Extraction of foreground objects from a MPEG2 video stream in "rough –indexing" framework

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 15300351, Cours de la Libération, 33405 Talence, France

ABSTRACT

In the domain of video indexing, one of the research topics is the automatic extraction of information to reach the objective of automatically describing and organizing the content. Thinking of a video stream, different kinds of information can be taken into account, but we can suppose that most of the information is contained in the foreground objects so that number of objects, their shape, their contours and so on, can constitute a good guess for the content description.
This paper describes a new approach to extract foreground objects in MPEG2 video stream, in the framework of "rough indexing paradigm" we define. This paradigm leads us to reach the purpose in near real time, nevertheless maintaining a good level of details.

Keywords: object-based indexing of multimedia content, compressed streams, motion analysis, foreground object segmentation.

1.INTRODUCTION

The creation of large databases of audio-visual content in professional world and extensively increasing use of home multimedia devices able to store hundreds of hours of multimedia content from broadcast have led to the development of new methods for processing and indexing multimedia documents. The objective here consists in extracting of meaningful information and organising multimedia content for an easy user navigation in the multimedia information pool.
A variety of methods[1] have recently been developed to fulfil this objective, mainly using the global features of multimedia content such as dominant colour of images or key-frames in video, global motion etc... The new multimedia description standard MPEG7[2] proposes much detailed schemes for multimedia content indexing and description comprising the descriptors for specific objects inside visual scenes. Thus it becomes interesting to index the multimedia documents by objects, searching for the similarity of their behaviour (trajectory), spatial features (shape and texture descriptors) and so on. Nevertheless, the problem of an efficient object extraction from raw or compressed video still remains a challenge. It is clear that the requirements for the precision and complexity of object extraction tools are very much application dependent.
In this paper we propose a motion- and colour–based method for extraction of foreground objects from MPEG2 compressed streams in the framework of "rough indexing" paradigm we introduce.
In Section 2 general principles of the method are presented. In section 3 we will explain how, from motion information related to P-frames, rough object mask can be extracted and how this result is combined with rough low-resolution colour segmentation of I–frames in order to refine object shape and capture meaningful objects at I-frame temporal resolution. Some results and perspectives are presented in Section 4.

2."ROUGH INDEXING" PARADIGM AND PRINCIPLES OF FOREGROUND OBJECT EXTRACTION FROM COMPRESSED VIDEO

Recently, a new trend in analysis methods for indexing multimedia content has appeared which can be qualified as "rough indexing" paradigm. Many authors[3] are interested in fast and approximate analysis of multimedia content at poor

* jenny.benois@labri.fr
* {francesca.manerba, riccardo.leonardi}@unibs.it

Storage and Retrieval Methods and Applications for Multimedia 2004, edited by Minerva M. Yeung, Rainer W. Lienhart, Chung-Sheng Li, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5307 © 2004 SPIE and IS&T · 0277-786X/04/$15

(or intentionally degraded) resolution. Coded multimedia streams give a rich background for development of these methods, as low resolution data can be easily extracted from MPEG compressed streams without complete decoding. Thus many authors dealt with extraction of moving foreground objects from MPEG2 compressed video with still background[4], independently numerous works[5] have been devoted to the estimation of global camera model from compressed video. Thus the "rough data" – that is noisy motion vectors and DC images - have been used for fine indexing. Our "rough indexing" paradigm can be expressed as "the most complete model" on rough data and at rough 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 MPEG2 stream inside each GOP limited by intra-coded I-frames we utilize noisy macro-block motion vectors in P-frames to estimate camera motion and separate "foreground blocks" which do not follow it. These blocks serve for motion mask extraction from P-frames. From the I-frame, instead, we extract all colour information, that is we apply a colour segmentation algorithm to the DCT coefficients of the I-frame to subdivide the image in colour homogeneous regions.
Once obtained, colour and motion information are merged together at I-frame moment of time to extract the foreground objects.

3. OBJECT MASK EXTRACTION BASED ON MOTION INFORMATION

The idea here is to extract foreground blocks which do not follow global camera motion in each P-frame and then to apply a 3D segmentation on a whole GOP in order to extract a 3D "motion shape". Then the cross-section of this 3D motion shape at each P-frame will correspond to objects masks.

3.1 Motion mask extraction from single P-frame

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

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

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 macro-block centre in the previous image. The obtained estimator vector is $\theta=(a_1, a_2, a_3, a_4, a_5, a_6)^T$ and describes the different camera movements (pan, tilt, zoom, rotation).
To estimate the camera motion parameters from MPEG2 macro-block optical flow we used a robust weighted least-square estimator[6] taking the MPEG2 macro-block motion vectors as measures. The robustness of the method is based on a good outliers rejection scheme and on the use of Tuckey bi-weight estimator[9] as cost function $\rho(r)$ instead of classical cost function $\rho(r)=r^2$, where r is the residual between the measured values and those obtained by the model (1). The Tukey's bi-weight function $\rho(r,C)$ and its derivative $\psi(r,C)$, are defined as follows:

