Scalable Object-Based Indexing of HD Videos: A JPEG2000-Oriented solution

Cl. Morand*, J. Benois-Pineau, J-Ph. Domenger

LaBRI UMR CNRS/ University of Bordeaux 1, 351 Cours de la Liberation, F-33405 Talence

Abstract

Video indexing technique is crucial in multimedia applications. In the case of HD (High Definition) Video, the principle of scalability is of great importance. The wavelet decomposition used in the JPEG2000 standard provides this property. In this paper, we propose a scalable descriptor based on objects. First, a scalable moving object extraction method is constructed. Using the wavelet data, it relies on the combination of a robust global motion estimation with a morphological color segmentation at a low spatial resolution. It is then refined using the scalable order of data. Second, a descriptor is built only on the objects found at the previous step. This descriptor is based on multiscale histograms of wavelet coefficients of moving objects.

Key words: JPEG2000, object-based indexing, scalable indexing, scalable objet extraction

1. Introduction

Created in march 2002, the DCI (Digital Cinema Initiative, LLC [1]) is the joint venture of six American cinema majors. Its primary purpose is to establish voluntary specifications for an open architecture for digital cinema that ensures a uniform and high level of technical performance, reliability and quality control. The specifications make JPEG2000 [2] the digital cinema compression standard. Besides, using JPEG2000 is studying for video archiving [3], to conserve cultural patrimony with the greater compromise quality/compression possible. Nowadays, data coded with this standard constitute databases of audio-visual content so large that access to digital content requires the development of automatic methods for processing and indexing multimedia documents. One of the key features consists in extracting meaningful information allowing organizing the multimedia content for easy manipulation and/or retrieval tasks. A variety of methods [4, 5] have recently been developed to fulfill this objective, mainly using global features of the multimedia content such as the dominant color in still images or video key-frames. Object-based video indexing still remains a challenge as extraction of semantic objects from video is an open problem. Moreover, it is clear that the precision requirements and complexity constraints of object extraction methods are strongly application dependent. In the context of digital libraries containing compressed video, an effective object indexing from compressed video still remains a challenge.

The first step in object-oriented indexing is the foreground object-extraction. Several approaches have been proposed in the past, and most of them can be roughly classified either as *intra-frame segmentation based* or as *motion segmentation based* methods. In the former approach, each frame of the video sequence is independently segmented into regions of homogeneous intensity or texture, using traditional image segmentation techniques [6], while in the latter approach, a dense motion field is used for segmentation and pixels with homogeneous motion field are grouped together [7]. Since both approaches have their drawbacks, most object extraction tools combine spatial and temporal segmentation techniques [8, 9]. Challenge resides in applying this scheme without decompressing the video, as the low-level content descriptors, such as coefficients in the transform domain, can be efficiently re-used for the content analysis task [10].

In the case of the MJPEG2000, one difficulty is that the standard does not provide motion descriptors and so they have to be estimated. The ME (Motion Estimation) problem in the wavelet domain has largely been studied

*Corresponding author: Tel +33(0)5 40 00 38 80; fax +33(0)5 40 00 66 69

Email addresses: morand@labri.fr (Cl. Morand), jenny.benois@labri.fr (J. Benois-Pineau), domenger@labri.fr (J-Ph. Domenger)

in the literature [11]. In the case of the RI (rough indexing) paradigm [10] we are working in, the only data available are those contained in the compressed stream, which means that the wavelet basis is not analysis-oriented. From the scalability point of view, only part of the stream and not the entire stream can be available. Then, initially, only low resolution wavelet coefficients are avalaible, i.e the ones of the base layer. The decimation operation in the DWT (Discrete Wavelet Transform) makes it shift-variant. Hence, the BM (Block Matching) can be very inefficient in the wavelet domain. The low-band signal is usually smooth and the difference in the low-band coefficients between the original and the shifted signal is small. However, there is a big difference between the high-band coefficients of the shifted signal and those of the original signal. Such phenomena will happen frequently around the image edges. The signal difference in the high-pass signal depends on the amount of shift and the analysis filters for the DWT. The prediction errors of the high-band signal makes it difficult to estimate the motion vectors in the wavelet domain when the conventional block-matching is used. Several motion estimation methods in the wavelet domain have been developed [12, 13]. Among these, direct band-to-band motion estimation of the wavelet coefficients is not efficient because of the shift-variant property of the DWT. There is another approach that performs the motion estimation for only low-band signal, where the motion compensation of the high-band signal is performed with the motion vectors found in the corresponding low-band signal. In order to overcome the shift-variant property, a low-band-shift method has been proposed [12]. It consits in decomposing the reference image not with the DWT but with the ODWT (Overcomplete Discrete Wavelet Transform). The decimation no longer occurs in the reference image and the BM becomes more efficient. In this paper, we present a method for estimating motion in the scalable wavelet domain that uses only low LL resolution information. The method is based on both BM ME and GM (Global Motion) Estimation. This estimation will serve both for GM Estimation and Indexing of HD Video.

