REGION-BASED VIDEO CONTENT INDEXING AND RETRIEVAL

Fabrice Souvannavong, Bernard Merialdo and Benoit Huet

Département Communications Multimédia
Institut Eurécom
2229, route des crêtes
06904 Sophia-Antipolis - France
(souvanna, merialdo, huet)@eurecom.fr

ABSTRACT

In this paper we propose to compare two region-based approaches to content-based video indexing and retrieval. Namely a comparison of a system using the Earth Mover's Distance and a system using the Latent Semantic Indexing is provided. Region-based methods allow to keep the local information in a way that reflects the human perception of the content. Thus, they are very attractive to design efficient Content Based Video Retrieval systems. We presented a region based approach using Latent Semantic Indexing (LSI) in previous work. And now we compare performances of our system with a method using the Earth Mover's Distance that have the property to keep the original features describing regions. This paper shows that LSA performs better on the task of object retrieval despite the quantification process implied.

1. INTRODUCTION

The growth of numerical storage facilities enables large quantities of documents to be archived in huge databases or to be extensively shared over the Internet. The advantage of such mass storage is undeniable. However the challenging tasks of multimedia content indexing and retrieval remain unsolved without expensive human intervention to archive and annotate contents. Many researchers are currently investigating methods to automatically analyze, organize, index and retrieve video information [3, 12, 18, 1]. On one hand this effort is further stressed by the emerging MPEG-7 standard that provides a rich and common description tool of multimedia contents. On the other hand it is encouraged by TRECVID [1] which aims at evaluating state of the art developments in video content analysis and retrieval tools.

We propose to compare a system using the Earth Mover's Distance (EMD) and a system using the Latent Semantic Indexing (LSI) on the task of content-based information retrieval. These two systems have the property to compare video shots at the granularity of the region. Contrasting to traditional approaches which compute global features, these region-based methods extract features of segmented frames. The main objective is then to keep the local information in a way that reflects the human perception of the content [2, 19]. Thus, such methods are very attractive to design efficient Content-Based Video Retrieval (CBVR) systems. Following this idea, we proposed in previous works [16, 17] to use Latent Semantic Indexing for video shot retrieval.

LSI has been proven effective for text document analysis, indexing and retrieval [4]. Some extensions to audio and image features were proposed in the literature [9, 20]. The adaptation we presented models video shots by a count vector in a similar way as for text documents. This representation is defined as the Image Vector Space Model (IVSM). Key frames of shots are described by the occurrence of a set of predefined *visual terms*. *Visual terms* are based on a perceptual segmentation of images. The underlying idea is that each region of an image carries a semantic information that influences the semantic content of the whole shot.

Contrasting to LSI, EMD-based systems directly compute a distance on region features. Furthermore, the distance measures the minimal cost that must be paid to transform a set of regions into the other, providing an interesting measure of image differences. For storage convenience, region features can be quantized. The Image Vector Space Model is then a common basis for LSI and EMD based systems.

The aim of the paper is finally to compare these two ways of indexing images. The paper is structured as follows: Section 2 is a short state-of-the-art of region-based indexing methods. Then, we present the Image Vector Space Model and its application to video shot representation. Section 4 presents the EMD that is followed by a presentation of the LSI. Next, they are evaluated and compared on the task of object retrieval. Finally, we conclude with a brief summary and future work.

[1] Text REtrieval Conference. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation. http://www-nlpir.nist.gov/projects/trecvid/

2. PREVIOUS WORK

Existing general purpose content-based image retrieval systems roughly fall into two categories depending on the feature extraction method used: frame level or region level feature extraction.

Frame level systems describe the entire frame content [5, 13] and visual descriptors are extracted on the complete frame. Unfortunately extracted descriptors such as histograms do not contain spatial information, thus differences are computed with few constraints.

Region-based retrieval systems attempt to overcome the deficiencies of previous systems by representing images at the object level. An image segmentation algorithm is then applied to decompose images into regions which correspond to objects in the ideal case. This representation at the granularity of the region is intented to be close to the perception of the human visual system by highlighting local features.

Region-based systems are mainly decomposed into two categories depending on the way query and target regions are matched: individual region or frame regions matching. In the first case the query is performed by merging single-region query results [2, 14]. A score between each query region and each target frame regions is computed. Next, individual scores are merged to order frame by relevance. In the second case the approach is slightly different since the information of all regions composing target images is used [19, 8, 10, 7].

