PROPOSITION OF A HUMAN MOTION TRACKING METHOD BY TEMPORAL-SPATIAL SEGMENTATION IN AN IMAGE SEQUENCE

Frédéric Elsner *, Khaled Hacine **, Naceur Kerkeni *, Jean-Claude Angué **, Michel Bourton *
* Laboratoire d'Informatique, Robotique et Reconnaissance des Formes
** Laboratoire d'Automatique Industrielle et Humaine

URIAH - UA CNRS 1118, Université de Valenciennes, Le Mont Houy, BP 311
59304 Valenciennes Cédex, France

ABSTRACT

In this paper, a new approach of a 3D human motion tracking method in real time is proposed. The basis of the method is the use of two segmentations, one temporal and the other spatial. It allows to extract the pertinent parts of the body (that are materialised by reflective markers) and to track them : the speed computing is only made on moving objects in the scene that are interesting to track. The use of stereo vision allows to calculate 3D position of each marker and a prediction of the 2D positions of all of them in the next pair of images is done to optimise the tracking. The implementation of the algorithm, which is realised on a transputer network, reduces the processing time.

INTRODUCTION

The domain of image sequence analysis, or dynamic scene analysis, is expanding widely. Applications such as moving target tracking, road traffic management, human motion analysis and, more generally, dynamic component analysis in a scene are varied ways of research [1-3].

This paper deals especially with human motion analysis (biomechanics), whose goal is to understand the mechanisms that rule human motions in areas such as medical rehabilitation, sport, conception of ergonomic interface, etc... The image is the basic tool which is used in motion analysis. The implemented means for human motion analysis were photography-based, cinema-based, and are now video-based. Therefore, the research in this domain has been greatly made easier since the appearance and the development of computer science, acquisition technics and data processing methods.

Because of the great quantity of data to treat in human motion analysis, it is necessary to define the needs and the constraints of an efficient system. So, we must first represent the subject in the way the most exploitable according to the chosen algorithms. We have to define the use conditions of the system and the technological limitations that prevent from getting the "ideal" analysis system : dynamic treatment in real time which allows the tracking of the fastest movements and the stocking of the fewest informations. Moreover, in every computer treatment, the greater the quantity of data, the more complex the treatment. In these circumstances, the parallel computers seem to give interesting solutions. Indeed, the parallel architecture allows to add processors when the computing power is not sufficient enough. The goal of this paper is to describe a human motion tracking method with a view to realise a human motion analysis system which will be parallelized and executed on a transputer-based parallel machine in order to satisfy the computing power needs.

This paper is divided in five parts. The first one reviews the features of the existing systems and establishes the schedule of conditions of the wanted system. We also give the means used to get an exploitable representation of the human body. So, the targets materialising the significant parts of the body are presented. The low-level treatment needed for the extraction of these targets in the images is explicited and an approach of a segmentation method, homogeneous in the mean of motion (spatial-temporal segmentation), is given with the architecture that results from all the treatment. In the second part, the spatial-temporal segmentation method is analysed in detail followed by the tracking method developed in the third section. The actual state of the work is given in the fourth part and the conclusions and prospects are detailed at the end of the paper.

A NEW APPROACH OF MOTION ANALYSIS

Human motion analysis is a need in domains such as medicine, ergonomics and sport. A study of the existing systems [4-7] reveals that all of them seem to suit to human motion analysis. This tends to prove that the most important features are those that can not be measurable (flexibility of use, easiness of implementation, fitness, noise immunity, working environment) but it shows that there is still some improvements to be done such as greater acquisition frequency, more cameras or more precision. A human motion analysis system must also be able to track fast and complex movements with precision. The solution for fast movements is to acquire images at a frequency as great as possible : the greater the frequency, the less differences in two successive images and so the better tracking because there is almost no differences between two successive images. For complex movements, when some parts of the body disappear for a moment and other reappear, the solution is to have sufficient cameras to always acquire all the parts with at least one pair of cameras : indeed, a 3D system implies the use of stereo vision. To minimise the constraints, the light which is used is the one of the working area. As shown in figure 1, the interesting parts of the human body (knee, elbow,

wrist) are materialised by bands (because regions are needed to compute the speed of the parts) coupled with small semi-spherical targets (to get the position of the pertinent part in the image).