A second step in object-oriented indexing consists of defining a global feature on the object effectively found. Following the rough indexing paradigm, this feature has to be defined in the wavelet domain. Several indexing techniques in the wavelet domain exist for JPEG2000 compressed still images. Among them, we can cite the histogram-based techniques. A histogram is computed for each subband and comparison is made subband by subband [14]. The main disadvantage of such a technique is that it works only with limited camera operations. To reduce complexity and improve the robustness to illumination changes, [15] proposes modeling the histograms using a generalized Gaussian density function. Other indexing techniques are texture-oriented [16]. In this work, we propose an histogram-based index that is defined not on the whole image but only on the wavelet coefficients of the object.

The paper is organized as follows. In section 2, the notion of Scalable descriptor is precised. The general framework of our approach is summarized in section 3. Section 4 describes the spatio-temporal object extraction we designed and section 5 the object-based descriptor we propose. Section 6 pesents the results based on retrieval tests. Finally, section 7 concludes our work.

2. Scalable Indexing of Video Content

In video coding, scalability is the ability for a single codestream to be sent to different users with different processing capabilities and network bandwidths by selectively transmitting and decoding the related part of the codestream. Two recent standards for motion pictures compression, H264 and JPEG2000 , present this property; the coded signal allows the easy extraction of sub-bitstreams corresponding to a reduced spatial resolution (spatial scalability), a reduced temporal resolution (temporal scalability), a reduced quality for a given spatio-temporal resolution and/or for a reduced flow (SNR scalability).

A new challenge in Content-Based Video Retrieval (CBVR) is to design scalable descriptors that suit scalable coded streams. Two cases can be distinguished, following the descriptor is computed at the encoder or the decoder end. At the encoder end, the aim is to embed the descriptor in the coded stream. The descriptor has then to be hierarchically structured in a base layer, which concentrates the most relevant information, and refining layers, which contains more detailed information. At the decoder end, the aim is to compute an descriptor with the only available data, i.e. the transmitted part of the stream. The descriptor has to follow the scalable hierarchy of the codestream to allow comparison between videos transmitted at different resolutions.

This paper proposes a scalable descriptor responding to the JPEG2000 decoder-end case. The main difficulty is that the construction should be made in only one direction from low resolution to high resolution.

