ENHANCING LATENT SEMANTIC ANALYSIS VIDEO OBJECT RETRIEVAL WITH STRUCTURAL INFORMATION

Lukas Hohl, Fabrice Souvannavong, Bernard Merialdo and Benoit Huet

Multimedia Departement
Institut Eurecom
2229 routes des Cretes
06904 Sophia-Antipolis, France
e-mail: {hohl, souvanna, merialdo, huet}@eurecom.fr

ABSTRACT

The work presented in this paper aims at reducing the semantic gap between low level video features and semantic video objects. The proposed method for finding associations between segmented frame region characteristics relies on the strength of Latent Semantic Analysis. Our previous experiments [1] have shown the potential of this approach but also uncovered some of its limitation. Here, we will present a method using the structural information within an LSA framework. Moreover, we will demonstrate the performance gain of combining visual (low level) and structural information.

1. INTRODUCTION

Multimedia digital documents are readily available, either through the internet, private archives or digital video broadcast. Traditional text based methodologies for annotation and retrieval have shown their limit and need to be enhanced with content based analysis tools. Research aimed at providing such tools have been very active over recent years [2]. Whereas most of these approaches focus on frame or shot retrieval, we propose a framework for effective retrieval of semantic video objects. By video object we mean a semantically meaningful spatio-temporal entity in a video.

Most traditional retrieval methods fail to overcome two well known problems called synonymy and polysemy, as they exist in natural language. Synonymy occurs when different words describing the same object, whereas polysemy corresponds to words that refer to more than one object. Latent Semantic Analysis (LSA) provides a way to weaken those two problems [3]. It has been primarily used in the field of natural language understanding, but has recently been applied to domains such as source code analysis or computer vision. Latent Semantic Analysis has also provided very promising results in finding the semantic meaning of multimedia documents [1, 4, 5, 6]. LSA is based on a Singular Value Decomposition (SVD) on a word by context matrix, containing the frequencies of occurrence of words in each context. One of the limitations of the LSA is that it does not take into account word order, which means it completely lacks the syntax of words. The analysis of text, using syntactical structure combined with LSA already has been studied [7, 8] and has shown improved results. For our object retrieval task the LSA is computed over a visual dictionary where region characteristics, either structurally enhanced or not, correspond to words.

The most common representation of visual content in retrieval system relies on global low level features.These techniques are not suited for object representation as they capture information from the entire image, merging characteristics of both the object and its surrounding. A solution is to segment the image in regions with homogenous properties and use a set of low level features of each region as global representation. An object is then referred to as a set of regions.Despite the improvement over the global approach, region based methods still lack important characteristics in order to uniquely define objects. Indeed it is possible to find sets of regions with similar low level features yet depicting very different content. The use of relational constraints, imposed by the region adjacency of the image itself, provides a richer and more discriminative representation of video object. There has only been limited publications employing attributed relational graph to describe and index into large collection of visual data [9, 10]. Here we will show that it is possible to achieve significant performance improvement using structural constraints.

This paper is organized as follows. The concept of adding structure to LSA and a short theoretical background on the algorithms used, are presented in Section 2. Section 3 provides the experimental results looking at several different aspects. The conclusion and future directions are discussed in Section 4.

2. ENHANCING LATENT SEMANTIC ANALYSIS WITH STRUCTURAL INFORMATION

As opposed to text documents there is no predefined dictionary for multimedia data. It is therefore necessary to create one to analyze the content of multimedia documents using Latent Semantic Analysis [3]. In the non-structural approach each frame region of the video is assigned to a class based on its properties. This class corresponds to a ”visual” word and the set of all classes is our visual dictionary. In the structural case the classes do not directly correspond to visual words. Pairs of adjacent regions classes are used to define the structural dictionary. We shall now detail the steps leading to dictionary construction.

2.1. Video preprocessing

Every 25th frame F_i of the video V is segmented using [11] into regions R_{ij} (the j-th region in the i-th frame). Each segmented region R_{ij} is characterized by its attributes, feature vectors that contain visual information about the region such as color, texture, size or spatial information. For this paper, the feature vector is the 32 bin color histograms in HS color space of the corresponding region.

2.2. Building the basic visual dictionary

The structure-less dictionary is constructed by grouping regions with similar feature vectors together. Here the k-means clustering algorithm [12] is employed with the Euclidean distance as similarity measure. As a result each region R_{ij} is mapped to a cluster C_l, represented by its cluster centroid. Thanks to the k-means clustering parameter kc controlling the number of clusters, the dictionary size may be adjusted to our needs.