$$\rho(r,C)=\begin{cases} \dfrac{r^6}{6}+\dfrac{C^2 r^4}{4}+\dfrac{C^4 r^2}{2}, & \text{if } |r|<C \\ \dfrac{C^6}{6}, & \text{otherwise} \end{cases} \qquad \psi(r,C)=\begin{cases} r\left(r^2-C^2\right)^2, & \text{if } |r|<C \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where C is a constant value.
The estimation process[6] not only gives the optimal values of model parameters but also assigns the weights w_i to the initial measures expressing their relevance to the estimated model. The weight w_i is defined as follow:

$$w_i = \frac{\psi(r_i)}{r_i}, \qquad 0 \le w_i \tag{3}$$

that is when w_i is close to 1 means that the macro-block we are considering follows camera movement, while w_i close to 0 means that the macro-block has a different motion. The weight of each macro-block is calculated in the two directions x and y to obtain the relevance of the macro-block to each of both motion direction (w_{dx}, w_{dy}). Weights (w_{dx}, w_{dy}) will

then be normalized to fit the interval [0,1].

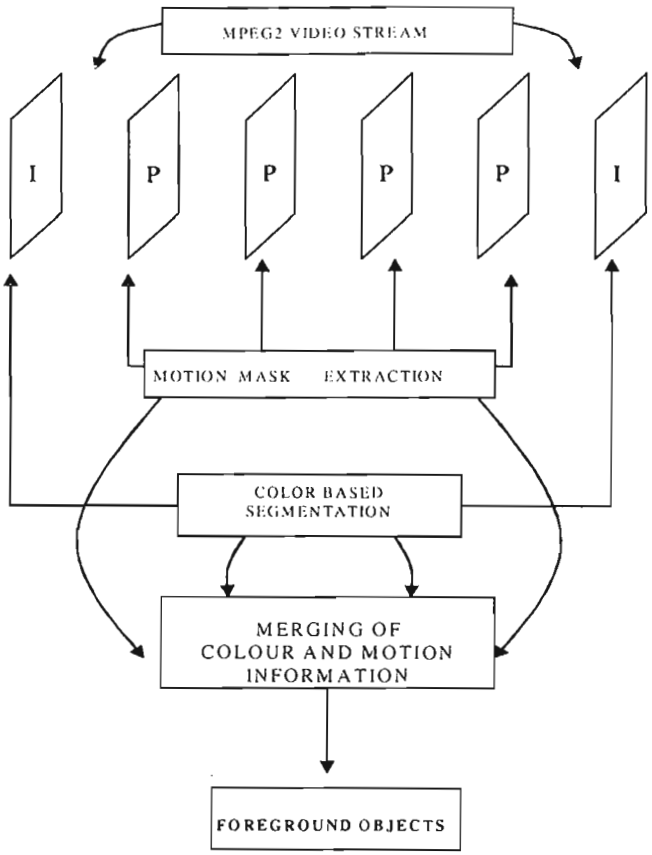


Fig. 1: Global scheme of the system

Once the estimation of camera motion model is fulfilled, the problem of object extraction can be formulated as separation of the macro-blocks with motion irrelevant to the estimated model. Hence objects with proper motion can be deleted.
The intra-coded macro-blocks can also be included in the set of "irrelevant" measures as they correspond to the failure of MPEG block-based motion estimator: so we can assign them a value of weight w_i close to 0 because we can suppose that the motion vectors of these macro-blocks do not follow the camera model.
Let us consider a grey-level image $I_{x,y}$ of resolution N/MacroBlockSize*M/MacroBlockSize defined as follows:

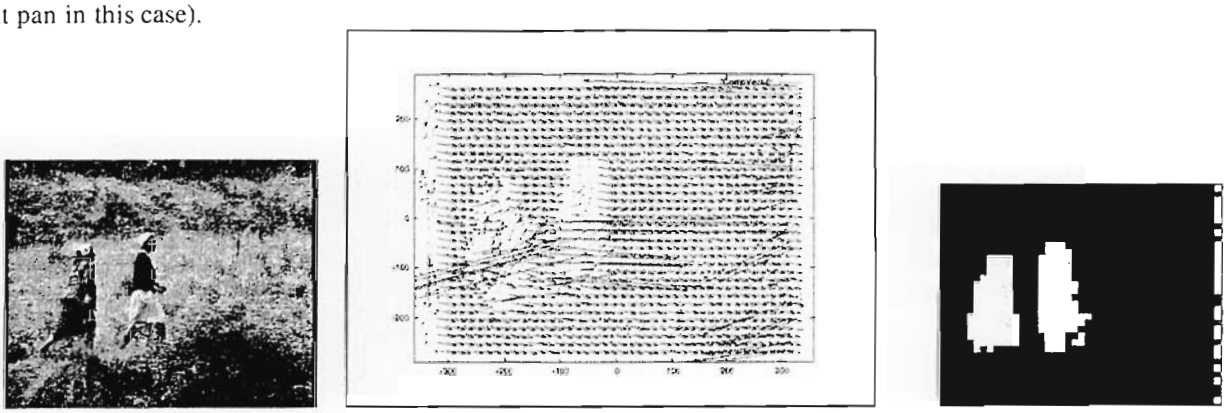
$$I_{x,y} = \lfloor (1 - \max(w_{dx}, w_{dy})) \cdot 255 \rfloor \tag{4}$$

Here the brighter pixels correspond to macro-blocks with low weights and thus would belong to the objects that do not follow the global movement. Pixels of $I_{x,y}$ with low values correspond to low weights scattered in P-frame due to textured and local motion deformation. Thus in order to get relevant pixels well representing objects with proper motion, a binary image $I^b_{x,y}$ will be now computed by thresholding:

$$I^b_{x,y} = \begin{cases} 1 \text{ if } I_{x,y} < s \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

The threshold s is based on typical "low value" of weights and was tested on various sequences in the range 5-10.