In this paper, we focus our interest on the last situation where all the information of all regions composing frames is used. In particular, we will have a closer look at the LSI [17] and EMD [8] based methods. We begin by presenting the Image Vector Space Model.

3. IMAGE VECTOR SPACE MODEL

The Vector Space Model of text processing [15] is the most widely used information retrieval model. In this model, each document is stored as a vector of terms. In practice these terms are extracted from the text itself subject to stemming and filtering. Finally a common vocabulary is defined to describe all documents.

Images, and more generally video shots, can be represented by such a vector space model [10, 8] that will be denoted Image Vector Space Model. For this purpose, images are first segmented into homogeneous regions that will be considered as the smallest entity describing the content, i.e. words. As illustrated in the scheme 1(a), features are extracted from segmented regions, next they are quantized to end up with *visual terms* composing a visual dictionary. The scheme 1(b) illustrates the workflow of the indexing process that allows to represent video shots in the Image Vector Space Model defined on the previously constructed dictionary. Each video shot is finally represented by a count vector of its composing *visual terms*.

3.1. Segmentation

Frames are automatically segmented thanks to the algorithm proposed by Felzenszwalb and Huttenlocher [6] to efficiently compute a good segmentation. The important advantage of the method is its ability to preserve details in low-variability images while ignoring details in high-variability images.

Moreover the algorithm is fast enough to deal with a large number of frames.

3.2. Region features

Regions are modeled by two types of features proven effective in their category [11] for content-based image retrieval:

- The color feature is described by a hue, saturation and value histogram with 4 bins for each channel,
- We use 24 Gabor's filters at 4 scales and 6 orientations to capture the texture characteristics in frequency and direction. The texture feature vector is composed of the output energy of each filter.

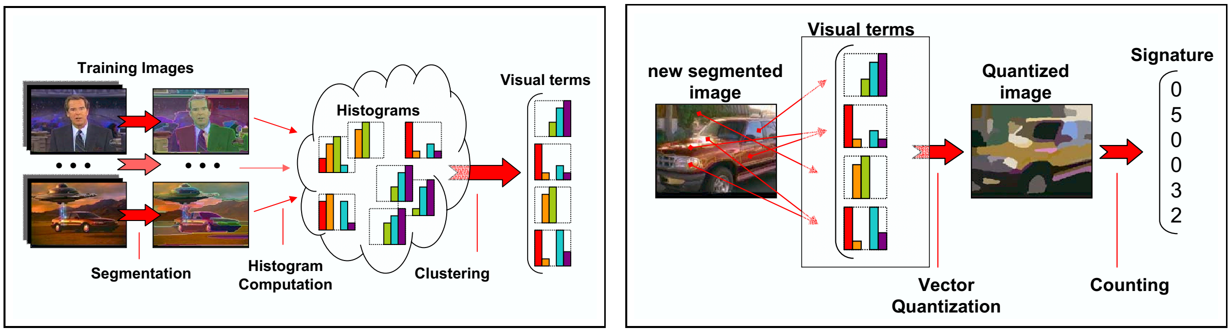
These visual features are then processed independently for two reasons. Firstly, combining features increases the variability of the data rendering more difficult the quantization task that follows. Secondly features can be more efficiently combined at the end with respect to the task. Different metrics can then be used or different weights can be assigned to different features by users, learning algorithms or a relevance feedback loop. Next sections are then presented for one feature and the same processing is applied to the other. The presented method can easily be extended to other features in order to complete the description of the content.

3.3. Quantization

This operation consists in gathering regions having a similar content with respect to low-level features. The objective is then to have a compact representation of the content without sacrificing much accuracy. For this purpose, the k-means algorithm is used with the euclidian distance. We call *visual terms* the representative regions obtained from the clustering and *visual dictionary* the set of *visual terms*. For each region of a frame, its closest visual term is identified and the corresponding index is stored discarding original features.

3.4. Indexing and Comparison

The indexation of new video shot is easy in this framework. First the video shot is segmented and region features are extracted. Each region is mapped to its closest *visual term*. Finally the video shot is indexed by the count vector of *visual*



(a) Quantization

(b) Indexing

Fig. 1. Image Vector Space Model principle for video content indexing.

terms composing the video shot. The natural measure to compare video shots in this framework is the scalar product that emphasizes common *visual terms*. However we would rather use its normalized form the cosine function that further highlights the relative amount of common content between video shots.

