

# TOWARDS A CONTENT-BASED VIDEO RETRIEVAL SYSTEM USING WAVELET-BASED SIGNATURES

Oscar D. Robles

Dpto. de Informatica, Estadística y Telemática.  
U. Rey Juan Carlos. C. Tulipán, s/n.  
28933 Móstoles. Madrid. Spain.  
orobles@escet.urjc.es

Angel Rodríguez

Dept. de Tecnología Fotónica.  
U. Politécnica de Madrid.  
Campus de Montegancedo s/n.  
28660 Boadilla del Monte. Madrid. Spain.  
arodri@dtf.fi.upm.es

Pablo Toharia

Dpto. de Informatica, Estadística y Telemática.  
U. Rey Juan Carlos. C. Tulipán, s/n.  
28933 Móstoles. Madrid. Spain.  
ptoharia@escet.urjc.es

Luis Pastor

Dpto. de Informatica, Estadística y Telemática.  
U. Rey Juan Carlos. C. Tulipán, s/n.  
28933 Móstoles. Madrid. Spain.  
lpastor@escet.urjc.es

## ABSTRACT

This paper presents two new primitives for representing the content of a video in order to be used in a Content-Based Video Retrieval System. The techniques presented here compute first a multiresolution representation using the Haar transform. Two types of signatures are extracted afterwards, one based on multiresolution global color histograms and the other one based on multiresolution local color histograms. The tests performed in the experiments include the recall measure achieved with the proposed primitives using a database composed by 62 videos with 817 shots.

## KEY WORDS

CBIR primitives, Wavelet transform

## 1 Introduction

The technological evolution of communication systems and the degree of maturity achieved in so different areas such as signal processing, databases or computer vision, has brought about the proliferation of information systems whose objective is the efficient storage and management of large amounts of multimedia data [1, 2, 3].

In this type of systems it is very important to search for information as a function of its content. The extended use of digital video and the need of an efficient management have given rise to Content-Based Video Retrieval Systems<sup>1</sup> [4, 5, 6, 7, 8]. The main purpose of this type of systems is to extract some kind of information that summarizes the video content as much as possible.

The process of extracting the video content can be usually decomposed in two basic steps:

- A time segmentation stage that allows the identification of meaningful units, like shots, episodes or scenes.

- A content analysis stage aiming to characterize regions, objects and movement over each one of the computed video shots [9].

In this paper we propose new signatures to represent the content of video data once we have previously made a segmentation process (described in [10]). Once the video has been temporally segmented in elemental unities named *shots*, we summarize the shot contents in a representative frame named *key frame*. The signatures collect multiresolution color information of the key frames extracted from the segmented shots.

## 2 CBVR system overview

Two main aspects must be considered in order to implement a CBVR system:

- The way video content is represented.
- What kind of queries are allowed in the system.

We will discuss both aspects in the following sections.

### 2.1 Video signature generation

As mentioned above, there is a process where the video content is summarized in a symbolic representation named *signature*. Once the video is segmented in shots, the content of each shot is processed to generate a set of concatenated symbolic representations, one per detected shot, that composes the whole video signature. There is no general agreement on the way key frames should be selected, so several approaches can be followed [4, 9, 11, 7, 8]. Some authors choose the first frame, the last one, or one of the inner frames randomly selected. Another approach is to select the key frame taking into consideration low level features extracted from the central frames.

<sup>1</sup>CBVR from now onwards.

In our case, we have chosen to make an automatic extraction of the key frame computing the differences between consecutive frames and selecting as the key frame the one that is the minimal of the whole sequence of differences inside the shot. This strategy has the advantage of being a complementary criterion of the shot detection process, so both computations can be done in a single step [10].

## 2.2 CBVR system operation

When the user introduces a query in a CBVR system, the result from the search is a list ordered according to the level of similarity of its signature with the signatures stored in the video database. This list contains the best  $n$  matches of the query. CBVR systems accept two different inputs taking into consideration the nature of the data involved in the query: static images and sequences of images that compose a shot extracted from a video. The problem of video matching, defined as the process that allows the comparison between pairs of video sequences, is far away from the scope of this paper.

In the first case, the problem to be solved consists on finding the video that best matches the searched frame, comparing the signature from the query with respect to the signatures of the key frames extracted from the database. The output of the system can be a sorted list of the  $n$  most similar key frames. When dealing with shots as input data, a preprocessing step is required in order to extract the key frame of the query. Then, it will be compared against the video database signatures in a similar way to the frame case. In both cases the user can select the detail of the output: video or shot level. At video level, the system shows only the video key frame, achieving only one key frame per video, but at shot level, the system returns a key frame per shot. In this case, it may appear several frames belonging to the same video.

