Tracking Densely Moving Markers

N. Alberto Borghese and Paolo Rigiroli[b)]
Laboratory of Human Motion Analysis and Virtual Reality (MAVR), Department of Computer Science, University of Milano, Via Comelico 39/41, 20135 Milano, INB-CNR, H.S.Raffaele, www.inb.mi.cnr.it/MAVR.html. b) Currently at Electronic Arts Canada, Content and Technology Group, Vancouver, Canada

Abstract

In this paper a new approach to the reconstruction of 3D trajectories of dense marker sets is proposed. Key element is the use of multiple passes to reconstruct the spatio-temporal structure of the movement with high reliability. First the tracking procedure computes a coarse structure of the motion, which is then recursively refined disambiguating difficult classification of the markers. The tracking procedure is based on integrating the temporal dimension in the matching process, by analyzing strings instead of points to derive more robust matches. Strings are analyzed using smoothness, n-focal constraints, and fitting of a skeleton to derive a proper matching. An innovative augmented-reality like interface greatly simplifies the labeling task. Lastly, a proper value for the critical parameters is automatically derived. Results on real data show that the system is able to produce a robust and largely complete set of trajectories, which greatly minimize the time required by post-processing.

1. Introduction

The acquisition and analysis of 3D human motion is expanding its influence from its native field, medicine (neurology, orthopaedics, neurosurgery and rehabilitation) to robotics, ergonomics, and computer animation. This prompts the continuous evolution of motion capture systems. When real-time is an issue, electro-magnetic technology competes with image processing; otherwise video camera based motion capture systems are preferred as they combine the least encumbrance to the subject with accuracy and reliability.

The use of natural video sequences is not mature yet to get fast accurate and reliable measurement [1]. In the marketed systems motion capture is facilitated markering the anatomical joints of the subject in motion (Fig. 1). Markers are detected by a suitable hardware [2], which feeds in real-time a host computer with their coordinates. These multiple sets of 2D points (one set for each camera) are then converted into the 3D motion of the markers by a procedure called *tracking* (second level).

A reliable tracking of natural body motion is still one of the most challenging tasks in computer vision [3-5]. As human skeleton is a highly articulated structure, twists and rotations make the movement fully three-dimensional. As a consequence, each body part continuously moves in and out occlusion from the view of the cameras, such that each of them can see only a chunk of the whole trajectory. Chunks from the different cameras have to be correctly matched and integrated to obtain a complete motion description. This difficulty is greatly enhanced when a dense set of markers is adopted.

Different criteria have been used to solve tracking. Epipolar, trifocal and quadrifocal constraints have been largely used to match the data points [6]. However, because of intrinsic limited accuracy in calibration and data measurement, these constraints do produce false matches, which, in dense markers configuration with many cameras, do make tracking quite difficult and prone to errors. These have to be manually corrected in a painful post-processing stage. Additional constraints must therefore be added. Temporal constraints based on the assumption that human motion is smooth [3, 7] have been introduced; a more robust approach integrates the previous criteria with solutions in which the skeleton-like structure of the body is taken into account [1].

We propose here a robust integration of these criteria by a new procedure, which is based on two key ideas: processing strings instead of points and multiple passes at different resolution. The resulting algorithm is not meant to work in real-time, but it does produce a robust and largely complete set of trajectories which greatly minimize the time required by painful post-tracking editing sessions. Moreover, an innovative augmented-reality like interface greatly simplifies the labeling task.

2. The tracking structure

Tracking is constituted of three sequential stages: pivotal tracking, holes filling and labeling. Each of these stages uses the tracking engine described in Section 3.

First, pivotal tracking uses the engine once to create reliable 3D strings. In this step only long 2D strings are used (coarse resolution). This allows obtaining a very robust 3D spatio-temporal structure of the motion. These 3D strings are then back projected over the cameras image plane to derive a robust re-initialization of 2D strings. The hole filling and labeling stages eventually succeed to complete the trajectories in a more principled way.

