical model of the query object. In addition, the relevant spatial context is learned in order to further improve the retrieval performance.

Arandjelovic and Zisserman [7] introduce another improvement of the query expansion from [3], so-called discriminative query expansion. Once the expanded BoW vectors computed, they are used as positive training data to train a linear SVM classifier. Images with low *tf-idf* similarity scores are used as negative examples in the learning process. The weights learned by the classifier from the small training dataset are then used to re-query the database and to rank images according to their distance from the decision boundary.

The main drawback of such methods is related to their dependency on the quality of the first query considered and of the first retrieved results. If no good matches are found in the first result set, the query can be altered and the recall rate will be degraded.

We propose a novel approach for query expansion that does not rely on the first retrieved results, but instead

[1] http://www.bing.com
[2] http://www.flickr.com
[3] http://images.google.com

leverages on the vast amount of manually annotated content from image portals and search engines such as Flickr and Google Images. Relevant content can be retrieved with text-based queries and the retrieval results and data from these portals can be accessed automatically through their APIs. In our approach, we employ such data in order to enrich the existing query object or to generate a set of visual queries starting from a textual description of the query object.

Our approach is based on the assumption that an elementary textual description is available for each query and is inspired by a use-case considered in the TRECVID 2012 INS Task [1]. Here, a collection of queries consisting of multiple example frames (2-9 examples) containing the object of interest is provided. In addition, each query includes a simple textual description of the object to be retrieved (*e.g.*, *Brooklyn bridge tower*, *Hagia Sophia interior*). We propose to leverage on these textual descriptions (which contains keywords commonly used for a textual search) in order to retrieve additional content that can be exploited for enriching the given queries. Let us note that, in a more general case, the given visual queries can be completely discarded and a visual query can be constructed solely from a set of textual descriptions. Let us also underline the multimodal character of this query technique: text descriptions are used for retrieving images, which are used at their turn for retrieving video content.

The approach consists of the following successive steps:

i. Issue a text based query on a web image search engine and download the first $N_{download}$ results.
ii. Extract local features from the retrieved images and perform a one-to-one, exhaustive interest point matching, together with the geometric consistency verification between all images. Build then a query graph using the retrieved images as nodes. The graph edges are used here to connect images consistently matched. Identify different instances of the query object as connected components in the graph.
iii. Determine the most representative image from each connected component,
iv. Build query descriptors using the representative images together with their best matches,
v. Aggregate query descriptors, and formulate queries on the video dataset in order to retrieve relevant videos.

The various steps involved are described in details in the following sections.

3. Text-based queries on web search engines

Most of the general public image search engines provide public APIs which allow automatic querying and download of results. In a first stage, we employ the Flickr API which is the most popular choice for many datasets.

Figure 1 illustrates the results of a query performed on Flickr for retrieving *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. Concerning the number of images $N_{download}$ to consider and download from the result list, we experiment with values between 25 and 100.

In the same time, a certain number of outliers are also retrieved due to annotations error. Such outliers are filtered out in the following step.

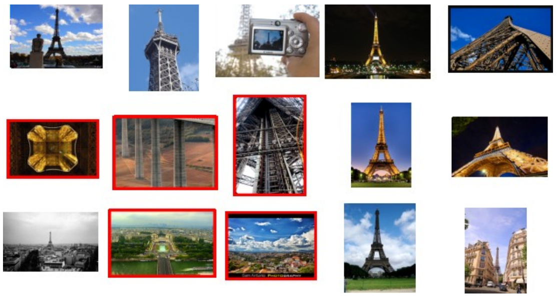

Figure 1. Flickr search results for *Eiffel tower*. False positives are highlighted in red.

4. Query graph

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 [8] and the RootSIFT descriptor [7]. All images from this set are matched one by one using the RootSIFT descriptor and Lowe's ratio test [8] for selecting the reliable matches. The matched points are checked for geometric consistency using a fast spatial consistency check 1.[10]. We consider that two images contain the same object if they have at least M_{min} geometrically verified matched interest points.

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 1). 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 1, the search engine has retrieved different instances of the object of interest (*e.g.*, different viewpoint, different weather conditions) 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, night view, …) to be further used as query examples.

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. The nodes of this graph are images and the edges connect similar images which have at least M_{min} geometrically verified matches.