Should the users then consider the search result to be unsatisfactory, they may select as a new input one of the displayed key frames which is most similar to the original query and then restart the retrieval process. Figure 1 depicts a scheme of the whole process. In both cases, the classification method used in the matching process has been a minimum distance classifier [12].

## 3 Description of the wavelet based signatures

The development of the wavelet transform theory has spurred new interest in multiresolution methods and has provided a more rigorous mathematical framework. Wavelets give the possibility of computing compact representations of functions or data. Additionally, they allow variable degrees of detail or resolution to be achieved, and they are attractive from the computational point of view [13, 14, 15].

The features selected to represent the content of

the key frames are color multiresolution histograms [16]. These histograms are computed over the analysis coefficients of the wavelet transform for each color channel of the multiresolution representation. The purpose of these features is to define a primitive that represents the color information of the original frame at different resolution levels.

Two complementary approaches have been tested: global and local multiresolution histogram computation. For the global approach, we extract the histograms from the analysis coefficients at each step of the key frame Haar transform over each one of the three color channels. In order to fuse the information proceeding from each resolution level, we propagate the values computed at the lower levels weighting this values by a factor that depends on the considered resolution level:

$$h(k)_{\{R,G,B\}} = \frac{1}{k} \sum_{x,y} I_{\{R,G,B\}}^{(i)}(x,y) \cdot 4^{j-i} \quad (1)$$

$$\forall (x,y) \mid I_{\{R,G,B\}}^{(i)}(x,y) = k, k = 0, \dots, 255,$$

being  $I_{\{R,G,B\}}^{(i)}$  the analysis coefficients of the transformed image at resolution level  $i$ ,  $j$  the resolution level of the original image and  $(x,y)$  the coordinates of each coefficient. The expression of the computed histograms is

$$\hat{h}(k) = \frac{h(k)}{n} \quad (2)$$

where  $n = \sum_k h(k)$ . Figure 2 shows an example comparing color histograms of a key frame and the multiresolution color histograms of the same frame. A visual analysis shows an offset of the multiresolution distribution around the mean value and an enhancement of the peaks for each color channel. The comparison of histograms from different key frames is done computing the area of the histograms intersection.

Local histograms are computed in order to increase the level of description detail and the discrimination power achieved by the global histograms. In this case, the original key frames are divided into 9 non-overlapping regions. This produces a  $3 \times 3$  lattice of adjacent cells which multiresolution histograms are computed in. The comparison between two key frames is done accumulating the similarity measure achieved by each color histogram for the nine region couples, i.e., 27 histograms.

Once the histograms have been computed for each of the video key frames, we have available all the information to generate the video signature. This signature is specified using the standard description language XML<sup>2</sup>. An XML file describing the signature of a video will include the following items:

- Video file name.
- General features like width and height of the frames, number of frames, etc.