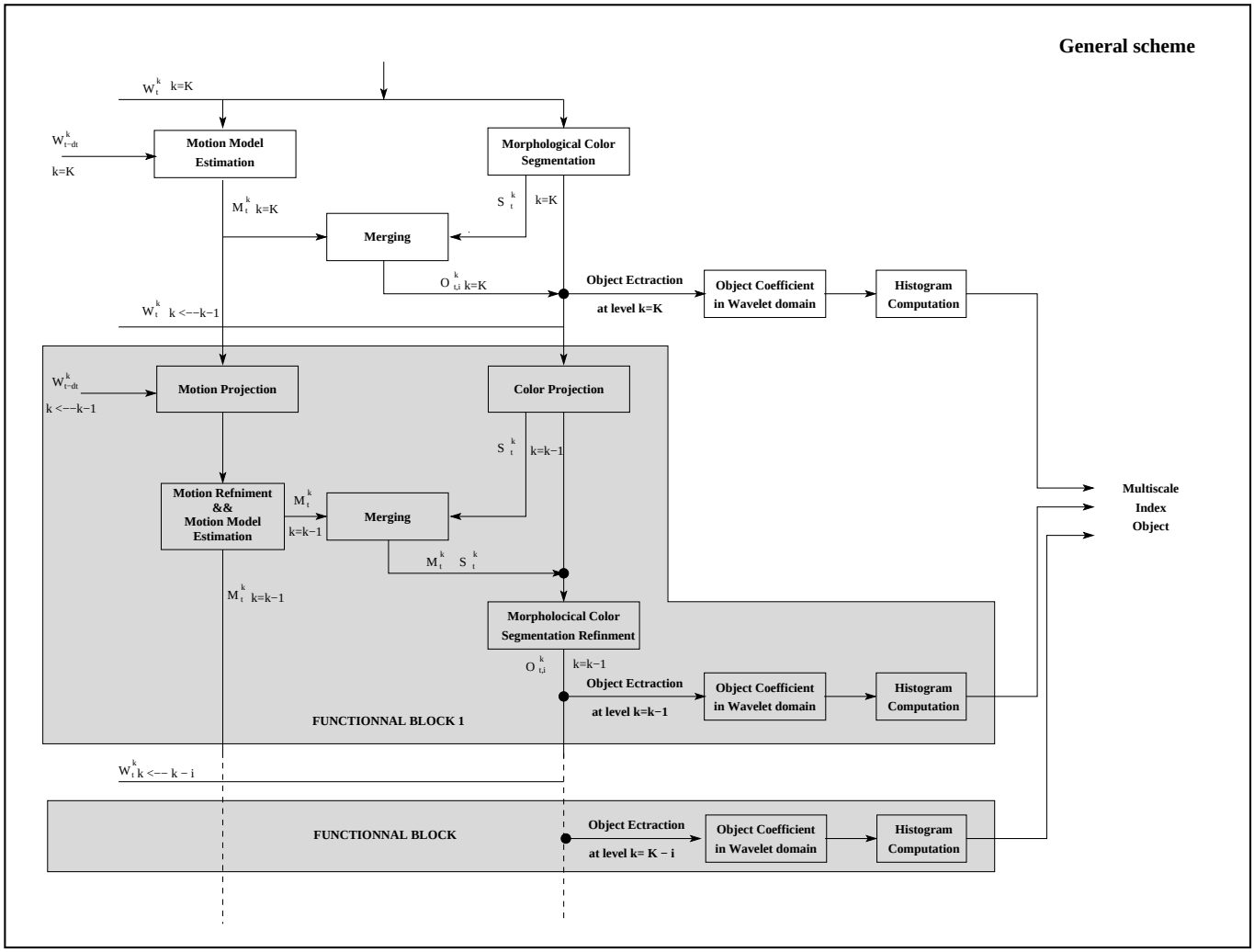

Figure 1: General Scheme of object-Based Scalable Video Indexing

3. General Scheme of Object-Based Scalable Video Indexing

3.1. Notations

Before describing the proposed general framework of the object-based scalable indexing, the notations and abbreviations used in the following of the paper are presented.

The wavelet transform used in the JPEG2000 standard leads to a multiscale representation. We denote K the total number of decomposition layers. k is the considered level with $k = 0$ the original image and $k = K$ the lowest resolution level. At a given level k, a wavelet frame is a combination of four subbands obtained by combination of a Low-pass and a High-pass filtering of the image being decomposed; the four subbands are conventionally noted as LL^k, LH^k, HL^k and HH^k. These subbands are grouped in $LF^k = \{LL^k\}$ -the set of Low Frequendy subbands- and $HF^k = \{LH^k, HL^k, HH^k\}$ -the set of High Frequency Subbands at a given level k.
Given this, W_t^k is the wavelet frame at instant t and level k, S_t^k is the result of a color segmentation, M_t^k is the motion mask, V_t^k is the set of Motion Vectors (MVs) of the whole frame W_t^K. Let $O_{i,t}^k = \{O_{t,i}^k, i = 1..n(k)\}$, where $n(k)$ is the number of objects at level k, be the set of wavelet coefficients representing object i in the wavelet frame at instant t and level k and let $O_{t,i} = \{O_{t,i}^k, k = K..0\}$ be the multiscale representation of wavelet coefficients of object i in wavelet frame at instant t.

3.2. Principle

Our approach in terms of video indexing is led by the RI paradigm we defined in [17]. It can be expressed as a general approach for indexing videos from a compressed stream whatever the compression basis is: MPEG, H.26x (DCT) or JPEG2000 (wavelets). In this paper, we focus on the JPEG2000 standard, that is, on the DWT (Discret Wavelet Transform) domain. Here, 9/7 Daubechies wavelets are used for lossy compression. The scalable streams are supposed not containing ROI (Region of Interest) option of JPEG2000. The global scheme of our scalable solution is depicted in Fig 1.