An example of such a graph is illustrated in Figure 2. This graph is computed from the first 50 images retrieve by Flickr for "Eiffel tower". Note how the false positives from Figure 1 have been discarded, as such images have been identified as isolated nodes with little similarity to other images in the data set. In addition, the number of images to be considered has been significantly reduced, since we keep only the images from the graph.

In Figure 2 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 from Figure 2, we can extract 6 connected components, each consisting of images with similar instances of the Eiffel tower. The less representative instances are either completely rejected in the matching sequence or compose small connected components with poor interconnectivity.

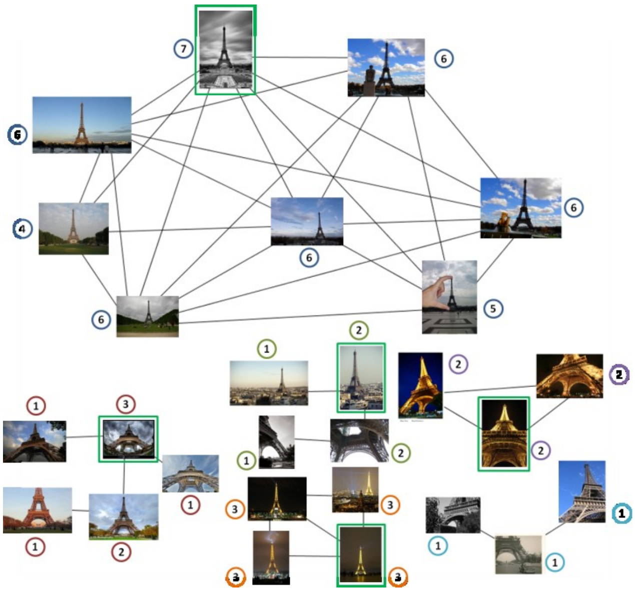

Figure 2. Query graph obtained for *Eiffel tower*. Each node has marked its degree in a circle. The colors of the circle indicate the connected component which the current node is a part of. The representative images of from each component are highlighted with a green bounding box.

We can notice that the representative 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.

5. Enriched query descriptors

Concerning the description of the expanded query, we propose to exploit the information from the images that the representative images have been matched with. A representative 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. We refer to this as the centered query descriptor. 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 3, we illustrate the features matched between a representative image and its similar images from the same connected component of the query graph. Figure 4c illustrates the weighted centered representative image. The thickness of the elliptical shapes is proportional with the number of verified matches of the respective region. These weighting mechanisms enrich the query descriptors and emphasize the most representative features of the current object, increasing accuracy of the retrieval.

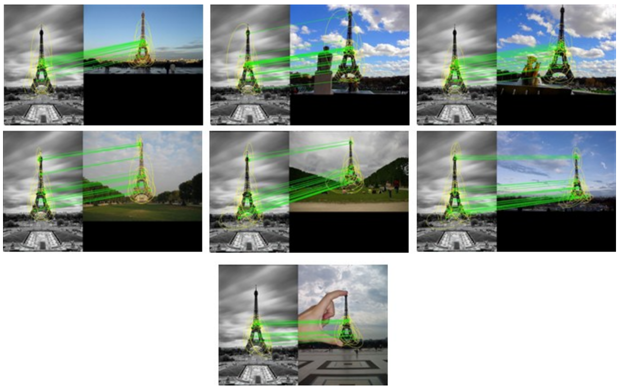

Figure 3. Representative image and its geometrically verified matches

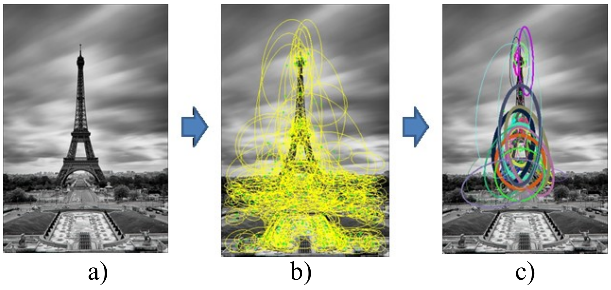

Figure 4. Enriched representative image with geometrically verified regions weighted proportionally with the number of matches from Figure 3. a) Original image, b) Detected Hessian-affine regions, c) Verified regions and their weights (the ellipse thickness is proportional with the weighting).