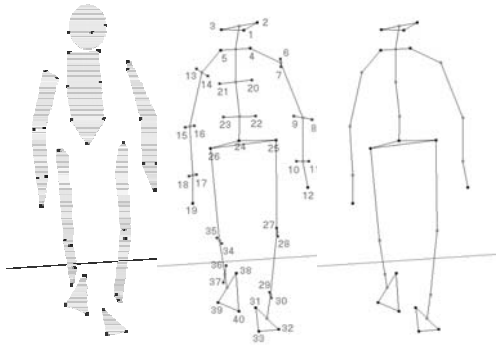

Fig. 1. (a,b) a typical marker disposition. A total of 40 markers are attached on the subject. Two markers are attached to the elbow, wrist, knee and ankle. This redundant set of markers prevents the loss of data and allows a better identification of the joint center. The joint centers can therefore be regarded as virtual markers (light gray panels (b,c)). The line segments identify the distances, which are almost constant during motion. The skeleton produced by the tracking procedure is obtained connecting measured and virtual markers (c).

3. The tracking engine

First, a 2D tracker builds 2D strings (segments of 2D trajectories) on each camera; these 2D strings are matched over pairs of cameras through the epipolar constraint to obtain a set of *candidate* 3D strings The novel idea of matching strings instead of points adds more reliability to this process. The candidate strings are then condensed and false matches are eliminated in the registration phase.

3.1 2D tracking

The 2D string, defined as the temporal concatenation of the 2D markers position, is the key element. The data are analysed separately on each camera, frame by frame. For each string, its position is predicted from the position in the previous frames. When a marker is close to this prediction *and* it is not close to any other marker, it is concatenated to that string, and the string is grown one more frame. If more than one marker is close to the prediction, their classification is postponed to a later stage and the string is interrupted at the previous frame. This happens when two or more trajectories come close one to the other on the image plane of the camera. When there is no 2D point close to the prediction (the marker is hidden to the view of that camera), the string is terminated also in the previous frame. These situations are very common when dense sets of markers are adopted. When a marker is not close to any existing 2D string, it is assumed the beginning of a new string.

At the end of this phase, the shortest strings are eliminated. In fact, the 2D tracker does generate spurious 2D strings. As these are two or, exceptionally three-four frames long, they can be easily eliminated filtering out the shortest strings. The shortest strings will be reconsidered in the hole filling and labelling tracking phases, where a robust initialization of 2D strings is obtained from 3D information.

3.2 Matching 2D strings

Matching features over multiple images is a very active area in computer vision and different solutions have been proposed [6]. These are mainly based on the multi-collinearity constraints applied on a frame by frame basis to obtain real-time performance. When dense marker sets are considered, these techniques do produce a lot of false matches which makes further processing hard. To avoid this, it is proposed here to integrate the temporal dimension into the matching process. This is obtained by matching pairs of strings instead of pairs of points: the epipolar condition:

$$\mathbf{p}_{il}(t_s)\mathbf{E}_{ij}\mathbf{p}_{jk}(t_s) = d_e \approx 0 \qquad (1)$$

has to be satisfied for all the t_s common to the two strings. $\mathbf{p}_{il}(t_s)$ and $\mathbf{p}_{jk}(t_s)$ are the points in the two strings, l and k, on camera i and j, measured at frame t_s. Condition (1) is applied only to pairs of 2D strings which have a large enough common temporal interval. Notice that (1) has not to be computed for all the points of the 2D strings: time sub-sampling can reduce the computational time. Moreover, when (1) is not satisfied for one pair of points, it automatically follows that matching is excluded also for all the other points in the two 2D strings. This procedure avoids most of the false matches. Instead, when (1) is verified for all the points of a pair of strings, the match is indeed much more robust.

4.3 Registration

At the end of the matching process, each matched pair of 2D strings produces a 3D string. When a marker is seen by M cameras, a total of [M(M-1)]/2 strings is produced for the same markers. Grouping these 3D strings into a single 3D string is the task of this step. To achieve this

scope, a 3D region with circular section is created around each 3D string. All the 3D strings contained inside this region are condensed into a single 3D string. The 3D string points are obtained as a weighted average of the 3D positions on the 3D strings involved, where the weight depends on the viewing angle of the cameras pair associated to a string, to achieve the best accuracy [8]. Notice that condensation of two strings can be safely carried out only when a reasonable number of frames is common to the two. Otherwise, if only a few frames are in common, condensation is postponed to a later tracking stage. This has two a-side advantages: the accuracy for that 3D string is increased and the string is lengthened up to the smallest starting frame and largest ending frame of all the constituent 3D strings.

