REAL-TIME ROUGH EXTRACTION OF FOREGROUND OBJECTS IN MPEG1,2 COMPRESSED VIDEO

Francesca Manerba[a], Jenny Benois-Pineau[b], Riccardo Leonardi[a]

[a]Dept. of Electronic for Automation, University of Brescia via Branze 38, I-25123 Brescia, Italy
[b]LaBRI, UMR CNRS/University Bordeaux 1 5300 351, Cours de la Libération, 33405 Talence,France
e-mail : francesca.manerba@unibs.it, jenny.benois@labri.fr, riccardo.leonardi@unibs.it.

ABSTRACT

This paper describes a new approach to extract foreground objects in MPEG1,2 video streams, in the framework of "rough indexing paradigm", that is starting from rough data obtained by only partially decoding the compressed stream. In this approach we use both P-frame motion information and I-frame colour information to identify and extract foreground objects. The particularity of our approach with regards to the state of the art methods consists in a robust estimation of camera motion and its use for localisation of real objects and filtering of parasite zones. Secondly, a spatio-temporal filtering of roughly segmented objects at DC resolution is fulfilled using motion trajectory and gaussian-like shape characteristic function. This paradigm results in content description in real time, maintaining a good level of details.

1. INTRODUCTION

The creation of large databases of audio-visual content in professional world and the growing use of intelligent home multimedia devices require the development of new methods for processing and indexing multimedia content [1].

The new multimedia standards MPEG4 and MPEG7 propose detailed schemes for multimedia content indexing and description, comprising the descriptors for specific objects inside visual scenes. Nevertheless the production of this object-based information for indexing multimedia content is out of the scope of standards and it is left to the content provider. So the problem of an efficient object extraction from raw or compressed video still remains a challenge. Several approaches have been proposed in the literature, which can broadly be classified either as intraframe segmentation, using the traditional image segmentation techniques [2], or as motion segmentation, where pixels with homogeneous motion field are grouped together for segmentation. Since both approaches have their own drawbacks, most of the video object segmentation tools integrate both spatial and temporal segmentation techniques [3].

However, most of combined approaches concentrate on segmentation in the pixel domain and cannot be performed until the data are reproduced in an uncompressed domain. A few compressed domain methods have been also proposed for spatio-temporal segmentation [4], but these approaches, although significantly faster than pixel-domain algorithms, cannot be executed in real-time. For this reason some methods have been developed to extract foreground objects in real time [5] working on compressed domain, but the problem is far from being resolved.

In this paper we propose a real-time method for extraction of foreground objects from MPEG1,2 "rough data". The adopted combined motion and colour-based techniques give proof of being an effective solution when working in the framework of the "rough indexing" paradigm we introduce in Section 2. The first step is to extract from P-frames, using a robust camera motion estimation algorithm, the regions, called "motion masks", which do not follow the motion of the camera. The camera model estimated is then used to filter the presence of parasite errors of MPEG1,2 encoding system. Then a morphological colour segmentation algorithm is performed on I-frames to refine the result of mask segmentation. Once foreground objects are obtained their shape is smoothed using a gaussian-like characteristic function. The trajectory of object is then computed. Thus the method allows for a simultaneous extraction of objects and description of a scene in terms of "parametric motion" of camera, trajectory of objects and their shape.

The paper is organized as follows: in Section 2 general principles of the method and the "rough indexing" paradigm are presented. In Section 3 we will explain how, from motion information related to P-frames, rough object masks can be extracted and then, in Section 4 how these results are combined with rough low-resolution colour segmentation of I–frames to refine object shape and capture meaningful objects at I-frame temporal resolution. In Section 5 object shapes and trajectories are calculated. Results are finally presented in Section 6.

2. ROUGH INDEXING FRAMEWORK AND PRINCIPLES OF OBJECT EXTRACTON

Recently, a new trend in analysis methods for indexing multimedia content has appeared which can be qualified as a "rough indexing" paradigm. This means a fast and approximate analysis of multimedia content at a poor resolution. Our "rough indexing" paradigm can be expressed as "the most complete model" on rough data - that is motion vectors and DC images - and at low resolution (both spatial and temporal). In this paradigm we combine both motion information – the complete 1st order camera motion descriptor of MPEG7 standard - and region-based colour segmentation to extract meaningful objects from compressed video with arbitrary camera and object motion.