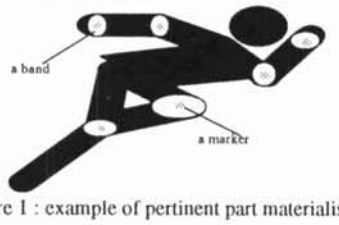


figure 1 : example of pertinent part materialising

To track them with precision, an image resolution of at least 512 x 512 pixels is needed. The low-level treatment that is necessary to extract the interesting moving objects in the scene is presented in figure 2. Only the bands are still in the image. When the low-level treatment is finished, a spatial-temporal segmentation enables to compute the bands speed and, in the same time, a target centre computing is done. The movement estimation is divided in three steps. The first step consists in obtaining regions by segmentation (these one being homogeneous in the mean of motion). The available information is the speed vector perpendicular to the region boundary, at a point (x,y). It's a local processing. The second step is the image structuration due to the segmentation in regions. It's called the intermediate structuration. Lastly, the third step consists in the estimation of each region speed. This is called the speed field estimation.

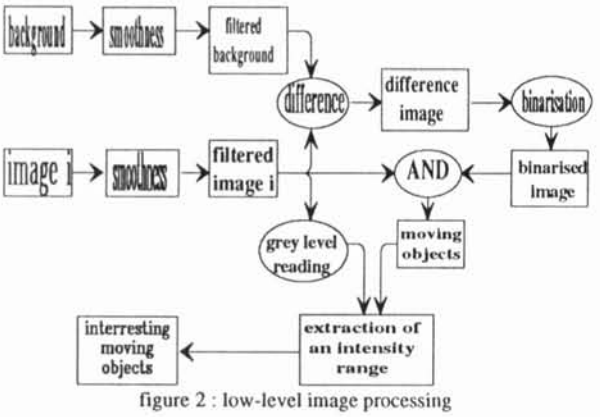


figure 2 : low-level image processing

The regions obtained with the low-level treatment enable the use of a spatial-temporal gradient based segmentation. The global speed of each region so given allows to treat the target loss problem .

The material architecture which results from our choice of processing is a parallel one. Indeed, only a such architecture allows a simultaneous tracking of several targets in a short time. Moreover, the huge sum of calculations is becoming important due to the nature of the treatment, so it is interesting to be able to add processors when necessary to work in real time.

THE SPATIAL-TEMPORAL SEGMENTATION

Let's remind that the segmentation algorithm used to compute the speed of the moving objects is based on a temporal-spatial gradient. All the segmentation treatment is defined in this section but the reader should refer to [8,9] for more details. The algorithm relies on a tri-dimensional modelisation (x,y,t) and a scheme of hypothesis.

The approach associates a constant speed model $\underline{T}=(a,b)$ to an homogeneous region in the mean of motion. We have a partial observation of the speed $\underline{V}$ at the point (x,y) at our disposal, namely the relation that links the speed $\underline{V}=(\frac{dx}{dt},\frac{dy}{dt})$ to the spatial gradient associated with the intensity function f, $\underline{\Delta}f(x,y)$:

$$\underline{\Delta}f(x,y) \,.\, \underline{V}(x,y) = -\frac{\delta f}{\delta t} \qquad (1),$$

which is an approximation at the first order. Let's consider the following expression :

$$e(x,y) = (\underline{T}-\underline{V}(x,y)).\underline{\Delta}f(x,y) \quad (2),$$

it shows the relation between the model and the real speed field at a point, i.e. e(x,y) is an error estimation function varying with the values of a and b. The projection of the real speed on the spatial gradient of the intensity can afford a directly measurable information about the motion.

Equation (1) represents a line perpendicular to the gradient at a point. So, we can assure that :

- in an uniformly lighted zone Z_i of the image, $\underline{\Delta}f=0$, and the motion is not detected or does not exist.
- in a zone Z_i where $\underline{\Delta}f\neq 0$, only the perpendicular to the region component of the motion is detected.

With the development of (1) and the introducing of (2), e(x,y) is clearly explicited by :

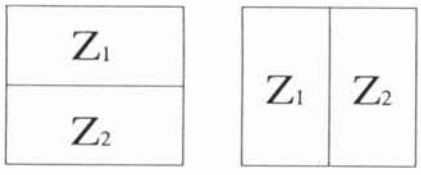
$$e_{a,b} = a.\frac{\delta f}{\delta x} + b.\frac{\delta f}{\delta y} + \frac{\delta f}{\delta t}$$

Once the model is defined by the parameters a and b, the homogeneity criterion must be given. It relies on a maximum likelihood test. Given a zone Z in the image, the problem is to select one of two cases :

- C_0 : for all the point in the zone Z, $e=e_{a_0b_0}$ (only one model M i.e. homogeneous zone)
- C_1 : for all the points in the zone Z_1, $e=e_{a_1b_1}$ and for all the points in the zone Z_2, $e=e_{a_2b_2}$ (two models M_1 and M_2 i.e. non-homogeneous zone)

To choose between the two cases, a likelihood function L_i is associated to each one and the log-ratio $\xi(a,b)$ of L_1 and L_0 is considered. If ξ is lower than a threshold λ, then the zone is homogeneous, else it is not and the zone is divided in two. The division is first made horizontally, then vertically (figure 3). Two log-ratio are computed and the chosen division is the one that gives the greater log-ratio. The treatment is applied again on the two zones. When there is only homogeneous zones in the image, an "intelligent" algorithm is needed to interpret the speed field and to define the regions of same speed.