2.3. Building a visual dictionary using structure

We now wish to construct a visual dictionary D_ν (of size ν) is containing words with structural information. This is achieved by considering every possible unordered pair of clusters as a visual word W,e.g. $C_3C_7 \equiv C_7C_3$.

$$D_\nu = \{W_1, \dots, W_\nu\}$$

$$(C_1C_1) \simeq W_1, (C_1C_2) \simeq W_2, \dots, (C_{kc}C_{kc}) \simeq W_\nu$$

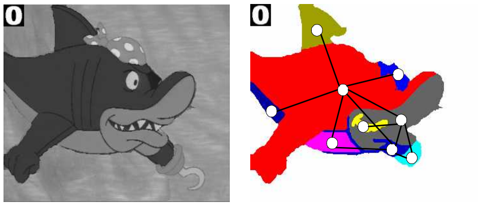
The size ν of the dictionary D_ν is also controlled by the clustering parameter kc but this time indirectly.

$$\nu = \frac{kc \cdot (kc-1)}{2} + kc \qquad (1)$$

To be able to build these pairs of clusters (words), we look at an abstract representation of the connectivity of the segmented regions for each frame. Each region is labeled with the cluster number it belongs to (e.g. C_{14}). If two regions are adjacent, they are linked in an abstract point of view, which results in a graph $G_i = (V, E)$ consisting of a set of vertices $V = \{v_1, v_2, \dots, v_n\}$ and edges $E = \{e_1, e_2, \dots, e_m\}$, whereas the vertices represent the cluster number labeled regions and the edges the connectivity of the regions (note: index i refers to the i-th frame). A region is called to be adjacent to another region if they are physically connected to each other, meaning that they share a common boundary (border) of at least one pixel. Thus each frame F_i of the video has its corresponding graph G_i which describes the frame as a set of elements (segmented regions R_{ij}) labeled with the cluster they belong to and their structural relations.

Every Graph G_i is described by its adjacency matrix. The matrix is a square matrix ($n \times n$) with both, rows and columns, representing the vertices from v_1 to v_n in an ascending order. The cell (i,j) contains the number of how many times vertex v_i is connected to vertex v_j.

Figure 1(b) shows a frame segmented into regions with is corresponding relational graph overlaid.



(a) (b)

Fig. 1. (a) The shark and (b) its corresponding ARG.

2.4. Latent Semantic Analysis

The LSA describes the semantic content of a context by mapping words (within this context) onto a semantic space. Singular Value Decomposition (SVD) is used to create such a semantic space. A co-occurrence matrix $\mathbf{A}$ containing words (rows) and contexts (columns) is built. The value of a cell a_{ij} of $\mathbf{A}$ contains the number of occurrence of the word i in the context j. Then, SVD is used to decompose the matrix $\mathbf{A}$ (of size $M \times N$, M words and N contexts) into three separate matrices.

$$\mathbf{A} = \mathbf{USV^T} \qquad (2)$$

The matrix $\mathbf{U}$ is of size $M \times L$, the matrix $\mathbf{S}$ is of dimension $L \times L$ and the matrix $\mathbf{V}$ is $N \times L$. $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices, thus $\mathbf{U^TU} = \mathbf{V^TV} = \mathbf{I_L}$ whereas $\mathbf{S}$ is a diagonal matrix of size $L = min(M, N)$ with singular values σ_1 to σ_L, where

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_L \qquad \mathrm{S} \approx diag(\sigma_1, \sigma_2, \dots, \sigma_L)$$

$\mathbf{A}$ can be approximated by reducing the size of $\mathbf{S}$ to some dimensionality of $k \times k$, where $\sigma_1, \sigma_2, \ldots, \sigma_k$ are the k highest singular values. By doing a reduction in dimensionality from L to k, the sizes of the matrices $\mathbf{U}$ and $\mathbf{V}$ have to be changed to $M \times k$ respectively $N \times k$. Thus, k is the dimension of the resulting semantic space.To measure the result of the query, the cosine measure (m_c) is used. The query vector $\mathbf{q}$ contains the words describing the object, in a particular frame where it appears.

$$\mathbf{q^T \hat{A}} = \mathbf{q^T U_k S_k V_k^T} = (\mathbf{q^T U_k})(\mathbf{S_k V_k^T}) \qquad (3)$$