The result is a binary image as in fig. 2c). In Fig. 2 the correspondence between the motion vectors and the region where the objects are located is shown. In 2a) is represented a P-frame taken from a video sequence where the object of interest is constituted by two women walking tracked by the camera and in 2b) are presented motion vectors associated to the frame. It is possible to see in the middle of the motion vectors field two regions with motion vectors completely different from the others due to object presence and in the zone at the right bordersome other "outliers" due in this case to camera motion. In fig. 2c) the obtained binary image $I^b_{x,y}$ is shown. The two white regions in the middle correspond to the zones where the foreground objects are located; on the right border instead are presented the outliers due to camera motion (a right pan in this case).



a) b) c)

Fig. 2:a) original frame. b) motion vectors of a P-frame. c) the corresponding grey level image

In fact, in each frame new macro-blocks enter in the frame in the direction opposite to the camera movement. The pixels of original video frame in this macro-blocks do not have their antecedent in reference frame. Therefore, motion vectors resulting from MPEG2 decoding process are erroneous and do not follow camera motion in most cases. In this case we have high irrelevance weight on these zones even if any object is present (see Fig. 3 and Fig. 2c)).

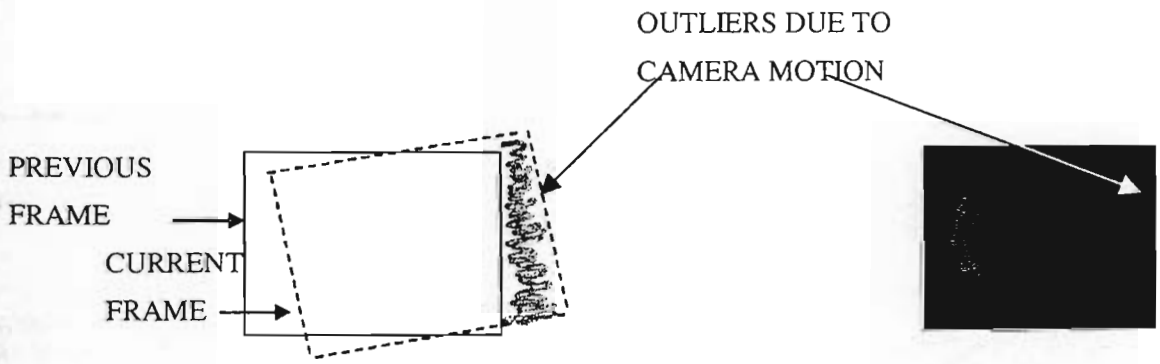
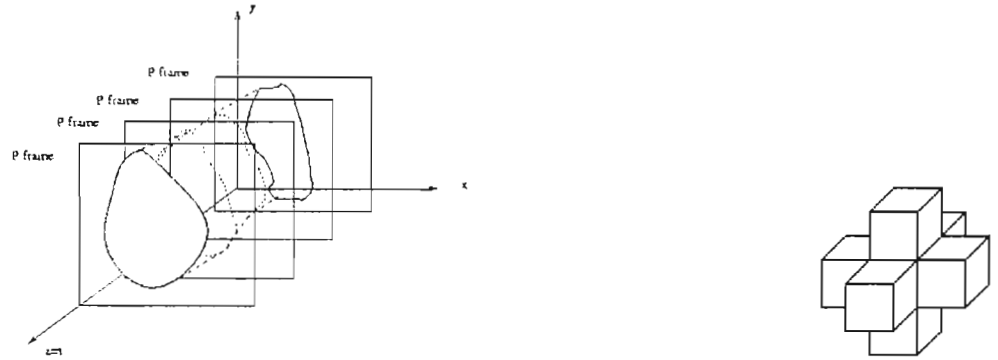


Fig. 3: High weights values caused by outliers on the frame boundaries

To filter the outliers on the frame border we use the estimated camera motion model. In fact with foreword prediction motion coding, the displacement vector $\bar{d} = (dx, dy)^T$ of a macro-block in the current frame is related to the coordinates of pixel $(x_c, y_c)^T$ in a current frame and its reference pixel $(x_p, y_p)^T$ in the reference frame as:

