ARTEMIS-UBIMEDIA at TRECVid 2011: Instance Search

Andrei Bursuc[1,2], Titus Zaharia[1], Olivier Martinot[2]

[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

Abstract. This paper describes the approach proposed by ARTEMIS-UBIMEDIA team at TRECVID 2011, Instance Search (INS) task. The method is based on a semi-global image representation relying on an over-segmentation of the key-frames. An aggregation mechanism was then applied in order to group a set of sub-regions into an object similar to the query, under a global similarity criterion.

1 Introduction

Object retrieval in videos is among the most challenging tasks up to date in computer vision. In the last years an increasing number of solutions have provided a variety of satisfying results for concept detection in video [1]. Yet, retrieving different instances of the same object in video sequences still remains an open issue. The main difficulty is related to the specification of semi-global image representations that need to be considered, together with the elaboration of efficient partial matching taken into account appropriately. Existing methods for object indexing and retrieval such as the discriminatively trained deformable-part based models [2] or the bag-of-words representations [3] cannot be applied successfully in all cases as they rely on classification and machine learning methods. While for specific object category retrieval the results of such techniques are encouraging, it is difficult to train an algorithm for any object the user might want to search for. This relatively recent topic of research has been considered in the TRECVID 2010 evaluation campaign [4], under the so-called instance search task, and TRECVID continued it in the 2011 edition.

This paper describes the work of ARTEMIS-UBIMEDIA in the Instance Search Task of the TRECVID 2011 campaign. In the following sections we will present in detail the applied algorithms and the evaluation for the runs we performed.

2 Instance Search Task Presentation

Instance Search (INS) is a pilot task introduced in the TRECVid 2010 campaign and continued in the 2011 campaign. Given a collection of test video clips and a collection of queries that delimit a person, object, or location in some example video, participant applications have to locate for each query up to 1000 clips most likely to contain a recognizable instance of the entity. A number of 25 queries have been specified, each consisting of a set of 2 to 6 example frame images drawn at intervals from video containing the item of interest. The BBC rushes dataset was proposed for this task with a total of 20982 short clips. Different transformations were applied to some random test clips in order to increase the difficulty of the task.

The main objective was to explore task definition and evaluation issues. It is only a rough estimate of searched instances locations was asked. Participants had only to find the clips where the instance appeared, but not the precise location and time stamp of the instance in the video clips.

3 Approach Overview

For our approach, we have considered a limited number of keyframes per shot (up to 4). We have then oversegmented each such keyframe in order to obtain a semi-global image representation. An aggregation mechanism was then applied in order to group a set of sub-regions into an object similar to the query, under a global similarity criterion. Our strategy relies on a greedy dynamic region construction method. The main aspects of our approach are presented in detail below. Let us start by detailing the color-based representation used.

3.1 DCD Representation

The object search process is performed uniquely upon the obtained keyframes in order to reduce the computational complexity. Each keyframe is segmented by applying the Mean Shift technique proposed in [5]. Let us mention that other segmentation methods can be used as well. Each region (or segment) determined is described by a unique, homogeneous color, defined as the mean value of the pixels of the given region. The set of colors, together with their percentage of occupation in the image (*i.e.*, the associated color histogram) are regrouped into a visual representation, which is similar to the MPEG-7 Dominant Color Descriptor (DCD). More precisely, let $C_I = \{c_1^I, c_2^I, \ldots c_{N_I}^I\}$ be the set of N_I colors obtained for image I, and $H_I = (p_1^I, p_2^I, \ldots p_{N_I}^I)$ the associated color histogram vector. The visual image representation is defined as the couple (C_I, H_I). An arbitrary number of dominant colors is supported, in contrast with the MPEG-7 DCD, where the maximal number of colors is limited to eight. In our experiments, we have used up to 32 dominant colors for each frame (Fig. 1). We can observe that despite the inherent loss in accuracy, the image content can still be visually recognized from the segmented images.

Fig. 2. Video frames (left) and their segmentations (right). A number of up to 5 colors are used

The query is, by definition, an object of arbitrary shape and is processed in the same manner in order to derive its visual representation. The advantage of the DCD representation comes from the fact that objects with arbitrary numbers of colors can be efficiently compared by using, for example, the Quadratic Form Distance Measure introduced in [6], which can be re-written for arbitrary length representations as described by the following equation:

$$D_h^2(H_Q, H_I) = \sum_{i=1}^{N_Q}\sum_{k=1}^{N_Q} a(c_i^Q, c_k^Q)\, p_i^Q p_k^Q + \sum_{j=1}^{N_I}\sum_{l=1}^{N_I} a(c_j^I, c_l^I)\, p_j^I p_l^I - \sum_{i=1}^{N_Q}\sum_{j=1}^{N_I} a(c_i^Q, c_j^I)\, p_i^Q p_j^I \quad , \tag{1}$$

where $H_Q = (p_1^Q, p_2^Q, \ldots p_{N_Q}^Q)$ and $H_I = (p_1^I, p_2^I, \ldots p_{N_I}^I)$ respectively denote the DCD histogram vectors of length N_Q, and N_I respectively associated to the query (Q) and candidate (I) images. The function a, describe the similarity between two colors c_i and c_j and is defined as

$$\mathrm{a}(c_i, c_j) = 1 - \frac{d(c_i, c_j)}{d_{max}} \tag{2}$$

where d is the Euclidean distance between colors c_i and c_j and d_{max} is the maximum Euclidean distance between any 2 colors in the considered color space (for the RGB color space $d_{max} \cong 442$).

