DYNAMIC DETECTION OF VISUAL ENTITIES

Andrei Bursuc[a, b], *Titus Zaharia*[a], *Françoise Prêteux*[c]

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

ABSTRACT

This paper tackles the issue of retrieving different instances of an object of interest within a given video document or in a video database. The principle consists of considering a semi-global image representation based on an over-segmentation of video frames. An aggregation mechanism is then applied in order to group a set of segments into an object similar to the query, under a global similarity criterion. We test the effectiveness of three different aggregation strategies, two of them based on a greedy approach, and the third one involving simulated annealing optimization. Experimental results on different color spaces show promising performances, with First Tier and Bull Eye detection rates of up to 70% and 88%, respectively. The integration of the method in a web-based video navigation system, allowing fast video object retrieval, is finally described.

Index Terms— object-based indexing and retrieval, multiple instance detection, partial matching, MPEG-7 visual descriptors, video indexing, color spaces.

1. INTRODUCTION

Object retrieval in videos is among the most challenging tasks in the field of computer vision applications. In the last years an increasing number of solutions have provided a variety of satisfying results for concept detection and object categorization in videos [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 strategies. In addition, variations in visual appearance and object's pose have to be taken into account appropriately. Popular existing methods for object indexing and retrieval such as the discriminatively trained deformable part-based models [2] or the bag-of-words (BoW) representations [3] show their limitations when considering the high variability of items that a user might want to retrieve. Such methods require the use of an off-line training phase, which is dependent of the considered training set and of the pre-defined categories.

2. RELATED WORK

Related work includes two families of approaches, the first one exploiting interest points and the second relying on region-based representations.

Currently, interest points are among the most popular tools for object recognition and classification for both images and videos. Early approaches for object retrieval, using interest points, have been developed by Sivic *et al.* in the Video Google system [3]. The BoW model is used for achieving fast and efficient retrieval of objects interactively selected by the user. Despite its simplicity and efficiency, BoW discards spatial information, which is essential for visual representation because of the ambiguity of visual words. Spatial information is usually re-introduced in the post-processing step through a geometry check (*e.g.* RANSAC [4], neighboring feature consistency [3]) or it is rigidly encoded by quantizing the image space (*e.g.* Spatial Pyramid Matching [5]). Geometry verification is computationally expensive and it is applied only to the top images in the initial ranking. Additionally, the number of interest points extracted from an image varies a lot with the image content with a low for less-textured images. While dense sampling ensures a constant number of interest points, it brings an extra computational burden.

Region-based approaches tackle the issue of less-textured images by leveraging on popular segmentation techniques to generate segments to be used for object representation. Gould *et al.* [6] propose to combine appearance-based features computed on superpixels patches with relative location priors in a two stage classification process. Claiming that no segmentation method is perfect, in [7] multiple image segmentations are combined and the resulted regions are described so-called Region Context Features. Kim *et al.* [8] merge fine overlapping entities into regions under spatial and similarity constraints. Such obtained regions are used as basic measure units for training an object category classifier. Recently, [9] used graphs for

discovering object instances, by grouping segmented regions in a graph of regions according to their similarity. Regions belonging to the instances of the same objects are thus discovered as connected components in a graph containing all the regions in the dataset.

The advantage of region-based approaches comes from the possibility of directly exploiting the connectivity information (*i.e.* adjacency between regions), which can be highly useful in the matching stage. Still, few approaches consider objects as consistent entities of adjacent regions, as only the most relevant regions from a certain object category are learned and weighted up from the considered pool of segments.

The method proposed in this paper also adopts a region-based representation strategy, which involves an over-segmentation of the image. Images are represented as graphs of regions and objects become sub-graphs. Objects are single connected component groups of adjacent regions, and their inner structure information can be thus exploited. The key ingredient of our approach is a dynamic region grouping procedure, which makes it possible to regroup different individual regions into a candidate object. A global matching score, which measures the similarity of the candidate object with the given query, needs to be minimized. Different matching strategies, including greedy algorithm and simulated annealing optimization, are considered.