$$dx = x_p - x_c$$
$$dy = y_p - y_c \qquad (6)$$

Now using the model (1) we can solve it for x_c and y_c taking as reference pixels the corners of the reference frame. Thus the reference frame will be warped to the current frame.
Thus we can obtain the geometry of the zone entered in the frame: if some "outliers" are present in that zone we can suppose that they are caused by camera motion and we do not consider them when searching for object masks.
Repeating the method described above for all P-frames inside a single video shot we can obtain motion masks for foreground objects in the shot. This method requires preliminary segmentation of video content into shots, as at the border of shots, MPEG motion vectors are massively erroneous and the estimated camera motion does not correspond to the reality. Hence we obtain the first guess of objects at reduced temporal resolution accordingly to our rough indexing paradigm. Nevertheless mask in each pair of P-frames were obtained independently from each other. This is why they remain noisy in time. To improve the detection we will profit of temporal coherence of video and smooth the detection along the time. To do this we can model a video segmentation as a 3D volume in (x,y,t) space. Here, the characteristic function of objects $f(x,y,t)$ is known at time moment corresponding to P -frames.
Let us consider two consecutive GOPs in MPEG2 stream inside a shot. To smooth $f(x,y,t)$ along the time we will apply a 3D segmentation algorithm to such pairs of GOPs (see fig. 4a).
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[7]. The algorithm follows a usual morphological schema: filtering, gradient computation and region growing by watershed in a 3D space. In morphological operations, a 3D 6-connected structuring element was used (see Fig. 4b)).



a) b)

Fig. 4:a) 3D segmentation along the volume composed by P frames. b) the 6-connected element used for morphological segmentation

The 3D segmentation allows for smoothing of initial noisy characteristic function of objects. Considering the volume slices $\tilde{I}^{P}_{x,y}$, we obtain a 2D mask for each P-frame we have analysed.

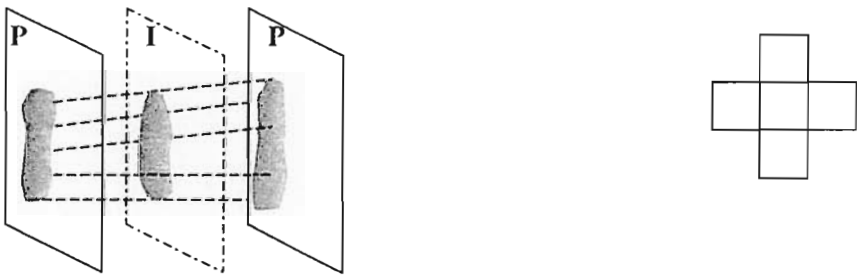
3.2. Extraction of motion mask in I -frames

The extracted motion masks represent a good guess to the object shape in P-frames. Nevertheless, only motion information is not sufficient for a robust object extraction. Thus we propose to merge the motion masks with the results of colour-based intra-frame segmentation of I-frames. In the framework of rough indexing paradigm we will only use DC-images of original I-frames.
Since motion masks have been obtained only for P-frames, we have to build the corresponding mask for the I-frame in order to overlap it to colour-based segmentation of the frame. As the MPEG2 decoder does not give motion vectors for the I frame we cannot build the mask starting from MPEG2 motion information.
Looking to the structure of the MPEG2 compression standard it can be seen that, considering two consecutive groups of pictures (GOP), the I-frame is situated between two B-frames, or, if we consider only P-frames, as in the case of this work, between two P-frames. Therefore, in order to calculate the mask for the I-frame, we can consider the P-frame that comes before the I-frame and the following one and then to interpolate the two images (Fig. 5a)). The interpolation can be fulfilled by two approaches: i) motion based one [11]. Here the region masks are projected into frame to be interpolated . ii) spatio-temporal segmentation without use of motion. For the sake of low computational cost we decided to use a spatial interpolation. We interpolate masks by morphological filtering. In fact the resulting binary mask in I-frame $\tilde{I}^{P}_{x,y}(t)$

will be computed as:

$$\tilde{I}^{b}_{x,y}(t) = \min\left(\delta\tilde{I}^{b}_{x,y}(t-1), \delta\tilde{I}^{b}_{x,y}(t+1)\right). \tag{7}$$

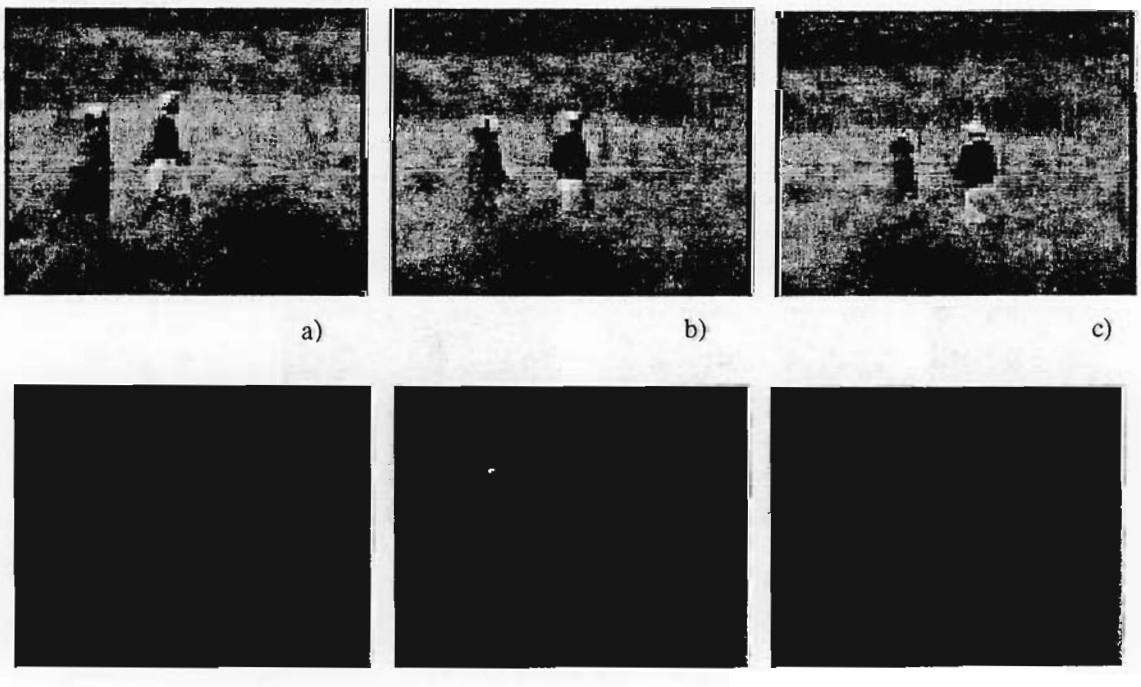
Here δ denotes the morphological dilation with 4-connected structural element of radius 1 (see fig. 5b)).
In this way we obtain the mask for the I-frame that exhibits the approximate position of the objects.