Figure 1 displays the global scheme of the approach. Considering an MPEG1,2 stream referring to given GOP limited by intra-coded I-frames we utilize macro-block motion vectors in P-frames to estimate camera motion and extract motion masks. From the I-frame, instead, we extract all color information, that is we apply a color segmentation algorithm to the DCT coefficients of the I-frame. Once obtained, color and motion information are projected onto the I-frame location to extract the foreground objects.

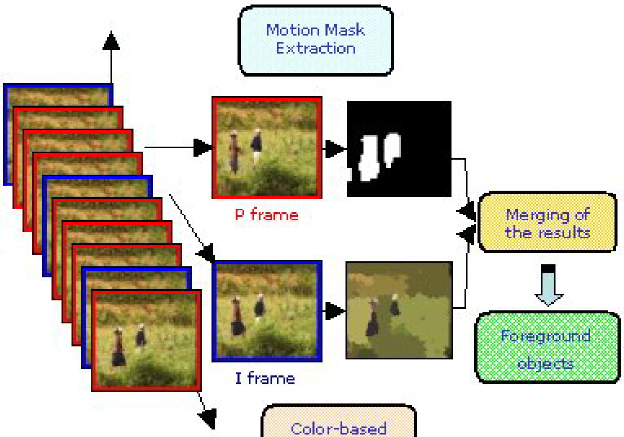

Figure 1: global scheme of the system

3. OBJECT MASK EXTRACTION

In order to detect "foreground blocks" which do not follow the global camera motion, we have to estimate it. Here we consider a parametric affine motion model with 6 parameters as admissible in MPEG7 "parametric motion" descriptor:

$$\begin{aligned} dx_i &= a_1 + a_2 x_i + a_3 y_i \\ dy_i &= a_4 + a_5 x_i + a_6 y_i \end{aligned} \qquad (1)$$

Here (x_i, y_i) is the position of the i-th macro-block centre in the current image and (dx_i, dy_i) is the motion vector pointing from the current position to the macro-block centre in the previous reference frame, only P-frames being used for motion model estimation. The estimated parameter vector $\theta=(a_1, a_2, a_3, a_4, a_5, a_6)^T$ can be related to the different camera movements (pan, tilt, zoom, rotation). To estimate the camera motion parameters from MPEG1,2 macro-block optical flow we use a robust weighted least-square estimator [6], based on an outlier rejection scheme and on the use of Tuckey bi-weight estimator as cost function.

So it is possibly to correctly estimate the global motion even in the presence of a large number of outliers. Therefore, unlike[5], practically no constraints have to be made on the size of object which can be close to the camera.

The estimation process [6] also assigns the weights w_i to the initial measures, i.e. motion vectors, expressing their relevance to the estimated model.

The weight of each macro-block motion vector is calculated separately in two directions and denoted as (w_{dx}, w_{dy}).

3.1. Motion mask extraction from single P-frame

Once the estimation of camera motion model is fulfilled, the problem of object extraction can be formulated as a classification problem trying to identify the macro-blocks whose motion is irrelevant with respect to the estimated model.

Let us consider a grey-level sequence W(x,y,t) of spatial resolution $(N*M)/(\text{MacroBlockSize})^2$ and temporal resolution as I-P distance in MPEG defined as follows:

$$W(x,y,t) = \left\lfloor \left(1 - \max\left(w_{dx}(t), w_{dy}(t)\right)\right) \cdot 255 \right\rfloor \qquad (2)$$

Here the brightest pixels correspond to macro-blocks with low weights and thus could belong to the objects that do not follow the global motion model. Thus in order to get relevant pixels well representing objects with proper motion, a binary image sequence $W^b(x,y,t)$ will be now computed by thresholding of W(x,y,t). The result is a "binary" volume in x,y,t space. An exemple of this volume is shown in Figure 2. After obtaining the binary volume, we have to smooth the result filtering the outliers which are not representing foreground objects (the ones only due to camera motion and MPEG1,2 coding errors).