The rest of the paper is organized as follows. The following section recalls the color-based visual representation adopted, based on the MPEG-7 Dominant Color Descriptor (DCD). The dynamic region grouping algorithms proposed are then detailed in Section 4. Section 5 presents and analyzes the experimental results obtained. The integration of the proposed method in a dedicated video object search and retrieval platform is presented in Section 6. Finally, Section 7 concludes the paper and opens some perspectives of future work.

3. DCD REPRESENTATION

The video document is decomposed in shots and a set of representative keyframes is determined for each shot, using the approach recently proposed in [10]. The keyframes are then segmented by applying standard algorithms. In our case, we have used the popular MeanShift technique [11]. Each region (or segment) 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 are regrouped into a visual representation, which is similar to the MPEG-7 DCD [12]. 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 percentages of occupation. The visual image representation is defined as the couple (C_I, H_I).

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 [13], defined as:

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

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 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*:*

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

where d is a distance between colors c_i and c_j in the considered color space, while d_{max} is its maximum value. The above-defined distance is used as a global criterion in the matching stage.

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:

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

We use equation 3 for retrieving the regions with the highest negative impact over the global criterion.

4. DYNAMIC REGION GROUPING

In a first stage, we perform a pre-filtering of each frame, aiming to eliminate individual regions with far-off colors, based on a color similarity criterion. Let us emphasize that the objective here is not to determine precisely the candidate object, but to roughly restrict the number of candidate regions, by eliminating far-off colors using a permissive threshold. The resulting regions are labeled into connected components and each such connected component is then considered as an initial candidate object to be matched with the query. Next, we consider a dynamic region grouping algorithm. Each candidate object is iteratively refined by removing and adding regions until the global matching distance is minimized. The candidate object presenting the minimum distance is selected and displayed as a result. An overview of the proposed approach is illustrated in Figure 1.

In this paper we have considered three optimization methods [14], so-called *greedy*, *relaxed greedy* and *simulated annealing*, and investigated their performances on different color spaces.

In the case of the *greedy* algorithm, at each step, we consider the current candidate object in image I and attempt to improve the current similarity measure between query and candidate objects. We recursively eliminate the color segment which provides the worst contribution to the global distance (*cf.* equation 3). We then check if the global

distance is decreasing or not. If positive, we eliminate the corresponding region, update the color frequency vector H_I, and re-iterate the algorithm on the new candidate object obtained by trying to eliminate the following noisy regions as long as the distance decreases. When the distance increases, we return to the previously obtained best score configuration and stop the algorithm.

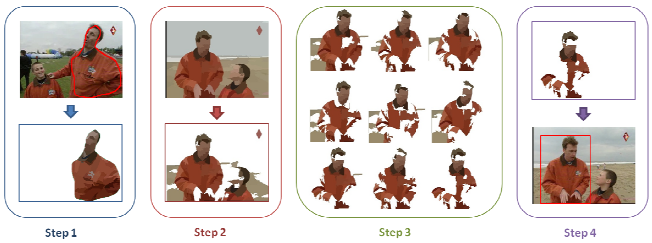

Figure 1. Overview of the algorithm: Step1: The user selects an object and the corresponding segmented region is extracted; Step.2: Regions with colors highly different from the query are filtered out. The remaining regions are separated in connected components, each representing a different initial candidate object; Step 3: Different configurations of candidate objects are generated by adding and removing candidate color segments; Step4: The object with the best score is selected and displayed in its bounding box.

When considering a greedy approach, the risk is to remain blocked in a local minimum, because whenever the distance is increasing, the algorithm stops. In order to overcome such a limitation, we consider a *relaxed greedy procedure*, which allows up-hill movements (in the sense of the functional in equation 3). If the current distance obtained is $\delta\%$ lower than the previous score, than the new configuration is accepted. We consider that if the current distance is with $\delta\%$ higher than the previous obtained one, the candidate object has a low probability of reaching a configuration with a better score. In this case, the algorithm should stop and return the current best distance.

Finally, in order to achieve asymptotic optimality, we have adopted a *simulated annealing* matching strategy. We associate a binary state $S = S(c)$ to each color segment c. If the color state S is equal to 0, the color segment c is considered as not belonging to the current candidate object. On the contrary, if the state S has value 1, c is considered as part of the current object. At each step, the algorithm attempts to change the stage of the current color segment, by investigating the variation of the global energy $\Delta E = E(S') - E(S)$ (with energy E defined as described in equation 1) when the state S is set to its complementary value S'.