a) b)

Fig. 5: a) Creation of the Mask for the I-frame by interpolation of two P-frames. b) The 4-connected structuring element used in morphological 2D operation.

Figure 6 depicts some I-frames extracted from a MPEG2 video stream and the corresponding masks obtained using the motion-based approach proposed.



a) b) c)

d) e) f)

Fig. 6: Extraction of motion masks for I-frames. "Hiragasy"® SFRS
a), b), c) Original I-frames at DC resolution d), e), f) the corresponding masks.

3.3. Object –mask refinement by colour segmentation

Hence motion masks interpolated for I-frames indicate locus of objects with proper motion. Now, if we use the colour information of I-frames inside masks, we can refine object shapes and estimate their textural and colour parameters, thus indexing video content by spatial features at I-frame temporal resolution. Hence we implement a colour segmentation process for the I frame to subdivide it into regions. Then regions are selected overlapping with motion mask we have calculated before. The set of overlapped region forms the objects of interest. Remaining in the rough indexing paradigm framework, we will use only DC coefficient of I-frames shaped into "DC images"[8]. The DC coefficients are easy to extract from MPEG2 stream without complete decoding of it. When working with 4:2:0 and 4:2:2 chroma formats we use a 0-order interpolation for U and V colour components.

For the colour segmentation we remain in morphological framework and apply it to the still colour DC frame. First task consists in gradient calculation and threshold application to obtain the borders of the objects (fig. 7b)) and then a region growing algorithm is performed by a modified watershed in YUV space.

The particularity of our morphological approach consists in the modified watershed. When the threshold as been applied to morphological gradient, all the connected regions present in the image are filled and labelled and for each of them the average colour value is calculated. We obtain, in this way, a map of the colour regions with some uncertainty zones corresponding to the zones with high gradient, typically the borders of the objects. To assign these zones to the connected regions we have applied a region growing algorithm developed using region-adapted threshold[10] (see eq. 8).

The threshold is calculated as function of the average luminance of the region and of a parameter Δ^i that grows with the iteration *i*.

$$\text{threshold} = F(\overline{m})g(\Delta^i). \tag{8}$$

where $F(\overline{m}) = |\overline{m} - 127| + 128$ and $\Delta^i = \Delta^{i-1} + 0.01$.

The function $F(\overline{m})$ shows that the threshold 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.

The function $g(\Delta^i)$ is an incremental term used to progressively relax the thresholds for merging boundary pixels with adjacent region. The initial value Δ^0 is computed as 1/F(127) and it is used to compute the first threshold value; so for regions in the middle of luminance range the threshold corresponds to only one level of luminance difference. In this first step the macro-blocks that differ from the adjacent region of a value smaller than this threshold are included in the region. When no macro-blocks can be added to any regions the value of the threshold is recalculated using the incremented term Δ^i. The threshold is relaxed until all the uncertainty zones are assigned to a region (fig. 7c)).

Once I-frame segmentation has been done, 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 (see Fig 7d)).

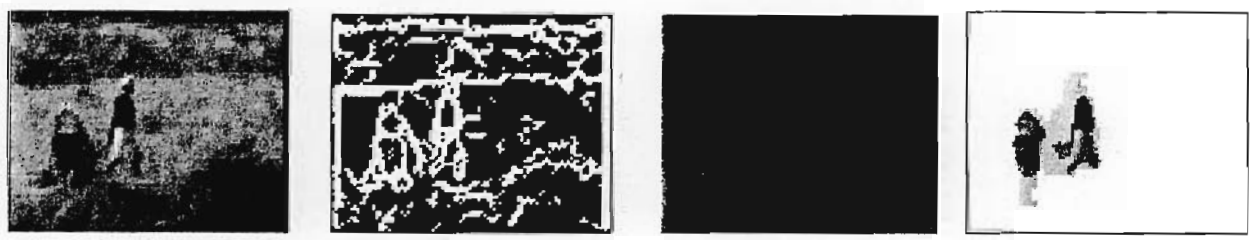


Fig 7: a) the original I-frame; b) the result of gradient calculation; c) the result of colour segmentation process visualized by its label map; d) superposition of motion mask and colour segmented regions.