The JPEG2000 coder is supposed to use K levels of decomposition, $k = 0$ being the original image and $k = K$ the lowest resolution level (In practice $K = 4$). The scheme of Fig 1 depicts both: scalable object Extraction (in grey) and computation of a scalable object descriptor. Scalable object extraction starts at the lowest level of the resolution in the wavelet pyramid ($k = K$) with motion estimation between wavelet frame at t (W_t^k) and $t - dt$ (W_{t-dt}^k). Rough MVs are determined and "motion masks"- in which moving foreground objects are contained- are extracted. In parallel, a morphological color segmentation is carried out. Color segmentation maps S_t^k and motion masks M_t^k are merged. This process is fulfilled independently for each pair of wavelet frames (W_t^k, W_{t-dt}^k), leading to extraction of Objects $O_t^k = \{O_{t,i}^k\}$. To ensure object temporal continuity of object masks, an object matching is fulfilled for each $O_{t,i}^k$ by back projection of $O_{t,i}^k$ to $O_{t-dt,i}^k$ object plan with estimated motion vectors. Hence, the set of objects is being available at the lowest level of wavelet pyramid and can be used for descriptor computation. For higher resolution levels of the pyramid $k = K-1, K-2, \ldots, 0$ the extraction process starts with the projection of the object mask O_t^k, of the segmentation map of the whole frame S_t^k and of the motion vectors V_t^k to W_t^{K-1} resulting respectively in $\hat{O}_t^(k-1)$, $\hat{S}_t^{k-1}$ and $\hat{V}_t^{k-1}$ (see blocks motion projection and Rough Color projection in Fig 1). The extraction of object O_t^k with $k \in [K-1 \ldots, 0]$ consists of refinement of projected color segmentation map $\hat{S}_t^{k-1}$ restricted to the projected object area $\hat{O}_t^{k-1}$: $\hat{S}_t^{k-1} \cap \hat{O}_t^{k-1}$. This refinement is done by Markov random field (MRF) modeling on the areas $\hat{S}_t^k \cap \hat{O}_t^k$ and $\hat{S}_t^{k-1} \cap M_t^k$, where M_t^k is the motion mask detected in W_t^{k-1}. The motion projection is adaptive to the type of region MVs V_t^k belong to: either object area O_t^k or background area $\bar{O}_t^k$. The refinement of motion vectors is done for all blocks. Global ME is done in order to extract new motion mask at level k, M_t^k. Hence, for each level of wavelet decomposition $W_t^k, k = K-1, K-2, \ldots, 0$ object masks $O_{t,i}^k$ are available. The scalable wavelet descriptor is then calculated on these masks (see block histogram computation). The object-based video retrieval can then be performed.

4. Scalable Object-Extraction in the JPEG2000 Wavelet Domain

Scalable Object-Extraction in the JPEG2000 Wavelet Domain follows the general principles we have previously developped [10]. This section summarizes the approach and presents minor changes that have been added to improve the quality of results.

4.1. Scalable Motion Estimation in the Wavelet Domain

The goal of ME here is twofold. Firstly, despite the wavelet pyramid is not a very suitable representation for ME, we need to estimate global, ie camera, motion in the sequence as precise as possible. Second, for object Extraction, we need to precisely label areas in the wavelet frame which do not follow the Global Camera Motion and hence are supposed to contain objects of interest. First, a BM is fulfilled on the Y-component of the LL subband of the *Kth* level of the decomposition. Second, the GM is estimated using a robust estimator which allows outlier rejection. Signal is then synthesized (ie, the LL subband is computed) going down the pyramid level by level. MVs at a current level are initialized using MVs predicted from the previous level according to the results of the GME (see Fig. 2). BM is performed at the current level to refine the MVs found at a previous level and is followed by a GME.

Block Matching To fulfill BM, a backward prediction is used. A full search is made in a fixed-size window. The comparison criterion used in each block is the MAD (Mean of Absolute Differences (1)) on the luminance channel of the LL subband.

$$MAD\,(dx, dy) = \frac{1}{N} \sum_{N} |\mathcal{L}_c\,(x, y) \; - \; \mathcal{L}_r\,(x + dx, y + dy)\,| \tag{1}$$

where N is the number of pixels in the block, (dx, dy) the displacement considered and Y_c and Y_r are the luminance values of the LL subband for the current image and the reference one. The search precision is of 1 pixel. Indeed, the 9/7 Daubechies basis is not designed for BM purpose, and, due to aliasing, a more precise search gives worse results.

Global Motion Estimation The GM, supposed to be equivalent to the Global Camera Motion, is assumed to

follow an affine 6-parameters model described by:

$$\begin{cases} dx(x,y) = a_1 + a_2 x + a_3 y \\ dy(x,y) = a_4 + a_5 x + a_6 y \end{cases} \quad (2)$$

where (x, y) is the position of the pixel in the current frame and (dx, dy) is the MV pointing from current position to position of the pixel in the previous image. The parameter vector to estimate is $\theta = (a_1, a_2, a_3, a_4, a_5, a_6)^T$.

In [18], we proposed to use the Tukey robust least-square estimator. The robustness of the method is based on an efficient outlier rejection three phases scheme. We adapt this scheme to the particular case of motion estimation in the wavelet domain. The first phase consists of an explicit rejection of the first row and column on the border of the image. The second phase is the rejection of the blocks having a low HF activity. For a block in the current image, the standard deviation vector $\sigma^{\mathbf{k}} = (\sigma^k_{LH}, \sigma^k_{HL}, \sigma^k_{HH})^T$ of the coefficients of the block in each subband is computed and compared to a given threshold. If criterion (3) is verified, the block corresponds to a flat region. Hence, BM has a strong probability of failing to find the real MV. This is why we do not consider these blocks as reliable for estimating the GM Model (GMM).

$$\sigma^{\mathbf{k}} < \mathbf{T}^{\mathbf{k}} <=> \begin{cases} \sigma^k_{LH} < T^k_{LH} \\ \sigma^k_{HL} < T^k_{HL} \\ \sigma^k_{HH} < T^k_{HH} \end{cases} \quad (3)$$

The threshold $\mathbf{T}^{\mathbf{k}}$ is adaptative with the level k of the pyramid to take into account the influence of the noise: at high resolution, main part of the noise is found in the HF subbands, whereas at lower resolution, signal has been low-passed filtered and noise reduced.