If the variation $\Delta E \leq 0$, then the current state S is replaced by its complementary state S'. If the energy variation ΔE has a positive value, then a random variable α, $0 \leq \alpha \leq 1$ is generated. The current state S is replaced by S' if the following condition is satisfied:

$$\alpha \leq e^{(-\Delta E/T)} \text{ and } E(S') < E(S) \qquad (4)$$

where T denotes the value of the temperature at the current state.

For each temperature, n_{it} iterations are performed. The temperature of the system is then iteratively lowered, according to the following given freezing scheme:

$$T_{n+1} = \tau T_n \qquad (5)$$

where τ is the (constant) cooling rate with value between 0 and 0.99, and T_n is the temperature at the n^{th} iteration. The algorithm starts at an initial temperature T_0 and stops when a freezing temperature T_f is reached. The considered values of SA method parameters are illustrated in Table 1.

Let us now describe the experimental results obtained.

5. EXPERIMENTAL RESULTS

For our experiments, we have considered the Sound and Vision dataset, used for the TRECVID 2010 Instance Search Task. Thus, we have considered 34 videos summing up to approximately 15 hours of video content. A total number of 5580 keyframes has been obtained. The MeanShift algorithm has generated an average number of 200 regions per frame, with a maximum of 330 regions per frame.

In order to test the influence of the choice of color space over the results, we have tested our methods in the RGB and CIE-Lab color spaces. For the RGB color space we have considered the Euclidean distance as a similarity measure between region colors. For the CIE-Lab color space we have retained the two distance similarity measures so-called CIE76 and CIE94 [15].

In order to objectively evaluate the performances of the algorithm, we have retained 12 query images (Figure 2), interactively selected by different users. Let us note that even though the considered queries represent persons, we pose the problem as a retrieval of generic objects one. Therefore the search is performed based on the visual appearance of the person (*e.g.* clothes) and we target the shots where the person is displayed with a simila appearance.

The query specification is not precise, the user being requested to select approximately the desired objects of interest. We have then established a ground truth set of relevant frames, by exhaustively selecting the set of all frames where the query object appears.

We have performed 2 sets of experiments for each considered query. Firstly, we have launched the query only within the video of the query object in order to test a typical use case for a regular user. Further on we have launched the query in the whole dataset in order to test the scalability of the proposed methods.

For the performance evaluation, we have adopted the First Tier (FT) and Bull Eye (BE) scores, respectively defined as the percentage of relevant results retrieved within the top P and (2P) first retrieved results, where P denotes the total number of relevant responses for a given query.

Figure 2. The query objects considered from the TRECVD 2010 data set and retained for our evaluation.

Experiments have been run on a PC with Intel Xeon CPU W3530 at 2.80 GHz and 12 Gb RAM. Computational times are display in Table 1.

Run	**Greedy**	**Relaxed Greedy** (δ=10%)	**SA** ($T_0 = 0.5$, n_{it}=5, $T_{fin} = 10^{-3}$)
Intra-Video time	6s	10s	816s
Inter-Video time	147s	270s	19388s

Table 1. Computational times for RGB color space

Figure 3 summarizes the results obtained for all the three proposed methods when the search is performed within the video of origin of the object of interest. The analysis of the results shows that the CIE76 runs perform best on all three methods, while RGB has the weakest performance. In the case of the greedy methods the improvement is more significant: +13% on both FT and BE for the *relaxed greedy* method and +12% and +9% for FT and BE, respectively, comparing to the RGB scores. In the case of simulated annealing, the gap is less significant, the algorithm approaching the optimal solution whatever the considered color space. Overall, the SA-based method yields the best retrieval performances, with FT and BE scores of 70% and 88%, respectively for CIE76. The SA optimization also presents a stable behavior, with quasi equivalent results for different values of the parameters T_0, T_{fin} and n_{it}. However, the main limitation of the SA approach is related to the high computational complexity, the average time per query being close to 800 seconds. The basic greedy method is much faster, with only 6 seconds in average per query. However, the associated retrieval performances are much inferior, with FT and BE scores of 51% and 69%, respectively for the RGB run. This drawback is partially eliminated by the relaxed greedy scheme, which achieves higher retrieval scores (FT = 67% and BE = 87%) close to the SA ones, while offering interesting performances in terms of computational complexity (with an average execution time of 11 seconds). Let us note that the influence of parameter δ is quite reduced, similar results (±2% in FT scores) being obtained for values of δ between 5% and 20%.

