ROBUST MOTION SEGMENTATION FOR HIGH DEFINITION VIDEO SEQUENCES USING A FAST MULTI-RESOLUTION MOTION ESTIMATION BASED ON SPATIO-TEMPORAL TUBES

Olivier Brouard, Fabrice Delannay, Vincent Ricordel and Dominique Barba

University of Nantes – IRCCyN laboratory – IVC team
Polytech' Nantes, rue Christian Pauc, 44306 Nantes, France
olivier.brouard@univ-nantes.fr

ABSTRACT

Motion segmentation methods are effective for tracking video objects. However, objects segmentation methods based on motion need to know the global motion of the video in order to back-compensate it before computing the segmentation. In this paper[1], we propose a method which estimates the global motion of a High Definition (HD) video shot and then segments it using the remaining motion information. First, we develop a fast method for multi-resolution motion estimation based on spatio-temporal tubes. So we get a homogeneous motion vectors field (one vector per tube). From this motion field, we use a robust approach to estimate the parameters of the affine model that characterizes the global motion of the shot. After back-compensation of the video shot global motion, the remaining motion vectors are used to achieve the motion segmentation and extract the video objects.

***Index Terms*—** Robust Motion Segmentation, Global Motion Estimation, Multi-Resolution Motion Estimation, Spatio-Temporal Tubes.

1. INTRODUCTION

The methods for tracking video objects are efficient to encode a video shot. Indeed, one possible strategy is to use adapted coding parameters for coding a given video object during all its *lifespan*. By video object, we mean typically, a spatio-temporal shape characterized by its texture, its color, and its own motion that differs from the global motion of the shot.

In order to track spatio-temporal objects in a video sequence, they need to be segmented. In the literature, several kinds of methods are described. These different methods use spatial and/or temporal [1, 2, 3] information to segment the objects. In the case of temporal information, it is necessary (when the camera moves) to know the global motion of the video to perform an effective video objects segmentation. Horn and Schunck [4] proposed to determine the optical flow between two frames. Thus, it allows to obtain information about the moving objects. Otherwise, the motion parametric model of the successive frames can be estimated [5]. Once the motion model is known, the global motion is back-compensated, and only the moving objects remain with their local motion information. In our method, we use a motion information per block, but for a group of frames (GOF), to estimate the global motion and realize the motion segmentation. This motion estimation needs to reflect as much as possible the *real motion* (*i.e.* the motion vectors of the scene real objects) of the block to give the right information to be processed. Exactly we use a motion estimation method which, considering several successive reference frames, estimates the motion of spatio-temporal tubes [6], with the hypothesis of an uniform motion along the GOF. Once, we have the motion vectors of the tubes, the second step is the estimation of the global motion. We adapt the motion vectors accumulation method described by Coudray [7] to estimate robustly the parameters of an affine motion model from the obtained motion vectors. Then, the global motion is compensated, and only the vectors which belong to moving objects remain non null. For the last step we realize the motion segmentation from them.

In the following section, we present the fast multi-resolution motion estimation method based on tubes. In section 3, we describe the robust computation of the global motion parameters, and in section 4, the motion segmentation method is given. Finally, we show the simulation results in section 5, and we conclude in section 6.

2. FAST MULTI-RESOLUTION MOTION ESTIMATION BASED ON TUBES

To obtain motion information correlated with the motions of real life objects in the video shot, we consider several successive frames and we make the assumption of an uniform motion between them. The fixing's time of the Human Visual System (SVH) is about $200\ ms$ [8], so as the next generation of HDTV will use 1920×1080 as frame definition in progressive mode with a frame rate of 50 Hz, our method will use a

[1] This research was carried out within the framework of the ArchiPEG project financed by the ANR (convention N° ANR05RIAM01401).

GOF composed of nine frames ($180ms$). Thus, we ensure the coherence of the motion along a perceptually significant duration. The current frame is at the center of the GOF. So, four past frames, and four future frames are located around it. We use the information along these nine frames to constrain the motion estimation, and obtain motion vectors more correlated with the real motion. In practice, we use only five frames from the GOF to evaluate the motion vectors, as illustrated in figure 1. We consider an uniform motion and we create a spatio-temporal tube. The tube allows to track a given block aligned on several successive frames. This constraint produces a motion vectors field more homogeneous (smoother) and more correlated with the real motion.

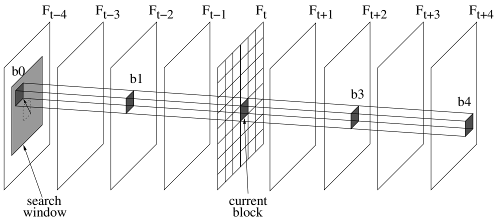


Fig. 1. Spatio-temporal tube: the five frames used to determine the motion vector of a given block.