<sup>2</sup>eXtensible Markup Language

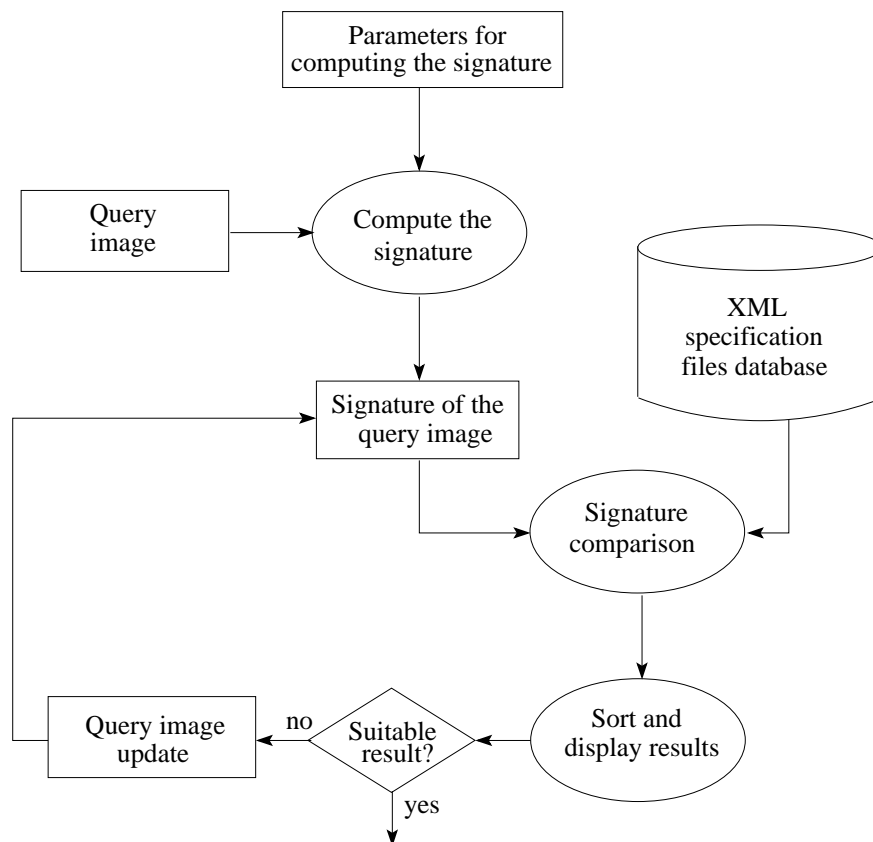


Figure 1. CBVR system operation.

- A list of the files keeping the key frames signatures.
- Finally, a list of video shots with the ordinals of both ends of the shot and the corresponding key frame.

A more detail description of the structure of this file can be found in [17].

## 4 Experimental Results

### 4.1 Experiments Setup

The main goals of the tests are to measure and analyze the recall value of the implemented signatures making both video and shot retrieval.

The recall is being accepted as one of the standard measures of performance. It can be defined as the ratio of relevant items retrieved in a query [2, 18, 19]:

$$\text{Recall} = \frac{\text{No. of relevant items retrieved}}{\text{Total relevant in the collection}} \quad (3)$$

The tests have been performed using a database composed by 62 general AVI videos. On a first stage, the videos have been manually segmented in 817 shots. Then, a frame has been randomly extracted from each segmented shot, developing a test set of 817 well-known classified query

frames. Finally, we have computed the response of the system for each one of the 817 frames at video or shot level output.

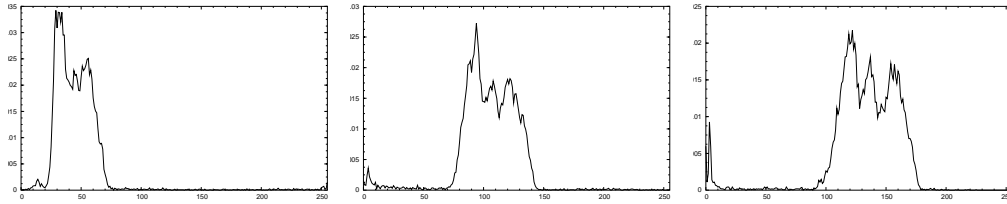
When dealing with the system evaluation it has to be noticed that it is very difficult to compare results from different retrieval systems. On one hand, the absence of a unified test set makes it impossible to check the system herein presented against different approaches from the literature in the same conditions. On the other hand, there are promising attempts in this way, like TRECVID [20]. TRECVID distributes among its participants a previously classified test set, but it has not been available for checking the primitives described above at the present moment. Anyway, the TRECVID project is mainly focused on interactive retrieval, preventing a direct comparison with the fully automated approach presented in this paper.

### 4.2 Results analysis

Figure 3(a) shows the accumulated recall value for the implemented signatures —global multiresolution histograms (labeled *Global histo.*), and local multiresolution histograms (labeled *Local multires. histo.*)— at video level, and Figure 3(b) presents the results achieved at shot level. The information represented by each curve of the Figures is the percentage of test set



(a) Original image

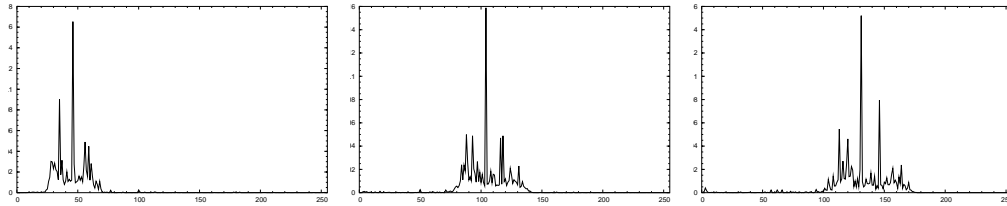


(b) Red

(c) Green

(d) Blue

Color histograms



(e) Red

(f) Green

(g) Blue

Multiresolution color histograms

Figure 2. Comparison between color and multiresolution color histograms.