Once we have extracted the objects, it is necessary to apply a gradient-based filtering to delete possible flat zones. A flat zone is region in the colour image characterized by very low colour gradient value, for example the sky is a typical flat

region. As the colour value is almost constant, the MPEG2 motion vectors cannot find the exact position of the current block in the previous image because there are a lot of macro-blocks of the same colour. So the macro-blocks of a flat region are characterized by motion vectors that do not follow camera movement and for this can be detected as moving objects.

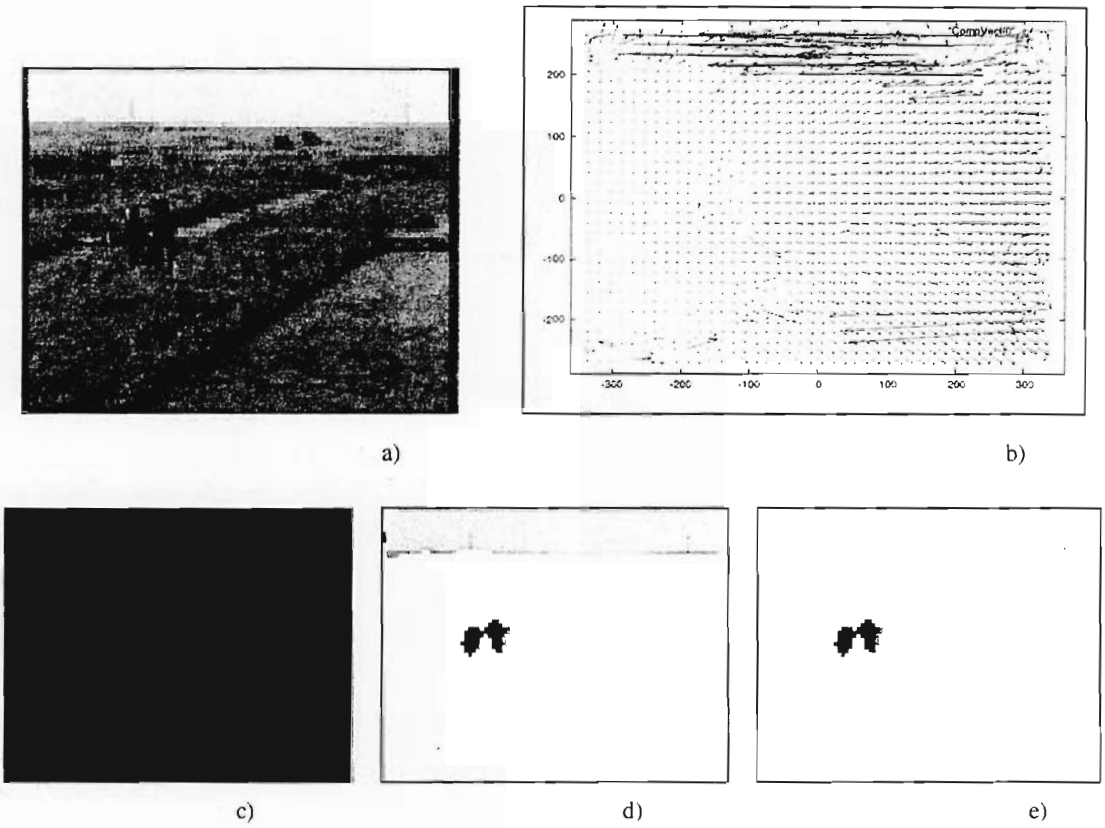


Fig 8: a)I-frame at DC resolution with a flat zone (the sky); b) the corresponding map of motion vectors; c) Object mask for the frame; d)objects obtained after superposition of mask and colour segmented I-frame; d) final result after flat zones filtering.

In Fig 8 a) an I-frame with a flat zone is presented. In this case the flat zone is represented by the sky. As it is shown in fig. 8 b) the presence of constant colour zones can led to motion vectors with incoherent direction and values. This situation is typical of foreground objects and, for this reason, a motion mask is detected even if any object is present (fig. 8.c). So after the superposition of motion mask with colour segmented DC image, the flat zone is present as a foreground object (fig. 8d). The approach we have chosen consist in filtering of flat zones according to mean gradient energy of region R (6).

$$\overline{E}(R)=\frac{1}{card(R)}\sum_{x,y}\left\|\overline{G}(R)\right\|^2 \qquad (6)$$

Here all connected component of masks which have the mean gradient energ y lower than a predefined threshold are excluded. Thus only objects with a colour heterogeneous content are preserved. In fact, it can be supposed that an object is composed of more than one region of constant colour and so its mean gradient energy value is higher that the one of a region composed by only a homogeneous region. For this reason, our flat zone filtering consists on calculating the colour gradient value of all the detected foreground objects and on deleting the regions with a very low value of it. Fig. 8e) shows the results of the filtering process. It can be noticed that even if, in this case, the objects of interest present a quite homogeneous colour, its gradient values are not as low as the one of a flat region and so they are preserved.