The third phase is related to the use of a robust Tukey bi-weight estimator. This allows assigning weights to block vectors expressing their relevance to the estimated model.

Outlier Caracteristic Function Determination The next step consists in determining for each MV $v \in V^k_t$ the value of the Outlier Caracteristic Function $f_o(v)$. $f_o(v) = 0$ indicates that v is a "non outlier" and $f_o(v) = 1$ means that v is an "outlier". For MVs resulting from the 3rd phase of the GME rejection scheme, relevance to the estimated model is given by an associated weight $w(v)$; the thresholding of these weights determines the status of the vector v (4).

$$f_o(v) = \begin{cases} 0 & \text{if } w(v) > T_w \\ 1 & \text{else} \end{cases} \quad (4)$$

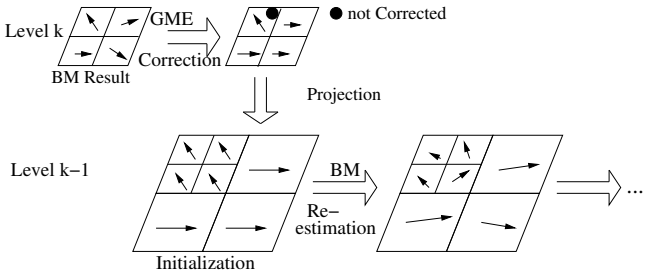

Figure 2: Motion Vectors projection using outlier information

For MVs that have been rejected during phases 1 or 2, no relevance to the GMM has been estimated. Nevertheless, we cannot consider that these blocks are part of a moving object. Hence we make a second pass on these blocks, in order to determine if they follow GM or not. We cannot use the MVs resulting from BM (because, as said previously, they are not reliable). Hence we use directly the MAD criterion of the BM (1). The MAD is computed for the MV estimated by the BM ($MAD(dx, dy)$) and for the MV computed thanks to the GMM found ($MAD(dx^m, dy^m)$). If the absolute difference between these two MAD values is less than a given threshold T_{MAD}, then the corresponding MV is marked as "non outlier" (5). Indeed, in this case, there is a strong probability that the BM has failed to estimate the real MV as the corresponding region is of low HF activity.

$$\begin{aligned} C(v) &= |MAD(dx, dy) - MAD(dx^m, dy^m)| \\ f_o(v) &= \begin{cases} 0 & \text{if } C(v) < T_{MAD} \\ 1 & \text{else} \end{cases} \end{aligned} \quad (5)$$

Thus constructed, "outlier" MVs correspond to blocks having their own motion whereas "non outlier" MVs correspond to blocks containing background samples and are well described by the GMM.

MV projection After all motion vectors have been estimated and marked at the top of the pyramid, it is necessary to obtain motion information at higher resolution levels. Thus we project and refine MVs. Here, two cases are distinguished following the MV is "outlier" or "non outlier" (Fig. 2). In case 1 ("outlier"), the MV estimated at level k is projected and $2p$ blocks at level $k-1$ are considered to correspond to the block of level k, here p is the subsampling factor (Note that we use a diadic subsampling as in JPEG2000, so $p = 2$). In case 2 ("non outlier"), the only one block at level $k-1$ corresponds to the block at level k. Its size is $2p$ the size of that one of level k and estimated vector is approximated by the value found using the model. These two

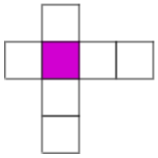

Figure 3: Dissymetric structural element used in uncertainty area determination

cases of projection are depicted in Fig. 2.
MVs corresponding to case 1 have to be accurately re-estimated. Indeed, in this category, one finds the blocks containing the edges of the moving objects.
In case 2, approximating by the model helps regularizing the MVs of the background. It also compensates the artefacts introduced by the shift-variance of the wavelet transform. Due to this ME process at a current pyramid level a new motion mask M_t^k is obtained composed of "outliers" blocks.

4.2. Scalable Morphological Color Segmentation

To obtain precise object shape, we propose improving Motion Masks by merging with a fine color-based segmentation. First a morphological color segmentation is applied to LL^K subband. This segmentation is superimposed on previously computed motion masks by majority vote. This result is then projected using data contained in the wavelet pyramid.