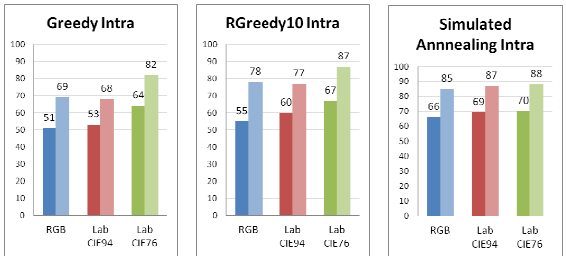

Figure 3.Experimental results for intra-video querying.

Figure 4 summarizes the results obtained for the three proposed methods when searching for the object of interest in the entire dataset. The analysis of the results shows that the CIE76 runs still perform best on all three methods, while CIE94 has the weakest performance. In the first case the CIE94 and RGB performances were similar, but in this situation the performance is considerably lower. The decrease in performance is less dramatic for the SA runs, when all three methods have similar results. The robustness of the SA approach is thus preserved even for larger datasets. Overall the improvements of the SA methods are stronger, excepting the CIE76 runs where the Relaxed Greedy scores are still close to SA ones.

The computational performances for the greedy methods remain within reasonable limits (147–270 seconds), while the price of the SA optimum solution takes up to 5 hours.

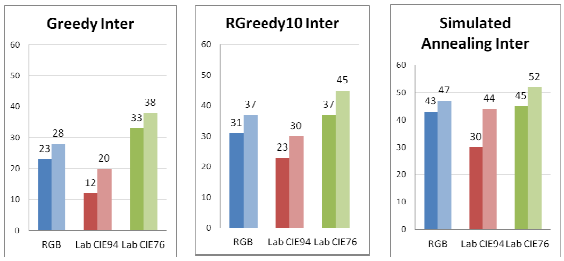

Figure 4.Experimental results for inter-video querying.

6. FAST OBJECT INSTANCES RETRIEVAL WITH THE OVIDIUS PLATFORM

In this section we illustrate the functionalities of the OVIDIUS platform for quick retrieval of shots and object instances within a video document. The OVIDIUS platform [16] achieves a good balance between such navigation, visualization and search features in an accessible web-based environment.

The video browsing process is based on the MPEG-7 structural approach for video description [17], which relies on the abstract concept of *segment*, defined as a generic piece of an audio-visual document. In combination with a recursive decomposition mechanism (which may be temporal, spatio-temporal or spatial), such an approach makes it possible to create a hierarchical and multi-granular video content description, adapted to both navigation/browsing and search functionalities.

The OVIDIUS GUI makes it possible to obtain a visual and interactive representation of the MPEG-7 descriptive structure and integrates the following components: video player, selector of the hierarchical level of each segment, rows of iconic representations of segments including the display of basic information such as (identifier, time stamp, type), navigation buttons and a timeline scrollbar for instant access to scenes from different parts of the video (Figure 5). A color code is used in order to inform the user about the current hierarchical level of access. While visualizing the video, the user can simultaneously skim through neighboring scenes and shots. The user can also add another navigation row of iconic representation from a different hierarchic level of segmentation. One thread contains the scene previews, while the other one displays the previews of the shots corresponding to the current scene. Users can switch between different segments horizontally (on the same hierarchical level) and vertically (between different hierarchical levels).

Concerning the search functionalities OVIDIUS integrates the above-described object retrieval method. Regions and objects of interest can be interactively selected by the user and retrieved inside the video. The retrieval results are displayed in decreasing similarity order in a dedicated thread.

7. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented a region-based object retrieval method which makes it possible to detect throughout the video multiple instances of an object selected by the user. Different color spaces have been evaluated.