Due to this segmentation method, the global speed of each region is given, and it allows to track the markers with the method in the following part.



(a) (b)

figure 3 : the two possible divisions of a zone Z
(a) horizontally (b) vertically

THE TARGET TRACKING METHOD

All the experimental scheme is represented by figure 4. Let's suppose that the motion is uniform [10] and the acquisition frequency is adapted to the speed of the motion. The tracking is made on two levels. There is a temporal tracking, which consists in the association of the corresponding markers in two successive images, and the spatial tracking, which consists in the association, by stereo vision with a pair of images, of the markers that represent the same physical point. There is also a prediction phase which is based on the computing of the 2D positions of the points in the next image pair. This calculation is useful to validate a choice in the temporal tracking and help in the correspondence problem when using stereo vision.

The goal of the temporal tracking is to define a bijection between the markers in the previous image and those of the current one. Indeed, each markers antecedent must be found in the previous image.

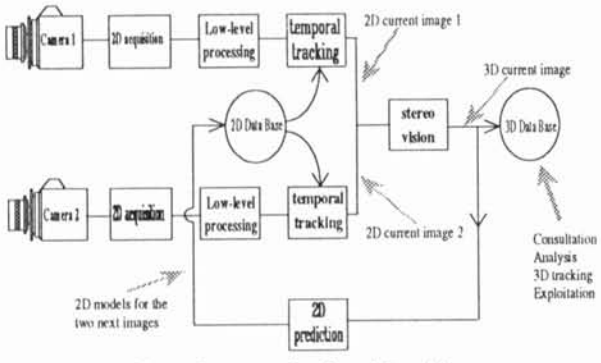


figure 4 : temporal and spatial tracking

This bijection is based on two criteria :

- "closer position" : this supposes that the distance between two markers is sufficient enough for them not to be merged and that the acquisition frequency is high enough in relation to the speed of the motion.
- "most identical optical flow" : this supposes a continuity of the speed field in the sequence of images (no sudden change of direction, orientation or value of the speed vector). The choice of this criterion is due to its wealth : it's a three dimension quantity (direction, orientation and value).

The change from the first to the second criterion is only done if the first one is not satisfactory. As mentioned above, a bijection between the markers of the current image and those of the previous one is needed. This supposes that the markers quantity is equal in two successive images. Unfortunately, this is not always true because of the momentary loss or reappearance of some markers. Moreover, there can be conflicts during the identification of the markers (correspondence making). For example, two markers of the current image (respectively the previous one) can have the same antecedent (respectively the same image) in the previous image (respectively the current one). All the conflicts are resolved by the creation of estimated, computed quantity (a priori) from the above criteria. The estimated quantity are computed with an interpolation of previous quantity from the image sequence.

Our method relies, just as the majority of the existing ones [11,12], on a three step scheme which can be formulated as follow :

First step - Creation of fictive markers according to the previous image (or images) : For all the points in the previous image, the fictive position is computed with the previous position and speed. The current speed is considered equal to the previous speed.

Second step - Hypothesis creation and likelihood computing : For all the points in the previous image and for all the points in the current image (even fictive ones), a value is associated to each couple of markers. This value is the possibility for two markers to represent the same one. This computing depends on the fact that a marker can be fictive or real (priority to the last one) and on the position and speed of each marker.

Third step - Hypothesis verification : All the couples of markers which value is greater than a given threshold must be rejected.

This scheme is done from image I-1 to image I and then from image I to image I-1. A global verification of the two partial computings assures the robustness of the final results

Let's remark that our method tries to keep a constant number of markers in the image by the creation of as many fictive markers as needed. This allows to get 3D positions by correspondence making by stereo vision easier. Indeed, the markers to associate choice is resolved by the temporal tracking because the markers are located identically and because all the markers have a projection in each image

ACTUAL STATE OF THE WORK

The low-level processing has been implemented on the Allen-Bradley Servovision Expert PVS2805 system and tested for static images. The results are quite good but it is necessary to get a real sequence of images, with changes of light, to improve the treatment.