Let $\mathbf{p_q}$ = $\mathbf{q^T U_k}$ and $\mathbf{p_j}$ to be the j-th context (frame) of $(\mathbf{S_k V_k^T})$

$$m_c(\mathbf{p_j}, \mathbf{q}) = \frac{\mathbf{p_q} \cdot \mathbf{p_j}}{||\mathbf{p_q}|| \cdot ||\mathbf{p_j}||} \qquad (4)$$

The dictionary size ought to remain ”small” to compute the SVD as its complexity is $O(P^2k^3)$, where P is the number of words plus contexts ($P = N + M$) and k the number of LSA factors.

3. EXPERIMENTAL RESULTS

Here, our object retrieval system is evaluated on a short cartoon (10 minutes long) taken from the MPEG7 dataset and created by D’Ocon Film Productions. A ground truth has been created by manually annotating some objects (figure 2) through the entire video. The query objects are chosen as diverse as possible and appear in 30 to 108 frames of the subsampled video.

A query object may be created by selecting a set of region from a video frame. Once the query is formed, the algorithm starts searching for frames which contain the query object. The query results are ordered so that the frame which most likely contains the query object (regarding the cosine measure m_c) comes first. The performance of our retrieval system are evaluated using the standard precision vs recall values.

We have selected 4 objects (figure 2) from the sequence. Some are rather simple with respect to the number of region they consist of, while others are more complex. Unless stated otherwise, the plots show the average (over 2 or 4 objects) precision values at given standard recall values [0.1, 0.2, . . . , 1.0].

3.1. Impact of the number of clusters

To show the impact on the size of clusters chosen during video preprocessing, we have built several dictionaries containing non-structural words. Figure 3 shows the precision/recall curves for three cluster size (32, 528, 1000). The two upper curves (528 and 1000 clusters) show rather steady high precision values for recall value smaller than 0.6. For 32 clusters the performance results are weaker. Using 528 clusters always delivers as good results as using 1000 clusters which indicates that after a certain number of clusters performances cannot be improved and may even start to decay. This is due to the fact that for large k the number of regions per cluster become smaller, meaning that similar content may be assigned to different clusters.

3.2. Structural versus non-structural words

For a given cluster size (kc=32) we compared two different ways of defining the visual words used for LSA. In the non-structural case, each cluster label represents one word, leading to a dictionary size of 32 words.In the structural case, every possible pair of cluster label is defining a word (as explained in 2.3), so that the number of words in the dictionary is 528. Figure 4 shows the results for both approaches when querying for four objects and two objects. The group of two objects contains the most complex ones. The structural approach clearly outperforms the non-structural methods. Even more so, as the objects are most complex. The structural approach is constantly delivering higher precision values, than the non-structural version, throughout the whole recall range.

3.3. Impact of LSA factors

An important parameter of the LSA method is the number of factors used to compute the similarity of the query regions with the video frames. Figure 5 shows the mean average precision values for different k using either a dictionary without structure or with structure. Note that k can vary in the range [1, . . . ,32] for the structure-less method and in the range [1, . . . ,528] for the approach using the structural representation. The maximum of the mean average precision values for the approach using structure is 0.78 at

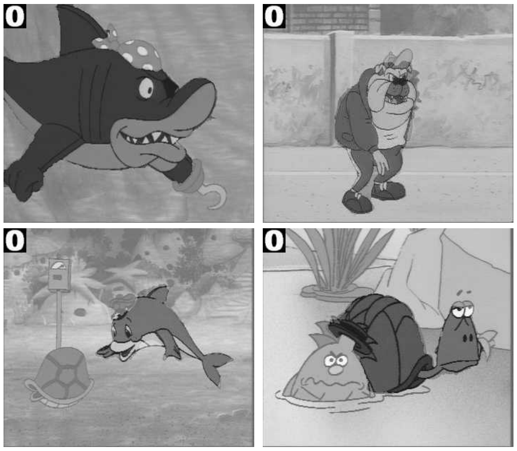


Fig. 2. The 4 query objects.