In oder to deal with both features that were processed independently, we propose to compute a unique similarity score between two shots as a weighted sum. Weights can then be used in a interactive environment to favor one feature type such as color over the other. In experiments, they will be set to one.

4. EARTH MOVER'S DISTANCE

The Earth Mover's Distance metric was introduced for image retrieval in [14]. It is based on the minimal cost that must be paid to transform a distribution into the other. It is more robust than histogram matching techniques, in the sense that it can operate on variable-length representations of the distribution, avoiding quantization. It also naturally extends the notion of a distance between single elements to that of a distance between sets or distributions of elements. Unfortunately, such a representation requires to have all feature vectors for all regions of the database. For large databases, it is hardly feasible and doing the segmentation and feature extraction on the fly would still increase dramatically computer resources. The solution is then to work on the Image Vector Space Model. Images are represented by the vector of *visual terms*. And the EMD uses *visual terms* features instead of the original features of the region [8]. Then, only features of *visual terms* have to be saved and regions are indexed by their sparse vector of *visual terms*. Since its introduction for image retrieval, this distance have been widely used in the field despite its expensive computation requirements.

Computing the EMD is based on a solution to the transportation problem. Let $P = \{(p_1, w_{p1}), ..., (p_m, w_{pm})\}$ be the first signature with m regions described by p_i and with a weight w_{pi}. Let $Q = \{(q_1, w_{q1}), ..., (q_n, w_{qn})\}$ be the second set of regions and $D = [d_{ij}]$ the ground distance matrix where d_{ij} is the ground distance between p_i and q_j. In this paper the ground distance is defined by the euclidian distance between two region feature vectors. We want to find a flow $F = [f_{ij}]$ with f_{ij} the flow between p_i and q_j that minimizes the overall cost:

$$C(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \tag{1}$$

subject to the following constraints:

$$f_{ij} \geq 0. \forall (i, j) \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{pi}, \forall i \tag{3}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{qj}, \forall j \tag{4}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}) \tag{5}$$

Once the transportation problem is solved [14], the EMD is defined as:

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{6}$$

As for the Image Vector Space Model, in order to get a unique distance value over color and texture features, a weighted sum is used.

5. LATENT SEMANTIC INDEXING

Latent Semantic Analysis (LSA) has been proven efficient for text document analysis and indexing. As opposed to early information retrieval approaches that used exact keyword matching techniques, it relies on the automatic discovery of synonyms and the polysemy of words to identify similar documents. We proposed in [16] an adaptation of LSA to model the visual content of a video sequence for object retrieval.

Let $V = \{S_i\}_{1<i<N}$ be a sequence of shots representing the video. Usually many shots contain the same information but expressed with some inherent visual changes and noise. The noise is generated by multiple sources from the visual acquisition system to the segmentation and clustering processes. Latent Semantic Analysis is a solution to remove some of the noise and find equivalences of the visual content to improve shot matching. It relies on the occurrence information of some features in different situations to discover synonyms and the polysemy of features. A common approach is to use the singular value decomposition (SVD) of the occurrence matrix of features in shots to achieve this task.

Shots are represented by the count vector of *visual terms* that describes the content of their regions. Let now denote q this feature vector. The singular value decomposition of the occurrence matrix C of visual terms in video shots gives:

$$C = UDV^t \quad \text{where} \quad U^tU = V^tV = I \tag{7}$$

With some simple linear algebra we can show that a shot (with a feature vector q) is indexed by p such that:

$$p = U^t q \tag{8}$$

U^t is then the transformation matrix to the latent space. The SVD allows to discover the latent semantic by keeping only the L highest singular values of the matrix D and the corresponding left and right singular vectors of U and V. Thus,

$$\hat{C} = U_L D_L C_L^t \quad and \quad p = U_L^t q \tag{9}$$

The number of singular values kept drives the LSA performance. On one hand if too many factors are kept, the noise will remain and the detection of synonyms and the polysemy of visual terms will fail. On the other hand if too few factors are kept, important information will be lost degrading performances. Unfortunately no solution has yet been found and only experiments allow to find the appropriate factor number. Figure 2 shows the process of LSI.