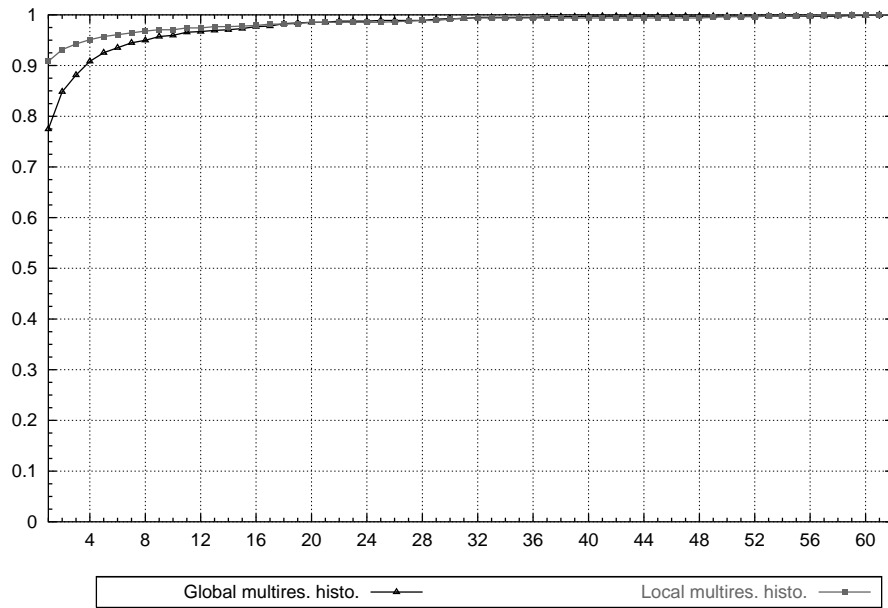
videos or shots which are placed up to the given position when a query is made. The abscissa axis shows the position obtained by the arrangement which produces the result of the query. The ordinate axis shows the accumulated frequencies histogram of the percentage of objects placed in a position up to that one pointed by the abscissa value. The values from Figure 3(a) are computed considering as the output value the minimum matching from all the key frames that compose the video signature.

Looking at Figure 3, it is possible to deduce that this solution achieves a high degree of success for both primitives, but local multiresolution surpasses global multiresolution histograms, achieving the 90% of the queries at video description level on first position, while at shot level, this value is achieved considering that the correct shot is included in a set composed by 4 elements.

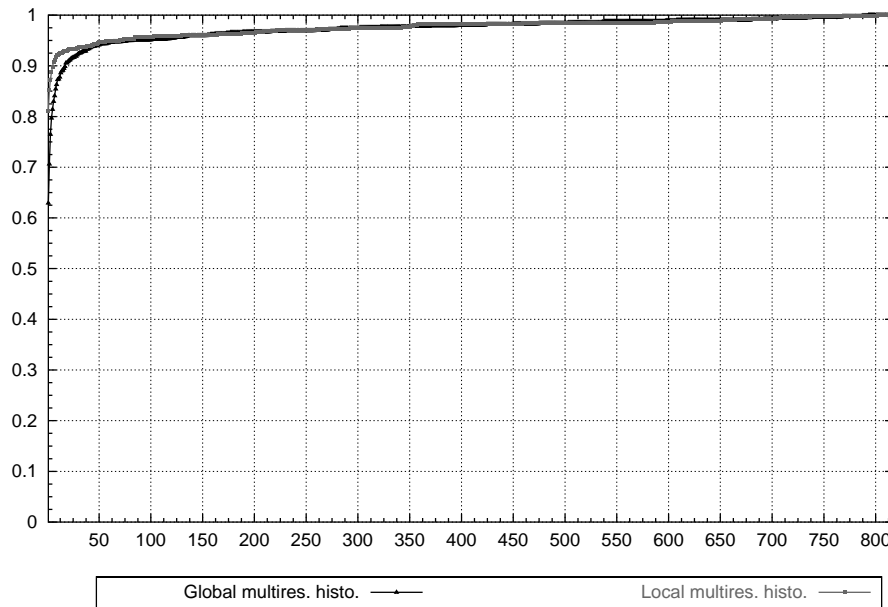
## 5 Conclusions and future work

In this paper, new multiresolution primitives for video content based retrieval have been presented. The primitives have been tested on a video database which contains 62 general thematic videos and 817 manually segmented shots. A remarkable feature of the implemented primitives are their good recall results.

Further research will take into consideration the problem of comparing pairs of video sequences, usually referred to as video matching. Another important point will be to provide the system with the capability to combine the described primitives with different new ones. Finally, parallel implementations of the studied primitives will also be considered.



(a) Video level.



(b) Shot level.