4. RESULTS AND PERSPECTIVES

The motion and colour based approach we have presented in this paper has been tested on different sequences from a set of natural video content. We considered feature documentaries "De l'Arbre à l'Ouvrage", "Hiragasy", "Aquaculture", "Cavitation", "Chancre" (SFRS ®), the length of each movie being of about 15 minutes.

We suppose that segmentation of video into shots has been already done, so we randomly select shots amongst those containing foreground objects.

As for camera motion for the set of processed content, pan, tilt, zoom and hand carried camera motion artefacts have been observed.

The examples of results obtained for movies "Hiragasy" and "Aquaculture" are shown in Figure 9. They have been validated by visual comparison between the original image and extracted foreground objects. Here our goal of detection of moving objects in I-frames is attained. The method shows some imprecision on the object border and also a slight over-detection due to the presence of false MPEG2 vectors in case of strongly textured image. The recall of the method is about 80% and a non –detection is observed if objects are too small that is cover of about 5-10 macro-blocks in a compact area. The over- detection is expressed as the merge of close objects and is observed in all situations when objects are situated at the relative distance of the order of macro-block size. In all I and P-frames, the outliers due to the camera motion on the frame borders were correctly removed even if the camera motion was not a pure translation.

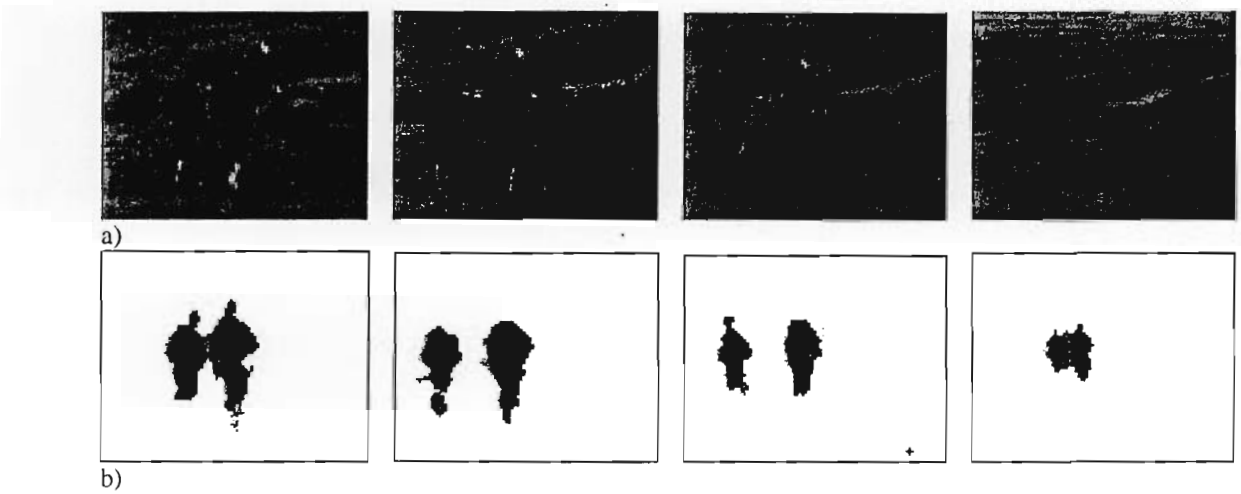
The algorithm has also been tested in limit situation, that is to say with objects so near to the camera to cover a large part of the background (30%) and in the case of no foreground objects. In the first case the robustness of the motion detection algorithm has led to a correct extraction of the camera motion parameters and consequently to a correct detection of the object in foreground. In the second case, even if the noise present on MPEG2 motion vectors has caused the presence of macro-blocks with high weight value in motion estimation, the filtering fulfilled by 3D segmentation (sec. 3.1), refinement by colour segmentation (sec. 3.3) and flat zone filtering has permitted to classify the "outliers" as pure noise.

The whole method performs at about 1 GOP (15 i/GOP) per second at Pentium 4 general computer without any fast processing instructions.

Thus in this paper we presented a method for foreground object extraction in the "rough indexing" paradigm which allows extraction of foreground objects in MPEG2 compressed video at I-P temporal resolution.

The method performs in a near real-time and gives promising results.

Taking into account clearly defined situations of difficulties for the method both in case of close objects and instantaneous weak relative motion with regard to camera motion, the actual work is devoted to the fast spatio-temporal filtering of detection results in order to smooth the detection in time and also to determine trajectory of objects along complete video shot.