Let us note that each color region in a candidate image has a specific contribution to the global distance. Thus, the contribution of color c_j^I in an image I to the global distance between image I and query Q is defined as:

$$C(c_j^I, Q) = \sum_{l=1}^{N_I} a(c_j^I, c_l^I)\, p_j^I p_l^I - \sum_{i=1}^{N_Q} a(c_i^Q, c_j^I)\, p_i^Q p_j^I \tag{3}$$

The above defined distance is used as a global criterion in the matching stage. Here, the objective is to determine, in each key-frame of the considered video sequence, candidate regions visually similar with the query.

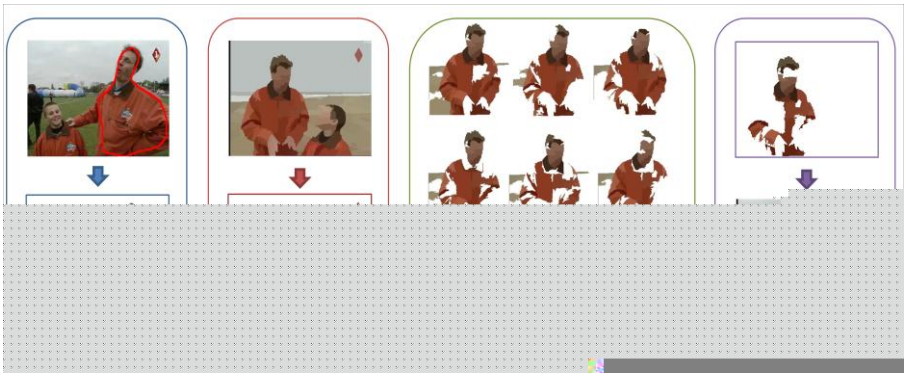

Fig. 3. Overview of the algorithm: (1): The query mask is used to crop out the corresponding segmented regions of the query; (2): Regions with colors highly different from the query are filtered out; (3): Different configurations of candidate objects are generated by adding/removing color segments; (4): The object with the minimal score is selected and displayed in its bounding box.

3.3.1 Relaxed Greedy Scheme

The algorithm starts from the initial set of regions obtained after the filtering stage described in the previous section. At each stage, we consider the current candidate object in image I and attempt to improve the current similarity measure between query Q and candidate objects. More precisely, we recursively eliminate the color segment which provides the highest contribution to the global distance (equation 3). We then check if the global distance is decreasing or not. If yes, we eliminate the corresponding region, update the color frequency vector H_I, and reiterate the algorithm on the new candidate object obtained. If not, the region is maintained and the algorithm successively tries to eliminate the following regions (sorted by decreasing order of their contribution to the global distance). Each time an attempt to eliminate a segment is performed, the region connectivity needs to be calculated in order to determine the eventually newly created connected components. Each connected component is then treated separately.

Concerning the exit condition, we constrain the algorithm to stop generating one. We consider that if the current distance is with $\check{u}$% higher than the previous obtained one, the candidate object has a low probability of reaching a configuration with a better score. The algorithm should in this case stop and return the current best distance. Otherwise, it should continue removing the regions with the highes configuration. In our submission, we have considered values of 10% and 20% for $\check{u}$, which provide a good trade-off between the number of generated configurations and the computational time [7].

The strategy of recursively eliminating the highest contributor to the global score increases the speed of the algorithm, by pruning the search space. However, the main limitation of the greedy-based approach is that it does not ensure the retrieval of an

5 Conclusion

In this paper we presented our experiments performed in the Instance Search task of the TRECVid 2011 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 retrieval.

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. Snoek, C.G.M., Worring, M.: Concept-Based Video Retrieval. Foundation and Trend in Information Retrieval, Vol.2, No.4 (2008) pp. 215-322.
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Discriminatively Trained Part Based Models. Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, September 2010
3. Sivic, J. and Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. IEEE International Conf. on Computer Vision (, 2003.
4. Smea ,
   Proc. 8th ACM International Workshop on Multimedia Information Retrieval (USA, October 26- 27, 2006). MIR '06. ACM Press, New York, NY, pp. 321-330.
5. Comaniciu, D., Meer, P.: "Mean Shift: A Robust Approach Toward Feature Space Analysis," IEEE Tran. on Pattern Analysis and Machine Intelligence, pp. 603-619, May, 2002.
6. histogram indexing for qua
   Machine Intell., vol. 17, pp. 729-736, July 1995.
7. Bursuc, A., Zaharia, T., Prêteux, F. ,
   In Proc. 7th ACM/IEEE International Conference on Signal Image Technology and Internet Based Systems (Dijon, France, November 28-December 1, 2011)
8. 220 (1983).