We adapt the multi-resolution motion estimation method introduced by Péchard *et al.* [6] because it is well suited to deal with HD videos. The HD frames are spatially filtered and sub-sampled by a factor of six (first the frames are sub-sampled by a factor of two, and then by a factor of three), because the cutting in radial frequencies of the HVS is not dyadic and the angular selectivity varies with the radial frequency. Before each sub-sampling step, an appropriate (half-bandwidth, and then one-third bandwidth) low-pass filter is performed to avoid aliasing. From the filtered and spatially sub-sampled frames, we compute the motion estimation. Each block is simultaneously compared to its potentially corresponding blocks aligned in the previous frames and in the next frames, as illustrated in figure 1. The global error, $MSE_G = \sum_k MSE_k$ with $k = -4, -2, +2, +4$, is the sum of the four Mean Square Errors (MSE) between the current block and its corresponding blocks in the previous frames and in the next frames. The index k represents the positions of the reference frames in comparison to the current frame. MSE_k takes into account the three YUV components of each block. The chosen motion vector of a tube gives the lowest MSE_G between the current block and its corresponding blocks in the four frames. The motion vectors are first estimated at the lowest resolution. Then, they are up-scaled appropriately to the higher resolution to be used as an initial search point for the motion estimation at the new resolution. These aligned blocks constitute a spatio-temporal tube.

Finally we get a motion vectors field with one vector per tube, each motion vector is applied to the block of the image F_t in the center of the tube. This motion vectors field is the input of the next process : the global motion estimation.

3. ROBUST GLOBAL MOTION ESTIMATION

Once we computed the motion estimation, the obtained motion vectors reflect more effectively the motion of the real life objects. The next step is to identify the parameters of the global motion of the GOF from this motion vectors field. Several models exist to estimate the global motion of a video, we use an affine model with six parameters as written in Eq. 1:

$$\begin{pmatrix} V_x \\ V_y \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \tag{1}$$

where a_i ($i = 1 \dots 4$) are the deformation parameters (exactly a_1 and a_4 characterized the zoom, and a_2 and a_3 the rotation), t_x and t_y are the translation parameters. V_x, V_y, x and y are respectively the horizontal and vertical components of each motion vector, and the spatial position of the block in the image F_t.

3.1. Global motion parameters estimation

We adapt the method introduce by Coudray [7] who estimated the global motion from a MPEG2 stream. For us the basic information used to estimate the global motion is a motion vectors field, one motion vector per tube. We estimate the parameters of the affine model from the motion vectors field using the following equations:

$$\begin{array}{c} a_1 = \frac{\partial V_x}{\partial x}, \quad a_4 = \frac{\partial V_y}{\partial y}, \quad a_2 = \frac{\partial V_x}{\partial y}, \quad a_3 = \frac{\partial V_y}{\partial x}, \\ t_x = V_x - a_1.x - a_2.y, \quad t_y = V_y - a_3.x - a_4.y. \end{array} \tag{2}$$

The global motion is often due to the camera motion, which can be relatively complex. As a zoom or a rotation affects the estimation of the translation parameters (because all these parameters are correlated), the global motion estimation (GME) is realized in two steps. First, we compute the deformation parameters from each motion vector. Each calculated derivative produces an assumption for the corresponding global motion parameters. To find the main assumption (with the highest probability) that matches the global motion parameter, the unit assumptions are accumulated in an histogram (one respective histogram for each global parameter). In order to gather the close assumptions, they are weighted using a Gaussian distribution.

The localization of the main peak of a given histogram produces the value retained for the corresponding parameter of the global motion model. To refine the localization of a main parameter, a least squares method is used to estimate the bend around the found position. Once the deformation parameters have been identified, they are used to compensate the original motion vectors field. Thus, the remaining vectors correspond only to the translation motions. These remaining motion vectors are then accumulated in a two dimensions histogram using a weighted Gaussian distribution (see Eq. 3):

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{3}$$

where x and y represent the compensated components of the motion vectors V_x and V_y respectively. The main values of the translation parameters are obtained by the localization of the main peak in the two dimensions histogram. For the last step the motion vectors are also compensated with the obtained translation parameters.

3.2. Weights of confidence for a robust estimation

For a given block of F_t (see figure 1), the motion vector obtained from the motion estimation minimizes the MSE. If the blocks of the tube are located inside a uniform region in the images, the matching of blocks is not confident. Thus, the blocks of the tube are not necessarily the blocks that belong to the real object. In this case, the motion vector associated with the tube could not reflect the real motion. To be robust, such motion vectors should not contribute to the GME. For this reason the motion vector contribution has to be weighted in function of the spatio-temporal content of the blocks belonging to the tube. Briefly, the motion vectors from highly textured and oriented areas give more reliable information about the real motion than those which belong to uniform areas. Hence we exploit the spatial activity of the tube to qualify it. To compute the spatial activity of the blocks, we use spatial gradients. The higher the spatial gradients of the tube blocks, the more confident its motion vector is.