Rough Projection and uncertainty area. First, the object mask O_t^k and the segmentation mask S_t^k is roughly projected on the immediate higher resolution level $(k-1)$ using the location principle of wavelets. This leads to block effects and therefore results in a wrong assignment of pixels on the borders of the objects. Second, an uncertainty region is defined in which pixels will be assigned according to a refined criterion at the current level of the pyramid. We define the uncertainty area as the difference between dilation and erosion of the rough projected object shape by a 4-connected dissymetric structuring element (Fig. 3). This dissymetry allows to compensate the one due to rough projection.

Refinement using Markov random field style modeling. The main idea here is to take advantage of the specific information contained in the High Frequency subbands, i.e. the information on horizontal (LH), vertical (HL) and diagonal (HH) contours. We use this information in a Markov random field model. Using MAP criterion and assuming that color distribution inside each segmented region follows a Gaussian law, we come to the classical form of the problem [19] consisting of minimizing the sum of potential functions. We explicit here the potentials. U_1 is linked to the color values.

$$U_1 = (l-\mu)^t \Sigma^{-1} (l-\mu) \tag{6}$$

with Σ the covariance matrix of color vectors of LL subband and μ is the mean color vector. Hence, U_1 is the direct translation of the Gaussian law assumption. U_2 is the clique potential:

$$U_2 = \sum_{c \in C} (1 - \delta(x, x_c) + (2\delta(x, x_c) - 1)|HF|_n^c)\alpha \tag{7}$$

Here, C designates the set of all cliques of size 2 (we work in 8-connexity) and c designates one clique, x is the label to found, x_c is the label linked to the clique.$|HF|_n^c$ designates the normalized value of the high frequency coefficient associated with clique c, i.e. for an horizontal (respectively vertical, diagonal) clique we will use HL (respectively LH,HH) coefficient. α is a coefficient that must be fixed experimentally. The term $(1 - \delta(x, x_c))\alpha$ will privilege labels x conducting to homogeneous regions in terms of labels. The term $(2\delta(x, x_c) - 1)\alpha$ penalizes the configuration of two identical labels in a clique if the HF coefficient indicates that there is a contour and de-penalize the configuration of two different labels in a clique if the HF coefficient indicates that there is no contour. In this work we do not use a classical stochastic optimization method (e.g. ICM) but we consider the minimization of the energy as a deterministic region-growing method. We can justify this as the uncertainty area to assign is very narrow and thus the stochastic optimization would be costlier for a insufficient difference in the result.

5. Scalable Object-Based Descriptor

After the previous object-extraction step, a video is now represented as a set of objects, ie, a multiscale description of objects of interest is associated to each frame of the video (see Fig 4). Even if the amount of information available has been reduced, objects keep being meaningfull and can lead to an interpretation of high semantic level.

5.1. Descriptor Computation

The proposed descriptor is based on the normalized histograms of the objects' wavelet coefficients. For a frame in a video, the scalable descriptor of Object O_i, $\mathscr{H}_i$, is given by (8) (in the following, subscript t is omitted).

$$\mathscr{H}_i = \{(h_{LL,i}^k, h_{HF,i}^k), k \in [0..K]\} \tag{8}$$

For each object, two histograms are computed at each resolution level of the wavelet pyramid. $h_{LL,i}^k =$

Figure 4: Multiscale representation of an object

$h^k_{LL,i}(O^k_i, l_{LL})$ is the normalized YUV-joint histogram of the LL subband, ie, $h^k_{LL,i}(O^k_i, l_{LL})$ is the probability of appearance of the color $l_{LL} = (Y_{LL}, U_{LL}, V_{LL})^T$ in the object O^k_i. $h^k_{HF,i} = h^k_{HF,i}(O^k_i, l_{HF})$ is the YUV-joint histogram of the mean HF subband with l_{HF} defined as:

$$l = \begin{pmatrix} \frac{1}{3}(|LHY| + |HLY| + |HHY|) \\ \frac{1}{3}(|LHU| + |HLU| + |HHU|) \\ \frac{1}{3}(|LHV| + |HLV| + |HHV|) \end{pmatrix}$$

On the one hand, we choose to make a distinction between LF and HF subbands as they are of different physical meaning. On the other hand, only one histogram is used for the HF subbands in order to be more robust to object rotations.

5.1.1. Quantization

In order to make rapid comparisons, histograms are uniformly quantized. In histogram quantization, crucial point is the choice of class width as it may result in a over-smoothed, correct, or under-smoothed representation of the corresponding distribution. Moreover, a compromise has to be found. On the one hand, the histogram quantization has to be sufficiently general to apply to an object whatever the object is. On the other hand, it has to be adapted to the object in order to give it a relevant description.