Finally shots are directly compared in the singular space. Let $f_q = (f_{i,k_i})_{1\leq i\leq n}$ be the representation of a shot with different features such as color and texture. f_{i,k_i} is the feature vector of i projected on the singular space of i whose size is k_i. We compute the weighted sum of cosine values over each feature. Thus the similarity value between q and q' is,

$$sim(q,q') = \sum_{i\in\{color,texture\}} w_i \cos(f_{i,k_i}, f'_{i,k_i}) \tag{10}$$

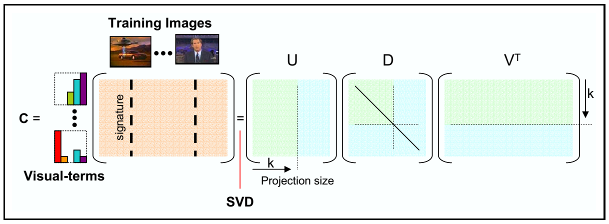


Fig. 2. Latent Semantic Indexing workflow.

This formulation is interesting since it does not only allow to dynamically select the weights between features but also to select the projection size.

6. COMPARISON

Selected approaches are compared in the framework of content-based video shot indexing and retrieval. In order to evaluate their ability to retrieve objects, we have manually selected seven characters through a video sequence (figure 3). Thus, the 130 possible queries are composed only of objects regions without the background. Performances are measured using average mean precision values that allow an easy comparison of systems.

First experiments (figures 4 and 5) show the impact of the dictionary size on system performances. These preliminary experiments are interesting for both approaches since they highlight the effect of the quantization which is generally required to reduce storage requirements.

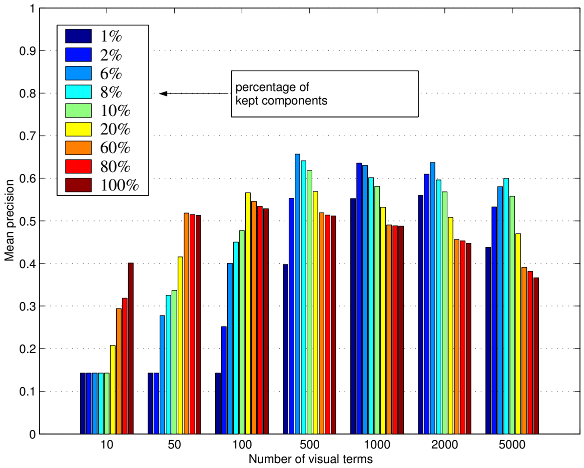
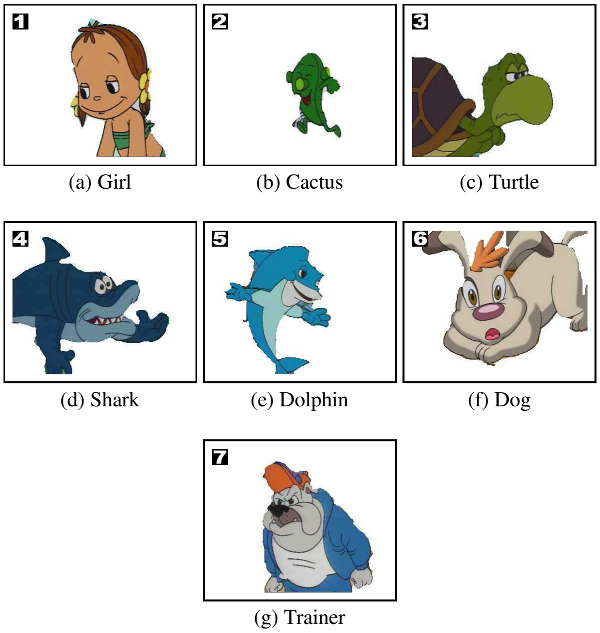


Fig. 4. Latent Semantic Indexing relying on the Image Vector Space Model

Figure 4 illustrates the evolution of the average mean precision using LSI. Results are obtained for all objects with respect to two variables: the dictionary and the projection sizes. This figure shows that performances are quite stable with respect to both variables. Furthermore, it shows



(a) Girl (b) Cactus (c) Turtle

(d) Shark (e) Dolphin (f) Dog

(g) Trainer

Fig. 3. Manually selected objects for the evaluation. Source: Docon Production donation to the MPEG-7 dataset

the positive effect of the projection that allows to improve retrieval performances. We can observe that the best stability and performances are given for 1000 clusters. Stability is an important criterion since both parameters are selected heuristically in practical. Thus, best performances should not be the result of a very fine tuning otherwise the comparison with other systems will not be reliable.