In fact, the section of the volume at time t_0 $W^b(x,y,t_0)$ only gives the positions of macro-blocks that do not follow the camera motions, but they can be caused not only by a foreground object but also by the presence of flat zones (see Sec. 4) or by new zones entered in the frame due to camera movements.

In this last case, in each frame new macro-blocks enter in the frame in the direction opposite to the camera movement. Their motion vectors are erroneous and represent outliers with regards to the camera motion (see Figure 2). It is necessary to filter them. These outliers cannot be filtered with objects tracking as proposed in [5] because they are present in more subsequent images. We propose to filter them using camera motion information. In [7] we proposed a method for warping the frames in the direction of camera motion in order to remove entering blocs. This warping uses estimated camera model to overlap current and previous frame in order to exclude entering blocks as illustrated in Figure 3. In all I and P-frames, the outliers due to camera motion on the frame borders were correctly removed even when camera motion was not a pure translation.

Then a 3D segmentation algorithm is applied to such volume $W^b(x,y,t)$ to filter MPEG1,2 noise and finally obtain the real object masks as illustrated in [7]. The result of this segmentation is a 3D volumetric mask that highlights the region inside which a foreground object is probably located and moves. In this work we used a 3D morphological segmentation algorithm developed in [8].

Since motion masks have been obtained only for P-frames, we have to build the corresponding mask for the I-frame. As the MPEG1,2 decoder does not give motion vectors for the I-frame we cannot build the mask starting from MPEG1,2 motion information. Therefore, we propose to interpolate the mask for I-frame from masks of previous and following P-frame. We map, in this way, the mask onto the I-frame that exhibits the approximate position of the objects.

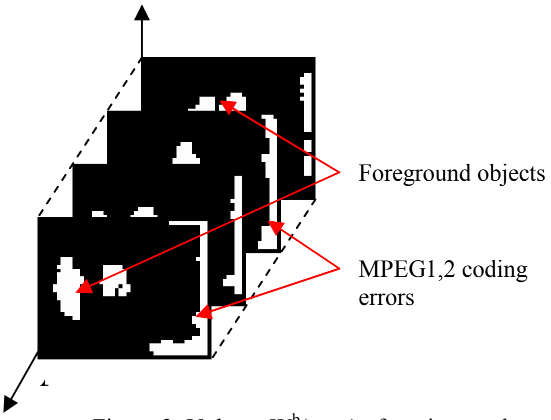

Figure 2: Volume $W^b(x,y,t)$ of motion mask.

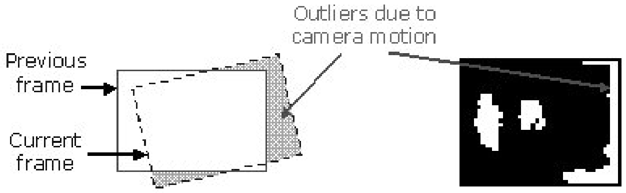

Figure 3: High weights values caused by outliers on the frame boundaries

4. OBJECT MASK REFINEMENT BY COLOR SEGMENTATION

Hence motion masks interpolated for I-frames indicate the locus of objects with their proper motion. Now, if we use the colour information of I-frames inside the masks, we can refine the object shapes and estimate their textural and colour parameters, thus indexing the video content by foreground object characterisation with an I-frame based temporal resolution. In the context of the rough-indexing framework, we will use only DC coefficients, thus only partial decoding the MPEG1,2 stream is needed.

The first task consists in a gradient calculation and thresholding to obtain the borders of the objects. A region growing algorithm is then performed with a modified watershed in YUV space [7].

The particularity of our morphological approach consists in the modified watershed and in the region growing algorithm developed using a region-adapted threshold [9]. The threshold is calculated as function of the average luminance of the region so that it depends on the mean grey level of the considered region. This function follows the principles of the “function of visual sensitivity” that shows how the difference between two grey level values is less perceived when the values are at the extremities of the range.