In statistics, a well-known rule of Sturges [20] relates the number of bins C to the sample size N :

$$C = 1 + log_2(N) \tag{9}$$

In this work we consider all vectors of wavelet coefficient picked-up in the object area at each level as a single statistical sample. Thus the number of classes of each level of pyramid will be

$$C^k = 1 + log_2(card(O^k_i)) \tag{10}$$

Knowing the data range Δ, bin width b is proportionnally related to C: $b = \frac{\Delta}{C}$. Theoretical analysis [21] shows that the resulting bin width provides an over-smoothed histogram which better corresponds to unknown Probability Density Function (pdf).

As stated in section 5.1, the considered descriptor is composed of YUV-joint histograms. Thus defined, C is the number of bins in the 3D space YUV. The next step consists in computing the marginal bin width for Y, U and V components. If C_Y (respectively C_U, C_v) is the marginal number of bins for component Y (resp. U, V), then C is expressed as follows :

$$C = C_Y * C_U * C_V \tag{11}$$

Based on the fact that human vision is less sensitive to small variations of U and V components than of Y component, we decide to take $C_U = C_V = \frac{1}{2}C_Y$ which leads to $C_Y = (4C)^{\frac{1}{3}}$. Given the data range on each component, marginal class width, b_Y, b_U and b_V, are deduced. Obviously, such a choice is not completely adapted to the inherent pdf of wavelet coefficients of an object. Nevertheless it seems to be sufficiently well adapted to allow reliable comparisons.

5.2. Similarity Metrics

In this paper, two metrics have been used: the histogram intersection and the Bhattacharrya coefficient [22]. Histogram intersection has been introduced by Swain [23] and largely used in computer vision. Given two histograms, the histogram intersection for a subband $SB = LL, HF$ is:

$$d_{SB}(f_i, f_j) = \Sigma_l min(h_{SB}(f_i, l), h_{SB}(f_i, l)) \tag{12}$$

The Bhattacharrya coefficient is defined as

$$\rho_{SB}(f_i, f_j) = \Sigma_l \sqrt{h_{SB}(f_i, l) * h_{SB}(f_j, l)} \tag{13}$$

As there are two histograms associated to one level, we define the similarity metrics as the combination of the similary metrics of each histogram. That is, for histogram intersection:

$$d^k = \alpha_1 d^k_{LL} + (1 - \alpha_1) d^k_{HF} \text{ with } \alpha_1 \in [0, 1] \tag{14}$$

and for Battacharrya coefficient:

$$\rho^k = \alpha_2 \rho^k_{LL} + (1 - \alpha_2) \rho^k_{HF} \text{ with } \alpha_2 \in [0, 1] \tag{15}$$

Figure 5: Low Resolution Motion Mask Extraction

Figure 6: Superposition of Motion Mask and Color Segmentation

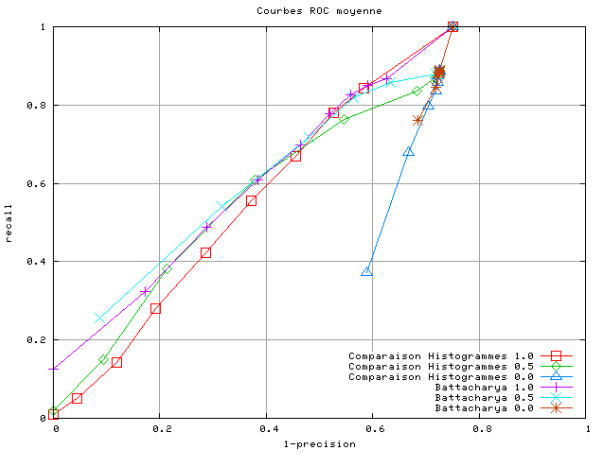

Figure 8: ROC Curves

6. Results

6.1. Object Extraction

In this section, results of Object Extraction are presented. Fig. 5 shows the Motion Mask $W^k_{t,i}$ obtained at the lower resolution level. Fig. 6 is $O^k_{t,i}$ the result of the fusion of Motion Mask and color Segmentation $S^k_{t,i}$. Then Fig. 7 shows the result of projection of this extraction at a higher resolution level. Even if not all the objects have been recovered, a significant part of them is available and no noisy information on backround has been added.

6.2. Object-Based Video Clip Retrieval

The test corpus we use for Video retrieval is constituted of 25 clips. Each clip has six copies made by geometrical transformations: rotation of 10 and 190 degrees, horizontal flip, resizing by factor 2 and 4, and croping. For a given original clip, aim is to retrieve all the transformed versions and uniquely those versions. It is a reduced problem of copy detection.

Figure 7: Projection of the result at Higher Resolution

For the original clip, the selection of the best segmented objects in one frame is manually determined. To compare to another clip, comparison of the corresponding histogram to all histograms of the other clip is fulfilled. Distance to this other clip is defined as the best score found.