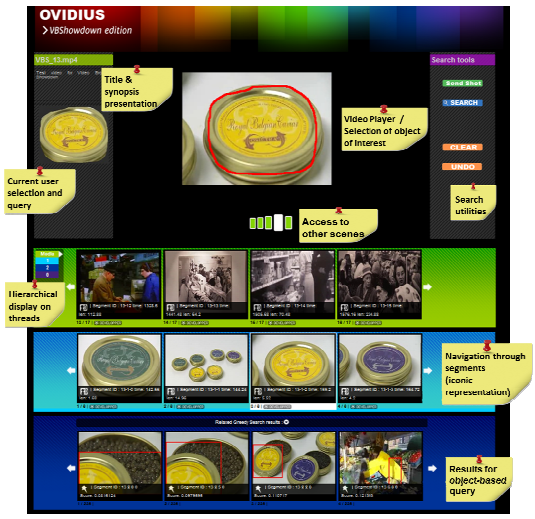

Figure 5.OVIDIUS GUI.

Promising retrieval results, with object detection rates of up to 70% (in FT score) and 88% (in BE score) have been obtained on the challenging TRECVID 2010 instance search track video database which involves various characters in visually different scenes. The CIE Lab color space combined with the CIE76 similarity measure yield the highest performances. Finally, the integration of the method within the OVIDIUS platform has been presented.

The perspectives of future work concern the integration in the same region grouping process of additional spatial information in the considered approach, in order to obtain a more discriminative representation. In this respect, the graph cut global optimization method offers interesting perspectives of future development.

AKNOWLEDGMENT

The current work has been carried out within the framework of the UBIMEDIA Joint Lab between Institut Télécom and Alcatel-Lucent Bell Labs.

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. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, “Object Detection with Discriminatively Trained Part Based Models”, IEEE Trans. on PAMI, vol. 32, no. 9, September 2010.

[3] J. Sivic, and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos”, Int. Conf. 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”, Int. Conf. on Computer Vision and Pattern Recognition, 2007.

[5] S. Lazebnik, C. Schmid, and J. Ponce. “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories”, Int. Conf. on Computer Vision and Pattern Recognition, 2006.

[6] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, “Multi-class segmentation with relative location prior”, Int. Journal on Computer Vision, 2008.

[7] C. Pantofaru, C. Schmid, and M. Hebert. “Object Recognition by Integrating Multiple Image Segmentations”, European Conf. on Computer Vision, 2008.

[8] K. Kim, K. Grauman, “Boundary Preserving Dense Local Regions”, Int. Conf. on Computer Vision and Pattern Recognition, 2011.

[9] H. Kang, M. Hebert, T. Kanade, “Discovering object instances from scenes of Daily Living”, Int. Conf. on Computer Vision, 2011.

[10] R. Tapu, T. Zaharia, “A complete framework for temporal video segmentation”, Int. Conf. on Consumer Electronics Berlin, 2011.

[11] D. Comaniciu, P. Meer, “Mean Shift: A Robust Approach Toward Feature Space Analysis”, IEEE Trans. on PAMI, pp. 603-619, May, 2002.

[12] International standard ISO/IEC 15938-3:2002, Information technology - Multimedia Content Description. Interface - Part 3: Visual. 2002.

[13] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, W. Niblack, “Efficient color histogram indexing for quadratic form distance functions”, IEEE Trans. on PAMI, pp. 729–736, July 1995.

[14] A. Bursuc, T. Zaharia, F. Preteux, “Detection of Multiple Instances of Video Objects”, Int. Conf. on Signal-Image Technologies and Internet-Based Systems, 2011.

[15] D.H. Brainard, ”Color Appearance and Color Difference Specification”, In Shevell, Steven K. *The Science of Color* (2 ed.). Elsevier. p. 206.

[16] A. Bursuc, T. Zaharia, F. Prêteux, “Mobile Video Browsing and Retrieval with the OVIDIUS Platform”, ACM Multimedia, 2010.

[17] International standard ISO/IEC 15938-5:2003, Information technology - MultimediaContent Description. Interface-Part 5: Multimedia Description Schemes. 2003.