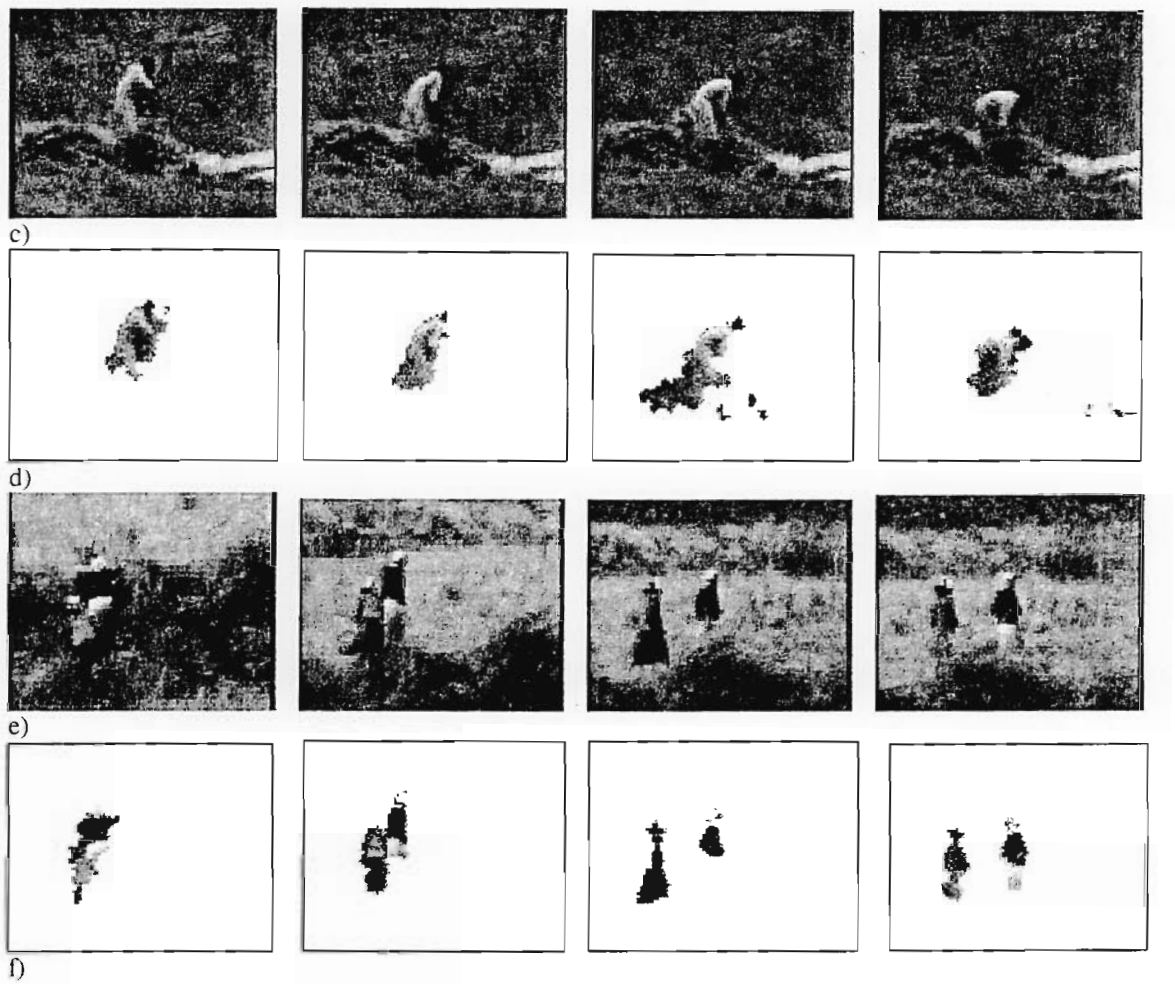


Fig. 9: Results of extraction of foreground objects from MPEG2 compressed video on I-frames (DC – images): a). b) "Aquaculture en Mediterranée", SFRS, c).f) "Hiragasy", SFRS.

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. ISO/IEC JTC 1/SC 29/WG 11/M6156, MPEG-7 Multimedia Description Schemes WD (Version 3.1), Beijing, July 2000.

3. Fauquier J., Boujemaa N., "Region-based retrieval : Coarse segmentation with fine signature", *IEEE ICIP'02*

4. N.H. AbouGhazaleh, Y. El Gamal, "Compressed video Indexing based on object motion", *VCIP'2000*, Perth, Australia, June 2000, pp. 986-993

5. Y.P. Tan, D.D. Saur, S.R. Kulkarni, P.J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation", *IEEE Trans. On CSVT*, 01/2000, pp 133-146

6. M. Durik, J. Benois-Pineau, "Robust Motion Characterisation for Video Indexing Based on MPEG2 Optical Flow", in *Content Based Multimedia Indexing*, Brescia, Italy, September 2001.

7. S. Benini, E. Boniotti, R. Leonardi and A. Signoroni, "Interactive Segmentation of Biomedical Images and Volumes using Connected Operators", *2000 International Conference on Image Processing, ICIP2000*, Vancouver, Canada September 2000.

8. B-L Yeo, B. Liu, "On the Extraction of DC Sequence from MPEG Compressed Video", in *1995 International Confer-*

ence on Image Processing, ICIP95, **vol. 2**, Washington DC.
9. P. Bouthemy, M. Geldon, F. Ganansia, "A unified approach to shot change detection and camera motion characterisation", *IEEE Circuits & Systems for Video Technology*, **vol. 9**, num. 7, 01/10/99, pp. 1030-1044.
10. 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
11. J. Benois-Pineau, H. Nicolas, "A new method for region-based depth ordering in a video sequence: application to frame interpolation", *Journal of Video Communication and Video Representation*, **vol. 13**, pagg. 363-385.