At the end of this phase, possible residual false matches are eliminated as follows. First the 3D strings are sorted by multiplicity obtaining an ordered list. All the 2D strings associated with the first 3D string in the list are labeled as used. The other 3D strings are processed sequentially analyzing their constituting 2D strings. When a 2D string is already labeled as used (by a higher level 3D string) is taken out from the pool and eventually the 3D string is recomputed with the remaining 2D strings. If less than two 2D strings remain, the 3D string vanishes: it was in fact a ghost marker.

5. The tracking procedure

Tracking starts with the creation of a robust spatio-temporal structure of the motion, created in the pivotal phase. At the end of this phase, data are incomplete and 3D strings are broken into chunks, with few or more frames in between. Holes have to be filled. To the scope, the 3D strings are back-projected onto the image plane of the cameras. The markers not labelled, which are almost coincident with a projection can be safely associated to the 3D string. When a marker in two consecutive frames are classified to the same 3D string, a robust initialization of a new 2D string is obtained. The tracking engine is run again on these new 2D strings, and eventually it succeeds in extending the 3D string. At the end of this stage a much more complete set of 3D strings is obtained.

5.1 The body skeleton

At this stage the 3D trajectory has been reconstructed for the whole movement for most of the markers, but the concatenation of the multiple constituent 3D strings has not been carried out yet. Moreover, there are few blinking reflexes may exist, which have to be correctly identified and eliminated. To the scope, a 3D model of the skeleton is introduced (Figs. 1). This is constituted of a set of links connected by hinges, which represent respectively the bony segments and the body joints. Joints are hypothesized to produce a pure rotation and links to keep their length constant throughout the motion. The rotation center is approximately positioned in the center of the joint articulation and to get a better identification of it, multiple markers have been introduced in the biomechanics community [8]. Here pairs of markers are attached laterally on the elbow, wrist, knee and ankle joints (Figs. 1a-b). The mean position of the pair of 3D measured markers identifies the corresponding joint centers. These can be regarded as a virtual marker and are plotted in light gray in Figs. 1b-c. The skeleton output by the tracking procedure is obtained connecting virtual markers and real markers (Fig. 1c).

5.2 The augmented reality like user interface

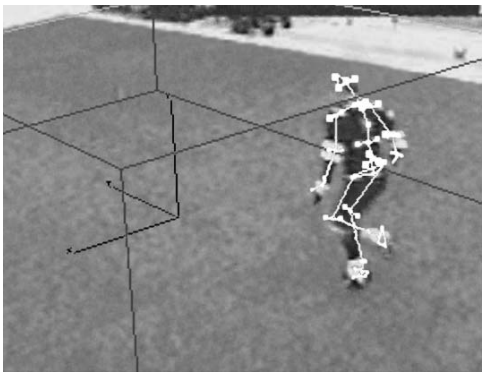

Fig. 2. The markers of the 3D strings are back-projected over the corresponding image of a video obtained with standard video-cameras. The power of this technique in helping the classification is evident.

The user has to initialize the model by associating once each body marker with a 3D string. A large help is given by an augmented reality like, display of the data as proposed here. One or more standard video cameras, temporally synchronized with the mocap cameras and spatially registered survey the scene. The 3D strings obtained by the tracking can therefore be back projected on the image plane of these cameras, and displayed superimposed to the video captured (Fig. 2). Classification becomes a trivial manual task also for dense markers sets. This is a very powerful tool to produce a reliable classification of dense markers ensemble in a very short time.

5.3 Labeling and interpolation

Once initial classification has been completed, tracking automatically extends the classification forwards and backwards to the unclassified strings.

To the scope, the classified strings are sorted by number of connected markers. For instance, in fig. 1c, most of the virtual markers are connected to four markers, the markers on the head to two and the markers on the hand to only the virtual marker on the wrist joint. The statistics of the associated links is then computed for the time interval in which they are present, to define boundaries to the variation in length of each link. For each classified 3D string, extension strings are searched forward and backwards, checking for compatibility with the measured length.

6. Parameter setting

Parameter setting is a crucial issue. A system, which requires to manually tune them, is of little help to real applications.

The critical parameters here are the error in the epipolar condition, d_e, ((1) in Section 3.2) and the radius of the circular section of the condensation region, r_c (Section 3.3). These are derived from the calibration data. Here the cameras are calibrated by using a rigid bar as described in [9]. This procedure allows collecting a large set of measurements, which can be used to derive a reliable experimental measure of the system accuracy. In particular, the computation of (1) for each pair of bar extremes matched over two cameras, allows determining a reliable statistics for the epipolar error and a proper value for d_e. r_c is related to the registration error. This can be evaluated by computing the statistics of the difference in the 3D position of the bar extremes reconstructed with two different pairs of cameras.