Once the I-frame segmentation has been performed, we finally obtain foreground objects extracted from I-frames and their spatial location in P-frames by superimposing and merging motion and colour masks at DC-frame resolution (Figure 4).

Even in this case it is necessary to filter the result; in fact if a flat region is present an object is erroneously detected; flat zones are characterized by very noisy motion vectors in many subsequent P-frames, due to ill-posed problem of motion estimation. The flatness of the zones, understood as a low energy of image gradient, allows for removing such areas [7].

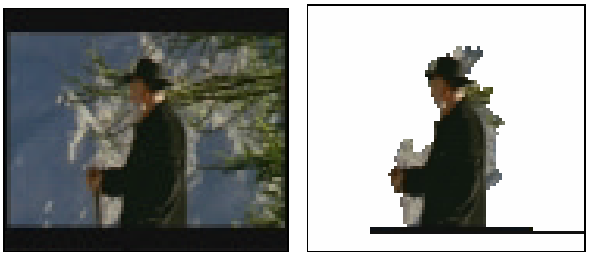

Figure 4: DC image and corresponding foreground object

5. EXTRACTION OF OBJECT SHAPE AND TRAJECTORY

Once foreground objects are obtained, the objective is to build the object trajectory in which each object of a I-frame is linked with the corresponding one in the previous reference frame and in the following reference frame.

In the case that more than one object is present in the scene, it is necessary to correctly link each object in the current frame with the corresponding one in the following frame. To do that, we suppose that an object is rigid and its motion can be well approximated by an affine model (1). Thus the object motion parameters are computed by a least square estimation to the motion vectors contained in the object masks. In this way the objects are projected, following their motion model, from an I-frame to the following one along the whole sequence and the masks corresponding to the same object taken in different moments of time are linked together. The trajectory can then be calculated linking object barycentres.

Nevertheless, due to the two factors influencing the objects extraction result in the framework of rough indexing paradigm, namely discontinuity of natural object motion along the time and color segmentation noise, the resulting trajectory can be jerky because of the shape variations. Thus the shape has to be filtered along the time. We propose to approximate the shape with an elliptic function in those frames where any object is detected. Hence our purpose now is to find the coefficient of a generic ellipse that can best approximate the contours of the objects in the frame we are considering.

We consider *t(x,y),* a normalized gaussian function centered in the object barycentre and with variance values set accordingly to object bounding box (3).

$$t(x,y)=\exp\left(-\frac{1}{2}\left(\left(\frac{(x-\mu_x)^2}{\sigma_x^2}\right)+\left(\frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right)\right) \qquad (3)$$

Fixing the value of *t*, the gaussian function gives a set of concentric ellipses with different extent. In fact this operation is equivalent to cut the gaussian function with a plane of equation

$t=t_0$. To choose the value of t that gives the best approximation of the shape for the selected object we minimize the function:

$$\min \sum_{(x,y)} \delta(x,y) \tag{4}$$

where

$$\delta(x,y) = \begin{cases} 1 \quad if \quad (x,y) \in ((\text{ellipse} \cup \text{mask}) - (\text{ellipse} \cap \text{mask})) \\ \qquad 0 \quad \text{otherwise} \end{cases}$$

An example is shown in Figure 5:

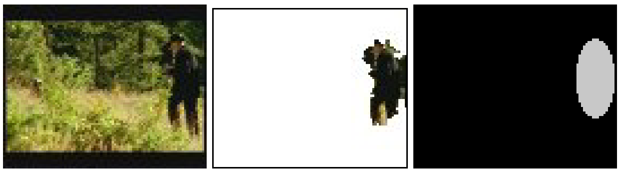

Figure 5: A DC image, the extracted object and the corresponding shape approximation.

6. RESULTS

In this paper we have presented a method for foreground objects extraction in the "rough indexing" paradigm, which allows extraction of foreground objects in MPEG1,2 compressed video at I-P temporal resolution. The method performs in real-time and gives promising results.