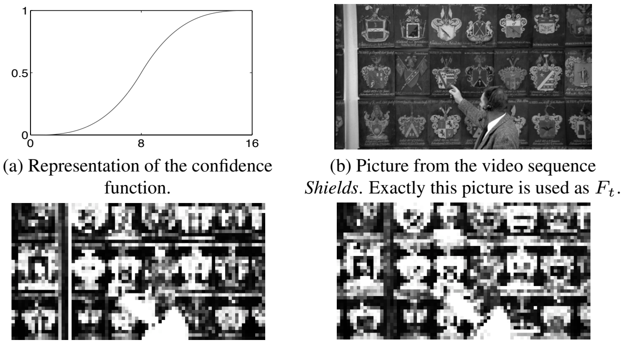
For each block, we use two spatial gradients, $\overline{\Delta V}$ and $\overline{\Delta H}$, which are respectively the average vertical and the average horizontal gradients. Depending on these features, a block may be labeled as a smooth, a fairly textured, or a highly textured area.

A block labeled as a highly textured area, can be it in only one direction, *i.e.* one of the two spatial gradients is high and the other is low. If the global motion is a translation in the same direction (vertical or horizontal) as the textured area, then the motion vectors of the blocks located in that region are not confident. So we distinguish the two spatial gradients. In practice, the confidence granted to the two components of the motion vector is calculated in a suitable way according to the blocks spatial gradients ($\Psi(\overline{\Delta H})$ and $\Psi(\overline{\Delta V})$). The function Ψ, to obtain the weights according to the spatial gradients, is computed as following (as illustrated in figure 2):

$$\Psi(x) = \begin{cases} (\frac{x}{8})^3/2, \ \ x \leq 8 \\ 1 - \Psi(16 - x), \ \ 8 < x < 16 \\ 1 \ otherwise \end{cases} \tag{4}$$

Figure 2 shows the weights of confidence for a GOF of the HD video sequence *Shields*. We observe that the maps that illustrate the weights of confidence of the GOF are different according to the orientations of the spatial gradients.

From now, to achieve a robust estimation of the global motion parameters, instead of the derivatives of the tube motion vectors, we use their weighted versions for their accumulation as described in the previous sub-section.



(a) Representation of the confidence function.

(b) Picture from the video sequence *Shields*. Exactly this picture is used as F_t.

(c) Map of the weights for the horizontal gradients (white = 1, black = 0).

(d) Map of the weights for the vertical gradients (white = 1, black = 0).

Fig. 2. Illustration of the confidence function.

4. MOTION-BASED SEGMENTATION

For the GME, we determined the translation parameters by the localization of the maximum in the accumulation histogram of the already compensated motion vectors. Assuming that each histogram peak represents an object motion, we do not retain only the main peak but all of them to segment the GOF.

In practice the first step consists of eliminating the noise. A reject threshold is defined empirically and all the values under it are fixed to zero. From the first detected peak which corresponds to the global maximum in the accumulation space, for all the positions connected to this peak, a local gradient is computed. This gradient is the difference between the populations of the two connected cells but in the direction of the maximum. As long as the gradient is positive, the tested position is considered as belonging to the peak, and the algorithm is iterated with the connected cells. At the end, all the positions belonging to the main peak are labeled. A new maximum is then detected among all the remaining (not labeled) cells, and the algorithm is iterated as long as there remain non null cells without label. If one cell is labeled as belonging to several peaks, it is linked to the closest peak.

5. SIMULATION RESULTS

We used one 1080p (*Tractor*), and two 720p (*Shields* and *New mobile calendar*) HD sequences from SVT [9]. These video sequences present different motions of the camera, and they contain one or several moving objects. We compare the results of our GME method to those from the Motion2D software [10]. Motion2D processes from a dense motion vectors field, and estimates robustly the parameters of the global motion using a multi-resolution least mean squares method (it is currently the reference method for the GME). As Motion2D computes the parameters of the global motion for two successive frames, we combine the obtained parameters for nine consecutive frames in order to compare them with ours. Figure 3 shows some characteristic results obtained with the *New mobile calendar* video sequence. The x-axis represents the GOF index. The global deformation parameters for the zoom

(a_1 and a_4) are not presented because they are estimated as near from zero according to the two methods. Excepted a slow zoom out, the only motions of the camera are translations. The two distinct objects (the calendar and the train) have their own translational motion. As result, Motion2D estimates the combination of the two objects translations as a global rotation, our method performs better and find only the objects translations. As a consequence, our estimation of the global translation is more accurate too. This result is specially due to the high regularity of our motion vectors field based on spatio-temporal tubes.