7. Results and Conclusion

The reconstruction of a running_out_of_balance sequence is plotted in Fig. 3. The success in tracking the whole body and in particular the hands and the feet is evident. The tracking procedures has required no intervention by the operator in the post-processing phase to correct marker swaps or wrong labeling. This has made this system, although the software was not optimized, overall faster than commercial packages in the production of 3D trajectories.

Although some techniques for automatic labeling are under study, based on a-priori model fitting a 3D skeleton [5] or statistical considerations [10], they may fail when complex gestures are recorded or dense markers sets are used. These considerations have prompted us to develop an efficient manual classification interface. The augmented-reality approach proposed here makes manual classification easy and fast for any operator.

References

[1] Kakadiaris and D. Metaxas, Model-Based Estimation of 3D Human Motion, IEEE Trans. Patt. Anal. Mach. Intell., 22(12), pp. 1453-1459, 2000.

[2] N.A. Borghese, M. Di Rienzo, G. Ferrigno and A. Pedotti, Elite: a goal-oriented vision system for moving objects detection, *Robotica,* 9, pp 275-282, 1990.

[3] S.B. Kang, R. Szeliski and H. Sum, A parallel feature tracker for extended image sequence, *Comp. Vision Image Underst*, 67(3), pp. 296-310, 1997.

[4] C.J. Veenman, M.J.T. Reinders, and E. Backer, Resolving Motion Correspondence for Densely Moving Points, *IEEE Trans. Patt. Anal. Mach. Intell.,* 23(1), pp. 54-72, 2001.

[5] L. Herda, P. Fua, R. Plankers, R. Boulic, and D. Thalmann, Using skeleton-based tracking to increase the reliability of optical motion capture, *Human Movement Science*, In press.

[6] R. Hartley and A. Zisserman. *Multiple view Geometry in Computer Vision.* Cambridge University Press, 2000.

[7] R. Mehrota, Esablishing motion-based feature point correspondence, *Pattern Recognition*, 31(1), pp. 23-30, 1998.

[8] . Davis, S. Ounpuu, D. Tyburski and J.R. Gage. A gait analysis data collection and reduction technique, *Clinical Biomech.* 1991*;* 575-587.

[9] N.A. Borghese and P. Cerveri, Calibrating a video camera pair with a rigid bar, *Pattern Recognition*, 33(1), pp. 81-95, 2000.

[10] Y. Song, X. Feng and P. Perona, Towards Detection of Human Motion, Proc. CVPR, pp. 722-728, 2000.

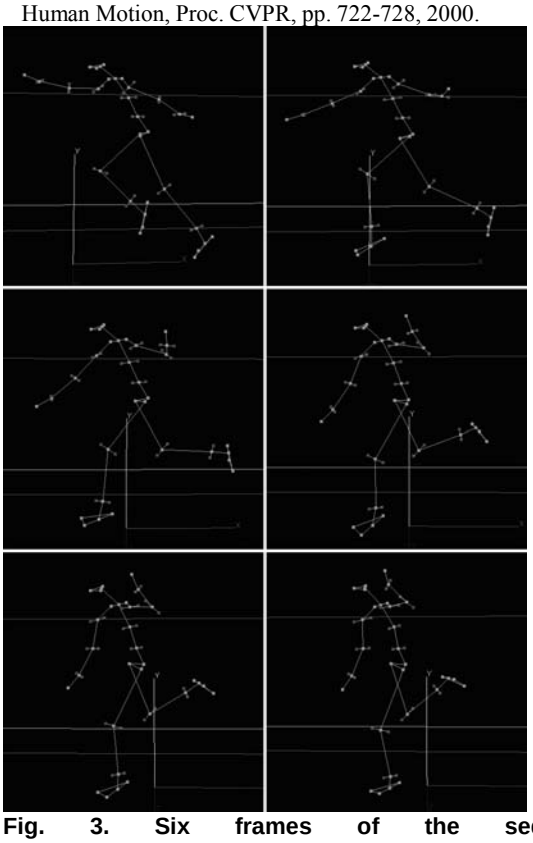

Fig. 3. Six frames of the sequence running_out_of_balance are plotted. Order: from left to right, from top to bottom.

....