The whole processing time is composed of: decoding time for motion vectors and DC coefficients, robust motion estimation, 3D morphological segmentation, 2D colour segmentation, shape approximation and trajectory calculation. The mean processing time for a MPEG2 video sequence is given in Table 1. The total time of extracting objects is 3.5 sec at Pentium 4 PC, that is in real time (3.6 sec for a sequence of 90 frames at 25 frames/sec). Trajectory calculation is not included in Table 1 because to compute it, it is necessary to have extracted object yet and to have processed all the sequence. Generally speaking, processing time for all steps except shape filtering does not depend on number of objects.

The algorithm proposed, based on a mixed motion and colour based approach, has been tested on different sequences from a set of natural video content. This test set considers the feature documentaries "De l'Arbre à l'Ouvrage", "Hiragasy", "Aquaculture", "Cavitation", "Chancre" (SFRS ®). An example of the results obtained for movie "De l'Arbre à l'Ouvrage" has been already shown in Figure 4 and Figure 5. The method shows some imprecision on the object border and also a slight over-detection due to the presence of erroneous MPEG1,2 motion vectors in the case of strongly textured image. An under-detection is expressed as the merge of close objects and occurs when objects are situated at a relative distance of the order of macro-block size.

Some limit conditions have been taken into account during the algorithm evaluation process. For example tests have been conducted when objects are so near to the camera that they cover a large part of the background (30%) and in the case of no foreground objects. In particular in the first case the robustness of the motion detection algorithm has led to a correct extraction of the camera motion parameters and consequently to a correct detection of foreground objects. In the second case, even if the noise present in MPEG1,2 motion vectors has caused the presence of macro-blocks with high weight value in motion estimation, the filtering fulfilled by 3D segmentation and refinement by colour segmentation has permitted to classify the "outliers" as pure noise.

Sequence (num of GOP)	**Decoding + Motion Est.**	**Object Extraction**	**Tot.**
De l'arbre à l'ouvrage (6 GOP, 90 frames)	0.7 sec	2.8 sec	3.5 sec

Table 1: computational time for extracting objects in the MPEG2 video "De l'arbre à l'ouvrage" (SFRS®).

7. REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, A. Zakhor "Applications of video content analysis and retrieval", *IEEE Multimedia*, pp. 42 – 55, Jul – Sep 02

[2] P. Salembier, F.Marques, "Region-based representations of image and video: segmentation tools for multimedia services", IEEE Trans. Circuits and Systems for Video Technology, vol. 9, pp. 1147-1169, Dec. 1999.

[3] D. Zong, S.F.Chang, "An integrated approach for content-based video object segmentation and retrieval", IEEE Trans. on Circuits and Systems for Video Technology, vol. 9, pp. 1259-1268, Dec. 1999.

[4] R.V. Babu, K.R. Ramakrishnan, S.H. Srinivasan, "Video object segmentation: a compressed domain approach", IEEE Trans. on Circuits and Systems for Video Technologies, vol.14, n.4, pp. 462-474, April 2004.

[5] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, M.G. Strintzis, "Real-Time Compressed-Domain Spatiotemporal Video Segmentation and Ontologies for Video Indexing and Retrieval, IEEE Trans. Circuits and Systems for Video Technology, vol. 14, No. 5, May 2004.

[6] M. Durik, J. Benois-Pineau, "Robust Motion Characterisation for Video Indexing Based on MPEG2 Optical Flow", in *CBMI'01*, Brescia, Italy, September 2001.

[7] F. Manerba, J. Benois-Pineau, R. Leonardi, "Extraction of foreground objects from a MPEG2 video stream in "rough indexing" framework", in SPIE Electronic Imaging '04, San Jose, United States, January 2004.

[8] S. Benini, E. Boniotti, R. Leonardi and A. Signoroni, "Interactive Segmentation of Biomedical Images and Volumes using Connected Operators", *2000, ICIP2000*, Vancouver, Canada September 2000.

[9] A. Mahboubi, J. Benois-Pineau, D. Barba ''Suivi et indexation des objets dans des séquences vidéo avec la mise-à-jour par confirmation retrograde'', CORESA'2001, Dijon, France, November 12-2001.