AVATARS2000, 30.11.-1.12.2000,
Lausanne, Switzerland

MOTION CAPTURE BY LEAST SQUARES MATCHING TRACKING ALGORITHM

Nicola D'Apuzzo
Institute of Geodesy and Photogrammetry (IGP)
Swiss Federal Institute of Technology (ETHZ), Zurich, Switzerland
nicola@geod.baug.ethz.ch

KEYWORDS: Motion Capture, Tracking, Video Sequences, CCD Camera, Least Squares Matching

ABSTRACT

In this paper we present a method to extract 3-D information of the shape and movement of the human body using video sequences acquired with three CCD cameras. This work is part of a project aimed at developing a highly automated system to model most realistically human bodies from video sequences. Our image acquisition system is currently composed of three synchronized CCD cameras and a frame grabber which acquires a sequence of triplet images. From the video sequences, we extract two kinds of 3-D information: a three dimensional surface measurement of the visible body parts for each triplet and 3-D trajectories of points on the body. Our approach to surface measurement is based on multi-image matching, using the adaptive least squares method. A semi automated matching process determines a dense set of corresponding points in the triplets, starting from few manually selected seed points. The tracking process is also based on least squares matching techniques, thus the name LSMTA (Least Squares Matching Tracking Algorithm). The spatial correspondences between the three images of the different views and the temporal correspondences between subsequent frames are determined with a least squares matching algorithm. The advantage of this tracking process is twofold: firstly, it can track natural points, without using markers; secondly, it can also track entire surface parts on the human body. In the last case, the tracking process is applied to all the points matched in the region of interest. The result can be seen as a vector field of trajectories (position, velocity and acceleration) which can be checked with thresholds and neighborhood-based filters. Key points on the body can be defined in the vector field of trajectories and tracked through the sequence leading to a result comparable to the conventional motion capture systems.

1. INTRODUCTION

In the field of computer animation, the quality of the defined motion acts an important role. To increase the level of realism, the motion can be digitized from an actor performing the desired movements. This is achieved by motion capture systems, which can be divided into three major groups: magnetic, optical and mechanic systems. Different characteristics can be taken into account to classify them, e.g. accuracy, processing time, method used, costs, portability of the system. The magnetic systems use electromagnetic sensors connected to a computer unit which can process the data and produce 3-D data in real time (Ascension, Polhemus). The major advantage of these systems is the direct access to the 3-D data without processing. For this reason they are very popular among the animation community. An heavy disadvantage is the restricted freedom of movement caused by the cabling. Optical systems are mostly based on photogrammetric methods where the trajectories of signalized target points on the body are measured very accurately (Boulic et al. 1998, Fua et al. 1998, Vicon, Qualisys, Northen Digital). They offer complete freedom of the movement and interaction of different actors is also possible. In the last years, many improvements have been introduced (smart cameras, CMOS sensors) to achieve real-time or nearly-real-time acquisition. Motorized video theodolites in combination with a digital video camera have also been used for human motion analysis (Anai et al. 2000). Electro-Mechanical systems have recently appeared in the market: in this case the person acting the movements has to wear special suits with integrated mechanical sensors that register the motion of the different articulations (Analogus). This method too has the advantage of real-time data transfer from the sensors to the computer without processing and is usually cheaper than the magnetic ones.

Motion capture can also be achieved by image-based methods. They can essentially be split into monocular and multi-image systems. Monocular systems use sequence of image acquired by a single camera. To gain three-dimensional information from 2-D video clips, knowledge of the human motion has to be used. Some systems gain this knowledge by learning from provided sample training data and applying statistical methods to get the 3-D motion (Mahoney 2000, Song et al. 2000, Rosales and Sclaroff 2000). Other systems perform the tracking of defined human body models with constrains by sophisticated filtering processes (Deutscher et al. 2000, Segawa and Totsuka 1999, Cham and Rehg 1999).

Multi-image system use sequence of images acquired simultaneously by two or more cameras. Some systems assume a simple 3-D human model which is fitted comparing its projections into the different images to the extracted silhouettes of the moving person (Cheung et al. 2000, Delamarre and Faugeras 1999) or the extracted edges (Gravila et al. 1996). Other systems use image based tracking algorithms to track in 3-D the surface of the human body (D'Apuzzo et al. 2000) or the different body parts (Ohno and Yamamoto 1999). Mathematical models of the human motion can also be used to track directly in the 3-D data, which can be trajectories of known key points (Iwai et al. 1999) or dense disparity maps (Jojic et al. 1999).

In this paper, we present a method to recover from video sequences both 3-D shape and 3-D motion information. The core of this paper is the description of the least squares matching tracking algorithm (LSMTA), which uses the least squares matching process to establish the correspondences between subsequent frames of the same view as well as correspondences between the images of the different views. LSMTA offer the opportunity to gain 3-D motion information from video sequences without using markers.