Figure 3. Recall measure. The abscissa axis represents the position obtained by the result of the query. The ordinate axis shows the accumulated frequencies histogram of the percentage of objects placed in a position up to that one pointed by the abscissa value.

## Acknowledgments

This work has been partially funded by the Spanish Commission for Science and Technology (grants CICYT TIC2002-04486-C02-02 and TIC2003-08933-C02-01).

## References

- [1] Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman, editors. *Proceedings of the SPIE/EI'2000 Symposium on Storage and Retrieval for Media*

- Databases*, volume 3972. SPIE, January 2000. ISBN 0-8194-3590-2.
- [2] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on PAMI*, 22(12):1349–1380, December 2000.
- [3] M. Luisa Córdoba, Oscar D. Robles, Angel Rodríguez, M. Isabel García, María S. Pérez, Manuel Nieto, Antonio Pérez, Luis Pastor, and José L. Bosque. PACOBIR: A parallel CBIR system. *Parallel Computing*, page 21. Accepted, in press.
- [4] R. Brunelli, O. Mich, and C. M. Modena. A survey on video indexing. *Journal of Visual Communication and Image Representation*, 10:78–112, 1999.
- [5] Y. Alp Aslandogan and Clement. T. Yu. Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):56–63, Jan. - Feb. 1999.
- [6] Rainer Lienhart, Wolfgang Effelsberg, and Ramesh Jain. VisualGREP: A systematic method to compare and retrieve video sequences. *Multimedia Tools and Applications*, 10(1):47–72, January 2000.
- [7] Sameer Antani, Rangachar Kasturi, and Ramesh Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945–965, April 2002.
- [8] Oge Marques and Borko Furht. *Content-Based Image and Video Retrieval*. Multimedia Systems and Application Series. Kluwer Academic, 2002. ISBN 1-4020-7004-7.
- [9] Alberto del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, California, 1999. ISBN 1-55860-624-6.
- [10] Oscar D. Robles, Pablo Toharia, and Angel Rodríguez. Automatic video cut detection using adaptive thresholds. In *4th IASTED International Conference on Visualization, Imaging, and Image Processing - VIIP 2004*, Marbella, Spain, September 2004. IASTED. To be submitted.
- [11] Hongjiang Zhang. Video content analysis and retrieval. In C. H. Chen, L. F. Pau, and P. S. P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, chapter 5.5, pages 945–977. World Scientific Publishing Company, 1998.
- [12] Angel Rodríguez, Oscar D. Robles, and Luis Pastor. New features for Content-Based Image Retrieval using wavelets. In Fernando Muge, Rogério Caldas Pinto, and Moisés Piedade, editors, *V Ibero-american Symposium on Pattern Recognition, SIARP 2000*, pages 517–528, Lisbon, Portugal, September 2000. ISBN 972-97711-1-1.
- [13] Adam Finkelstein, Charles E. Jacobs, and David H. Salesin. Multiresolution video. pages 281–290, August 1996.
- [14] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1997.
- [15] W. Sweldens. The lifting scheme: A constructin of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, 1997.
- [16] Oscar D. Robles, Angel Rodríguez, and M. Luisa Córdoba. A study about multiresolution primitives for content-based image retrieval using wavelets. In M. H. Hamza, editor, *IASTED International Conference On Visualization, Imaging, and Image Processing (VIIP 2001)*, pages 506–511, Marbella, Spain, September 2001. IASTED, ACTA Press. ISBN 0-88986-309-1.
- [17] Pablo Toharia, Angel Rodríguez, and Oscar D. Robles. Xml specification for avi files in a content-based video retrieval system. In *4th IASTED International Conference on Visualization, Imaging, and Image Processing - VIIP 2004*, Marbella, Spain, September 2004. IASTED. To be submitted.
- [18] Stephen Robertson. Evaluation in information retrieval. In *Fourth European Summer School in Information Retrieval, ESSIR'03. Student Booklet*. ESSIR, 31 August – 5 September 2003.
- [19] Fuhui Long, Hongjiang Zhang, and David Dagan Feng. Fundamentals of content-based image retrieval. In D. Feng, W. C. Siu, and H. J. Zhang, editors, *Multimedia Information Retrieval and Management. Technological Fundamentals and Applications*, Multimedia Signal Processing Book, chapter 1, pages 1–26. Springer-Verlag, Berlin Heidelberg New York, 2003. ISBN 3-540-00244-8.
- [20] Alan F. Smeaton. TRECVID 2003 video evaluation overview. Presented at the TRECVID 2003 Conference, National Institute for Standards and Technology, November 2003. <http://www-nlpir.nist.gov/projects/tvpubs/papers/tv3.overview.slides.pdf>.