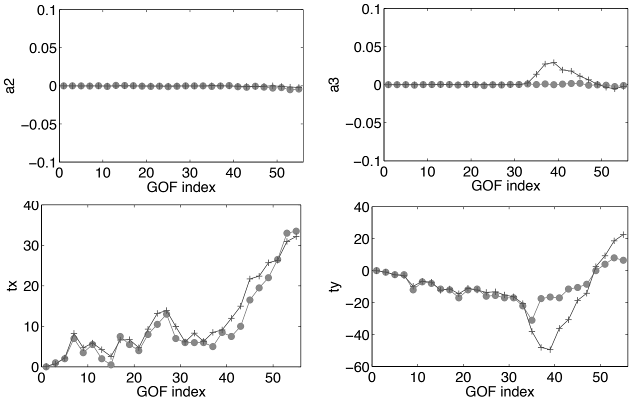
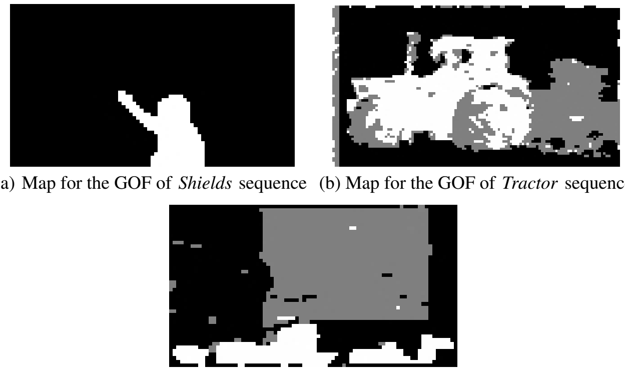


Fig. 3. Estimation of the global rotation parameters and the global translation parameters using our GME method (●) or Motion2D (+) for the *New mobile calendar* video sequence.

Figure 4 shows the results of the motion segmentation after our GME method for the three GOF. In the map (a), we observe clearly two different regions, which are the background and the moving man. The map (b) shows the result of the motion segmentation for one GOF of the video sequence *Tractor*. As the motion of the camera is a translation, uncovered areas appear on the sides of the map. So, the blocks located on these sides are not correctly segmented. But we observe that the tractor is clearly segmented in respect to the background. The tractor itself is segmented in different regions which are its "body" and the seeder. However, the wheels are not correctly segmented. Indeed, their motion is complex, as it is a combination of a rotation and a translation. In the last map (c), we observe clearly the different objects of the GOF (the background, the calendar and the train), but there are still some isolated blocks. Indeed, the motion vectors from these uniform areas are not confident for motion segmentation.

6. CONCLUSIONS

In this paper, we presented a motion segmentation method for HD sequences. First, a fast multi-resolution motion estimation based on spatio-temporal tubes using several reference frames is realized in order to obtain a field of motion vectors (one vector per tube) for the GOF (nine frames, with the assumption that the motion is uniform along it). The (derivatives) motion vectors are weighted according to the spatial activity of the tube, and accumulated in histograms. Then, we estimate robustly the global motion parameters of an affine model. The global motion is then back-compensated, and from the remaining accumulated motion vectors (with the assumption that each peak represents one object motion) the other peaks are extracted in order to achieve the motion based segmentation.



(a) Map for the GOF of *Shields* sequence (b) Map for the GOF of *Tractor* sequence

(c) Map for the GOF of *New mobile calendar* sequence

Fig. 4. Motion segmentation maps for the three GOF.

7. REFERENCES

[1] R. Megret and D. DeMenthon, "A survey of spatio-temporal grouping techniques," Tech. Rep., LAMP-TR-094/CS-TR-4403, University of Maryland, 1994.

[2] F. Porikli and Y. Wang, "Automatic Video Object Segmentation Using Volume Growing and Hierarchical Clustering," vol. 3, pp. 442 – 453, March 2004.

[3] Y. Wang, J. F. Doherty, and R. E. Van Dyck, "Moving Object Tracking in Video," Washington DC, USA, October 2000, Proc. IEEE Applied Imagery Pattern Recognition Workshop.

[4] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, no. 1 – 3, pp. 185 – 203, 1981.

[5] J. M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, December 1995.

[6] S. Péchard, P. Le Callet, M. Carnec, and D. Barba, "A new methodology to estimate the impact of H.264 artefacts on subjective video quality," Scottsdale, VPQM 2007.

[7] R. Coudray and B. Besserer, "Global Motion Estimation for MPEG-Encoded Streams," Singapore, in Proc. ICIP 2004.

[8] O. Le Meur, P. Le Callet, and D. Barba, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802 – 817, May 2006.

[9] SVT, "Overall-quality assessment when targeting wide xga flat panel displays," Tech. Rep., SVT corporate development technology, 2002.

[10] IRISA, "Motion2D," Tech. Rep., http://www.irisa.fr/Vista/Motion2D/.