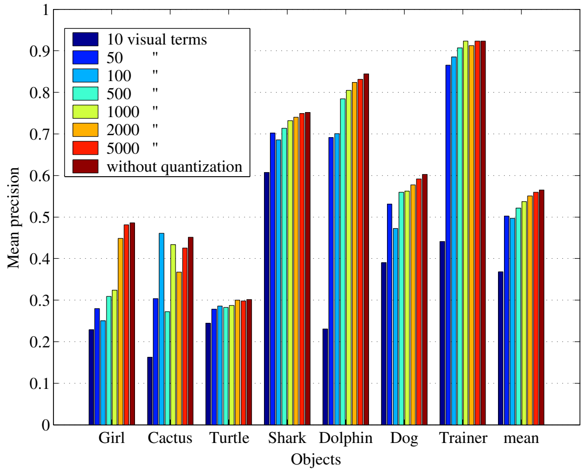


Fig. 5. Earth Mover's Distance relying on the Image Vector Space Model

Figure 5 illustrates the decrease in performance due to the quantization process before the EMD computation. The quantization is not required by the method itself but by storage constraints. As we can see, a quantization with more than 500 *visual terms* does not greatly influence performances. The quantization can be realized with a limited impact. This remark is important since it is particularly expensive to save extracted features for all regions and all frames, and the main solution is to quantized features.

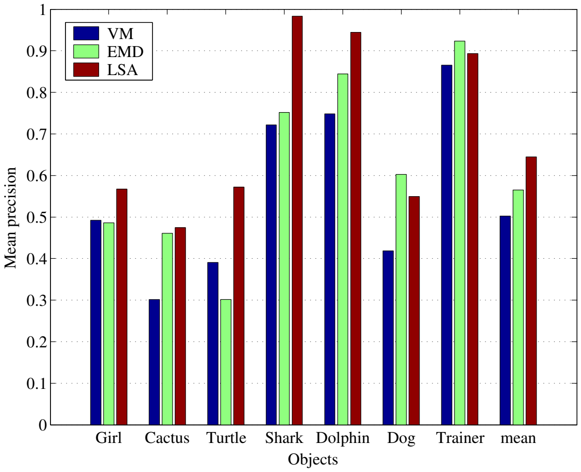
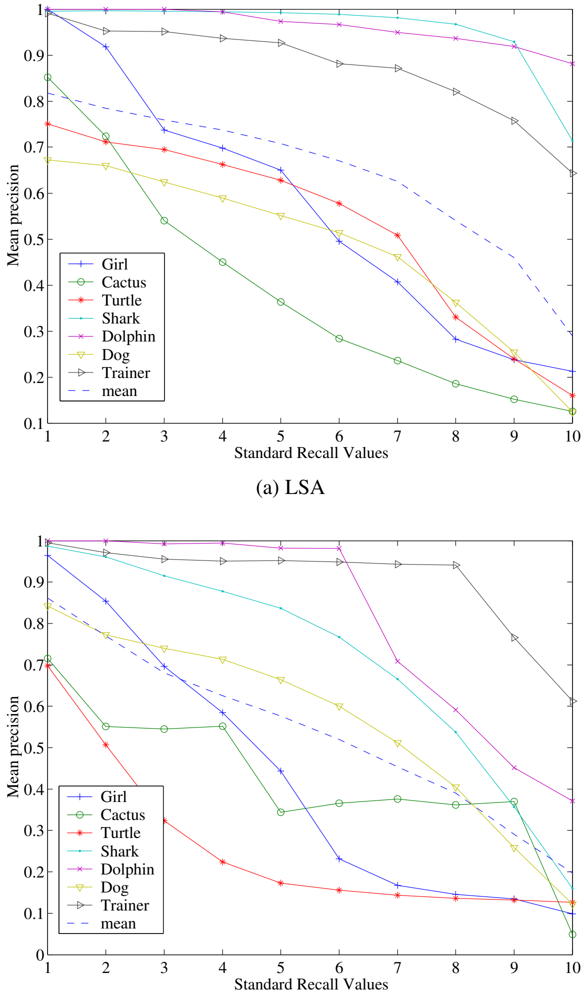