The ROC curves are given in Fig. 6.2. Our descriptor shows promising results. For both measure, the adjonction of the measure on histograms HF ($\alpha_{1,2} = 0.5$) improves the quality of recall and precision in comparison to the use of LF histograms only.

7. Conclusion

This paper presents a complete method for extraction and design of a scalable object-based descriptor in the JPEG2000 wavelet domain. First, the extraction of object is achieved using the direct information obtained by wavelet transformation. A difficulty is that the approach is made from low resolution to higher resolutions, adapting to the scalable codestream. So, only a rough extraction can be expected. Second, a descriptor is build on object wavelet histogram thus extracted.

Even if the overcomplete scheme is done under the rough indexing paradigm, quality of the results shows to be sufficiently refined to allow a quite precise description. The adjonction of a HF wavelet histogram in comparison to the LF histogram alone conducts to an improvement of recall/precision results. All these results are really promising.

The next step of our research will be to work on a larger

Data Base and to test the descriptor with different types of queries.

8. Acknoledgements

This work is supported by French National grant ANR ACI MD “ICOS-HD”

References

[1] D. C. Initiative, url: http://www.dcimovies.com/.
[2] I. 15444-3:2002, Information technology. jpeg2000 image coding system: Motion jpeg2000.
[3] G. Pearson, M. Gill, An evaluation of motion jpeg2000 for video archiving, Proc. Archiving 2005 29 (2005) 237–243.
[4] Y. Wang, Z. Liu, J.-C. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Processing Magazine.
[5] Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. ali, M. Shah, Video understanding and content-based retrieval, TREC Video Retrieval Evaluation Online Proceedings, TRECVID05.
[6] P. Salembier, F. Marques, Region-based representations of image and video: segmentation tools for multimedia services, IEEE Trans. Circuits ans Systems for Video Technologies 9 (1999) 1147–1169.
[7] T. Meier, S. Ngan, Video segmentation for content-based coding, IEEE Trans. On Circuits and Systems for Video Technologies 9 (1999) 1190–1203.
[8] D. Zong, S. Chang, An integrated approach for content-based video object segmentation and retrieval, IEEE Trans. On Circuits and Systems for Video Technologies 9 (1999) 1259–1268.
[9] M. Kim, J. Choi, D. Kim, H. Lee, M. Lee, C. Ahn, Y.-S. Ho, A vop generation tool: automatic segmentation of moving objects in image sequences based on spatio-temporal information, IEEE Trans. on Circuits and Systems for Video Technologies 9 (1999) 1216–1226.
[10] F. Manerba, J. Benois-Pineau, R. Leonardi, Extraction of foreground objects from mpeg2 video stream in rough indexing framework, in: SPIE (Ed.), Storage and Retrieval Methods and Applications for Multimedia, 2004.
[11] C. Demonceaux, D. Kachi-Akkouche, Motion detection using wavelet analysis and hierarchical markov model, Lecture Notes in computer science 3667 (2006) 64–75.
[12] Y. Liu, K. Ngi Ngan, Fast multiresolution motion estimation algorithms for wavelet-based scalable video coding, Signal Processing: Image Communication 22 (2007) 448–465.
[13] D. Maestroni, A. Sarti, M. Tagliasacchi, S. Tubaro, Fast in-band motion estimation with variable size block matching, in: ICIP, 2004.
[14] M. Mandal, T. Aboulnasr, S. Panchanathan, Image indexing using moments and wavelets, IEEE Trans. Consumer Electronics 42.
[15] M. Mandal, T. Aboulnasr, S. Panchanathan, Fast wavelet histogram techniques for image indexing, Computer Vision and Understanding 75 (1999) 99–110.
[16] J. Smith, S. Chang, Transform features for texture classification and discrimination in large image databases, Proc. IEEE Int. Conf. Image Processing 3 (1994) 407–411.
[17] C. Morand, J. Benois-Pineau, J.-P. Domenger, B. Mansencal, Object-based indexing of compressed video content: from sd to hd video, in: 14th International Conference on Image Analysis and Processing, 2007.
[18] M. Durik, J. Benois-Pineau, Robust global motion characterisation for video indexing based on mepg2 optical flow, in: CBMI2001, Brescia, Italy, 2001, pp. 57–64.
[19] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions and the bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.
[20] H. Sturges, The choice of a class interval, J. AM. Stat. Assoc. 21 (1926) 65–66.
[21] D. Scott, Multivariate Density Estimation, John Wiley, New York, 1992.
[22] A. Battacharyya, On a measure of divergence between two statistical populations defined by probability distributions, Bull. Calcutta Math. Soc. 35 (1943) 99–109.
[23] M. Swain, D. Ballard, Color indexing, International Journal of Computer Vision 7 (1991) 11–32.