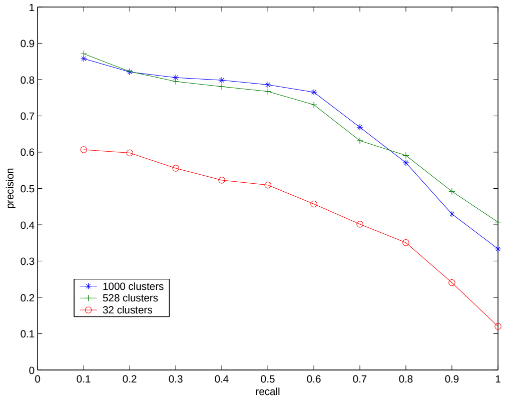


Fig. 3. Retrieval performance w.r.t. number of clusters.

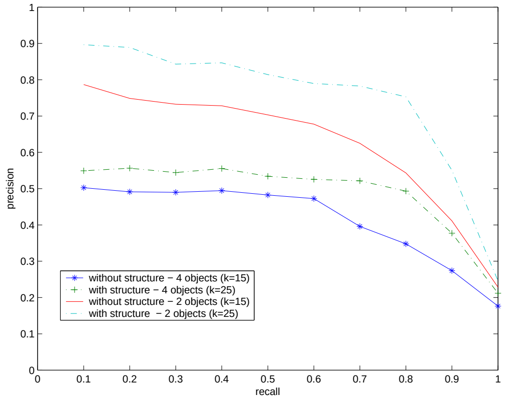


Fig. 4. Retrieval performance for 2 and 4 objects queries with structure and without.

$k = 25$, while for the non-structural approach the best precision (0.66) is obtained with $k = 25$. This concludes in 18% increase of the mean average precision value using structure for a given number of clusters (32 clusters in this case).

4. CONCLUSION AND FUTURE WORK

In this paper we have presented a method for enhancing an LSA based video object retrieval system with structural constraints obtained from the object visual properties. The method was compared to a similar method [1] which did not make use of the relational information between adjacent regions. Our results show the importance of structural constraints for region based object representation. This is demonstrated by a 18% performance increase in the optimal situation for a common number of region categories. We are currently investiguating the sensitivity of this representation to the segmentation process as well as other potential graph structures.

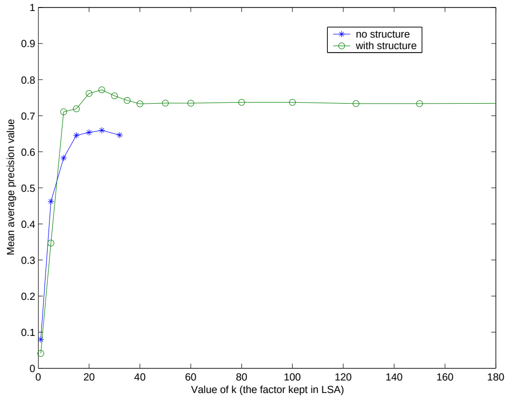


Fig. 5. The optimal k for both approaches.

5. REFERENCES

[1] Fabrice Souvannavong, Bernard Merialdo, and Benoît Huet, “Video content modeling with latent semantic analysis,” in *Int. Workshop on Content-Based Multimedia Indexing*, 2003.

[2] TREC Video Retrieval Workshop (TRECVID) http://www-nlpir.nist.gov/projects/trecvid/.

[3] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[4] Rong Zhao and William I. Grosky, *Video Shot Detection Using Color Anglogram and Latent Semantic Indexing: From Contents to Semantics*, CRC Press, 2003.

[5] F. Monay and D. Gatica-Perez, “On image auto-annotation with latent space models,” *ACM Multimedia*, 2003.

[6] Xin Liu Yihong Gong, “Video summarization and retrieval using singular value decomposition,” *Journal of ACM Multimedia Systems*, vol. 9, pp. 157–168, 2003.

[7] Peter Wiemer-Hastings, “Adding syntactic information to LSA,” in *Conf. of the Cognitive Science Society*, 2000, pp. 989–993.

[8] T. Landauer, D. Laham, B. Rehder, and M. Schreiner, “How well can passage meaning be derived without using word order,” *Conf. of the Cognitive Science Society*, 1997, pp. 412–417.

[9] K. Shearer, S. Venkatesh, and H. Bunke, “An efficient least common subgraph algorithm for video indexing,” *Int. Conf. on Pattern Recognition*, vol. 2, pp. 1241–1243, 1998.

[10] B. Huet and E.R. Hancock, “Line pattern retrieval using relational histograms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1363–1370, December 1999.

[11] P. Felzenszwalb and D. Huttenlocher, “Efficiently computing a good segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 98–104.

[12] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.