Fig. 6. Comparison of the IVSM, LSA and EMD

Finally, the plot 6 compares presented approaches. The Image Vector Space Model is the basic approach that consists in indexing the content with a count vector of predefined *visual terms*. The EMD-based method directly uses region features to compare contents. The LSI-based method enhances the IVSM approach by automatically inducing *visual terms* similarities. As expected the worst performances are provided by the IVSM approach. Indeed, this method implies a quantification process followed by the lost of quantized values that are replaced by their index. At the end, the similarity function just counts the number of common regions. The EMD-based method performs well. For most objects it outperforms the Image Vector Space Model representation and surprisingly it is not the case for queries on the turtle. LSI-based method outperforms both approaches with an average gain of 15% and 8%. For the mean queries on the dog and the trainer, LSI performances are slightly weaker than the one with EMD. In order to better understand what is happening, precision vs recall curves are plotted for each system and object.



(a) LSA

(b) EMD

Fig. 7. Mean precision vs recall curves for the different objects

Figure 7 consists in two group of plots that represent the evolution of precision and recall values for each object. We can observe that the EMD-based system has higher precision values for small recall values. Then, the precision quickly decreases. The LSI-based system begins with smallest precision values but keeps its good performances longer. Since the EMD computes a distance directly on region features, it is more selective and less tolerant to changes. At the opposite, the LSI-based system works on an approximated content that allows to retrieve content that are slightly different from the query but still relevant.

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed to compare two region-based indexing techniques that have different properties. The first technique is the Latent Semantic Indexing method that we have previously introduced to index video shots. This method is based on an Image Vector Space Model which is enhanced by the Latent Semantic Indexing. The second technique is an Earth Mover's Distance based system. Features of segmented frames are directly used by the EMD to compute their distance. To summarize, the first approach implies the quantization of features and then the lost of feature values while the second approach keeps the original content.

Surprisingly results have shown that the LSI-based system was outperforming the EMD-based system. We further explained this fact and their different behaviors with the help of precision vs recall curves.

Future work will concern the comparision of both methods on real video and the investigation of an hybrid system. The interest will be do combine the precision of an EMD-based system with the flexibility of an LSI-based system.

8. REFERENCES

[1] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia Tools Applications*, 4(1):29–56, 1997.

[2] Chad Carson, Megan Thomas, and Serge Belongie. Blobworld: A system for region-based image indexing and retrieval. In *Third international conference on visual information systems*, 1999.

[3] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 8, pages 602– 615, 1998.

[4] 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*, 41(6):391–407, 1990.

[5] Christos Faloutsos, Ron Barber, Myron Flickner, Jim Hafner, Wayne Niblack, Dragutin Petkovic, and William Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3/4):231–262, 1994.

[6] P. Felzenszwalb and D. Huttenlocher. Efficiently computing a good segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–104, 1998.

[7] Lukas Hohl, Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Enhancing latent semantic analysis video object retrieval with structural information. In *Proceedings of the IEEE International Conference on Image Processing*, 2004.

[8] Feng Jing, Mingling Li, Hong-Jiang Zhang, and Bo Zhang. An effective region-based image retrieval framework. In *Proceedings of the ACM International Conference on Multimedia*, 2002.

[9] Mikko Kurimo. Indexing audio documents by using latent semantic analysis and som. In Erkki Oja and Samuel Kaski, editors, *Kohonen Maps*, pages 363–374. Elsevier, 1999.

[10] Joo-Hwee Lim. Learning visual keywords for content-based retrieval. In *IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 169–173, 1999.

[11] Wei-Ying Ma and Hong Jiang Zhang. Benchmarking of image features for content-based image retrieval. In *Thirty-second Asilomar Conference on Signals, System and Computers*, volume 1, pages 253–257, 1998.

[12] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 536–540, 1998.

[13] A.P. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[14] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 59–66, january 1998.

[15] Gerard Salton, A. Wong, and C.S. Yang. A vector space model for information retrieval. *Communications of the ACM*, 18(11):613–620, November 1975.

[16] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Video content modeling with latent semantic analysis. In *Third International Workshop on Content-Based Multimedia Indexing*, 2003.

[17] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Latent semantic analysis for an effective region-based video shot retrieval system. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2004.

[18] Howard Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. Intelligent access to digital video: The informedia project. *IEEE Computer*, 29(5), 1996.

[19] James Z. Wang, Jia Li, and Gio Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[20] Rong Zhao and William I Grosky. From features to semantics: Some preliminary results. In *Proceedings of the International Conference on Multimedia and Expo*, 2000.