The temporal-spatial algorithm in itself has been implemented in 3L Parallel C on a T800-20 transputer network (INMOS B008 board). The processing has been made on whole images and with 4 processors. The processing time is 500ms (very far from the temporal need) and the speed-up ($\frac{time\_with\_1\_processor}{time\_with\_p\_processors}$) is 3,6. To reduce the processing time, the treatment is being implemented not on the whole image but on windows that include each a pertinent part. We estimate that, with 4

processors and according to the fact that the processing time is proportionnal to the pixel quantity to be treated, the processing time should be 25ms. It is closer to the temporal need and we assume that a greater quantity of processors, with the fact that the processing time should be proportionnal to the squarred number of pixels to be treated for example, will give better results and will allow us to do all the process in less than 20ms. The tracking algorithm is actually tested and being improved.

PROSPECTS AND CONCLUSIONS

In this paper, a new approach of a 3D human motion tracking method has been described. The use of two segmentations (one temporal, the other spatial) allows to extract the pertinent parts of the human body and to track them. The temporal-spatial segmentation only computes the speed of each moving regions in the scene that are interesting to track. Those latters are extracted from the scene by a low-level image processing. Finally, the future use of stereo vision will allow to compute the 3D positions of the markers and a prediction of the 2D positions in the next pair of images will be done to optimise the treatment.

The introduction of the bands allows us not to use a very strict environment (extra light, special background colour, etc...) because the low-level processing that extracts the bands eliminates noises in the images.

At that stage of the study, some treatments are only being implemented and the construction of a validation line is proposed, before implementation on a parallel transputer-based machine. The image sequence is recorded on two tapes (from two cameras) with a time code as shown in figure 5. The images are then replayed, stocked and treated by passage as in figure 6.

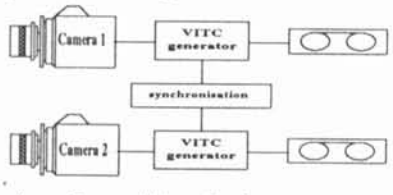


figure 5 : acquisition of an image sequence

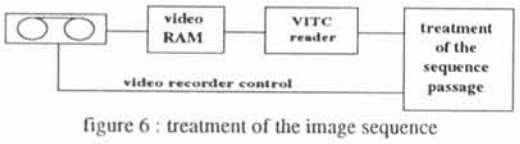


figure 6 : treatment of the image sequence

The advantages of a parallel architecture for human motion analysis is that the configuration of the processing unit can be optimised in relation to the motion parameters: speed of motion, needed precision, quantity of markers, kind of motion. It is also easy to increase the number of cameras in view to treat more complex motions.

REFERENCES

[1] W. Long and Y.-H. Yong
"Log-tracker : an attribute-based approach to tracking human body motion"
International Journal of Pattern Recognition and Artificial Intelligence, vol 5, n° 3, pp 439-458, 1991

[2] A. Cumani, A. Guiducci and P. Grattoni
"Image description of dynamic scene"
Pattern Recognition, vol 24, n° 7, pp 661-673, 1991

[3] V. Cappellini, editor
"Time-varying image processing and moving object recognition, 2"
Proceedings of the 3rd International Workshop, Florence, Italy, May 29-31, 1989

[4] P. Cloup, J.C. Angué and E.M. Laassel
"A new system of gestual automatic 3D analysis (SAGA-3)"
12th annual meeting of the American Society of Biomechanics, University of Illinois at Urbana Champaign, September 1988

[5] V. Macellari
"COSTEL : a computer peripheral remote sensing device for 3-D monitoring of human motion"
Medical & Biological Engineering & Computing, pp 311-318, May 1983

[6] J.A. Towle
"CODA-3 : A three-dimensionnal measurement system for use in kinematics"
Published by the Institution of Electronic and Radio Engineers, "Progress reports on Electronics in Medicine and Biology", 1986

[7] D.L. Mitchelson
"State of the art in automated 3-D motion monitoring systems"
Biomechanics at Unea, Sweeden, June 1985

[8] P. Bouthemy and J. Santillana-Rivero
"Segmentation en régions selon des critères de mouvements dans une séquence d'images"
Proceedings of the 6th Workshop on Pattern Recognition and Artificial Intelligence, pp 105-114, Antibes, France, November 16-20, 1987

[9] P. Bouthemy
"Estimation et structuration d'indices spatio-temporels pour l'analyse du mouvement dans une séquence d'images"
Traitement du signal, vol 4, n° 3, pp 239-257, 1987

[10] M. Jenkin
"Tracking 3D moving light display"
Proceeding in Workshop motion : representation control, pp 171-175, Toronto, Canada, 1983

[11] I.K. Sethi and R. Jain
"Finding trajectories of feature points in a monocular image sequence"
IEEE Transaction on Pattern Analysis and Machine Intelligence, vol 9, n° 1, January 1987

[12] J. Weng, T.S. Huang and N. Ahuja
"3D motion estimation, understanding and prediction from noisy image sequences"
IEEE Transaction on Pattern Analysis and Machine Intelligence, vol 9, n° 3, pp 370-389, May 1987