6. Experiments and results

We test our method on the *Flickr* corpus considered for the INS Task in TRECVID 2012. The corpus consists of 74,958 general public video clips downloaded from Flickr and summing up to 190 hours of content. A set of 21 query topics have been proposed for this dataset in the 2012 campaign [2]. We evaluate the performance, in terms of Mean Average Precision (MAP) computed over the 21 query topics.

We have sampled uniformly 1 frame per second from each video, resized them to a surface area close to 384x288 and then detected the Hessian Affine regions [8] and extracted RootSIFT descriptors [7]. For each video clip we have computed a BoW vector by quantizing all descriptors from its corresponding keyframes. Concerning the minimum number of verified matched interest points (M_{min}) for two images to be connected with an edge in the query graph, we have obtained better results for M_{min} ranging from 5 and 10. Here, M_{min} is set to 5.

We test both the centered and the representative query descriptors. In addition we test the effectiveness of using a single query descriptor by merging all the individual descriptors of the representative images into a single BoW vector. Such a representation will reduce significantly the computational cost as only one query needs to be performed instead of multiple queries for each of the representative images. In the case of non-merged vectors, we issue separate queries for all representative images descriptors and for each video we select its best score among all query runs.

The results are displayed in Table 1. We can notice that the distributed descriptors perform better due to the increased amount of information given by the descriptors from multiple images. The performance increases with the size of image set crawled from Flickr. In addition the merged BoW vectors perform better than the individual ones while performing a single search operation on the dataset. All results are comparable with the median result of the TRECVID 2012 INS submitted results. Let us note that the median result is obtained from runs using original query images from the *Flickr* corpus with query objects defined at the pixel level. In our case, solely textual queries have been used, with no visual information related to the considered queries, making this result remarkable. Moreover, the textual descriptions exploited are quite elementary [2].

Table 1. Query expansion using images from Flickr.

Expansion method	Image set	Merged BoW vectors	MAP
Centered representative query	25	No	0.0455
		Yes	0.0476
	50	No	0.0585
		Yes	0.0583
	100	No	0.0689
		Yes	0.0688
Distributed representative query	25	No	0.0540
		Yes	0.0558
	50	No	0.0756
		Yes	0.0787
	100	No	0.0871
		Yes	**0.0967**
TRECVID 2012 Median MAP			0.0795

Following, we test the generalization of our method to multiple search engines (*e.g.*, Bing, Flickr, Google). We have downloaded an even number of images from all engines and constructed a single query graph. The results are illustrated in Table 2. We can observe that the scores are slightly lower comparing to the previous runs. This is due to the fact that the search engines return various image types in the top list (*e.g.*, graphics, painting …), inserting outliers in the query graph which are more difficult to handle. In addition, duplicate images can be retrieved by different engines and all their descriptors, including the outliers are then included in the query graph.

Table 2. Query expansion using images from multiple search engines (Bing, Flickr, Google).

Expansion method	Image set	MAP
Distributed representative query	30	0.552
	45	0.0596
	60	**0.0725**

Further, we wanted to illustrate the pertinence of the newly computed query descriptors and performed a run using the original queries of the Flickr corpus and the new descriptors. We compare it against a baseline run using only the original TRECVID queries. The improvement in the results is up to 3% in MAP as seen in Table 3.

Table 3. Query expansion descriptors added to original TRECVID 2012 queries.

Image source	Image set	MAP
Flickr	50	0.121
Multiple sources	30	0.103
	45	0.107
	60	0.117
Baseline run		0.095

7. Conclusion

In this work we have introduced a novel multi-modal query definition and expansion method that makes it possible to retrieve objects in video content from simple tags. Positive instances of the objects are automatically identified from the Internet and used for the construction of a visual descriptor to be employed for retrieving video content.

References

[1] C.G.M. Snoek, M. Worring, “Concept-Based Video Retrieval”, Foundation and Trend in Information Retrieval, vol.2, no.4, pp. 215-322, 2008.

[2] 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”, *Proc. of TRECVID 2012*, 2012, NIST, USA.

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, “Total recall: Automatic query expansion with a generative feature model for object retrieval,” *Proc. IEEE International Conference on Computer Vision*, 2007.

[4] J. Sivic, and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos”, Proc. *International. Conference on Computer Vision*, 2003.

[5] M.A. Fischler and R.C. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”, *Comm. of the ACM,* vol. 24, no. 6, pp. 381–395, June 1981.

[6] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, “Total recall II: Query expansion revisited,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

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

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

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

[10] 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.