2. LEAST SQUARES MATCHING TRACKING ALGORITHM

The least squares matching tracking algorithm (LSMTA) is a process composed of 5 steps: (1) acquisition of video sequences, (2) calibration of the system, (3) surface measurement for each frame, (4) surface tracking and filtering, (5) tracking e-points.

2.1. Data Acquisition and Calibration

Three synchronized CCD cameras in a linear arrangement are used. A sequence of triplet images is acquired with a frame grabber and the images are stored with 768x576 pixels at 8 bit quantization. The CCD cameras are interlaced, i.e. a full frame is split into two fields which are recorded and read-out consecutively. As odd and even lines of an image are captured at different times, a saw pattern is created in the image when recording moving objects. For this reason only the odd lines of the images are processed, at the cost of reducing the resolution in vertical direction by 50 percent. In the future is planned the use of progressive scan cameras which acquire full frames.

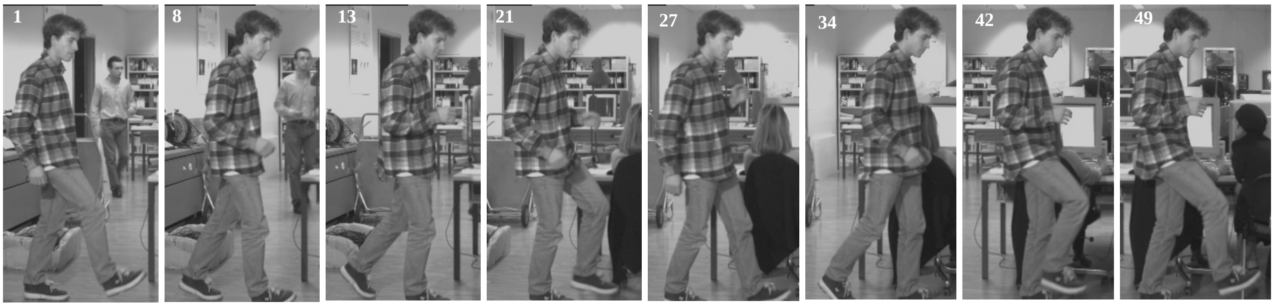

Figure 1: Acquired walking sequence (8 of totally 50 frames)

To calibrate the system, the reference bar method (Maas 1998) is used. A reference bar with two retroreflective target points is moved through the object space and at each location image triplets are acquired. The image coordinates of the two target points are automatically measured and tracked during the sequence with a least squares matching based process. The three camera system can then be calibrated by self-calibrating bundle adjustment with the additional information of the known distance between the two points at every location. The calibration process outputs are the exterior orientation of the three cameras (position and rotations: 6 parameters), the parameters of the interior orientation of the cameras (camera constant, principle point, sensor size, pixel size: 7 parameters), the parameters for the radial and decentring distortion of the lenses and optic systems (5 parameters) and 2 additional parameters modeling other effects as differential scaling and shearing (Brown 1971). A thorough determination of these parameters modeling distortions and other effects is required to achieve high accuracy.

2.2. Surface Measurement

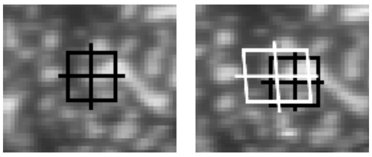

Figure 2: Least squares matching algorithm (LSM). Left: template image, right: search image

Our approach is based on multi-image photogrammetry. Three images are acquired simultaneously by three synchronized cameras. A multi-image matching process (D'Apuzzo 1998) establishes correspondences in the three images starting from a few seed points. It is based on the adaptive least squares method (Gruen 1985) which considers an image patch around a selected point. One image is used as template and the others as search images. The patches in the search images are modified by an affine transformation (translation, rotation, shearing and scaling). The algorithm finds the corresponding point in the neighborhood of the selected point in the search images by minimizing the sum of the squares of the differences between the grey levels in these patches. Figure 2 shows the result of the least squares matching with an image patch of 13x13 pixels. The black box represents the patches selected (initial location in the search image) and the white box represents the affinely transformed patch in the search image.

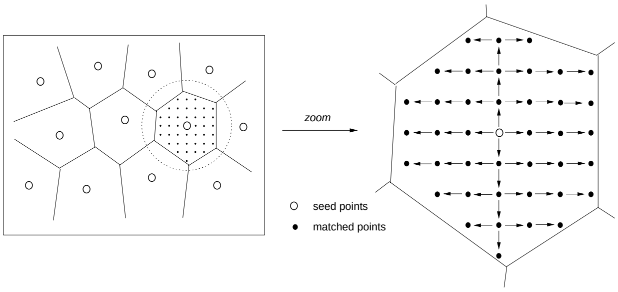

Figure 3: Search strategy for the establishment of correspondences between images

An automated process based on least squares matching determines a dense set of corresponding points. The process starts from a few seed points, which have to be manually selected in the three images. The template image is divided into polygonal regions according to which of the seed points is the closest one (Voronoi tessellation). Starting from the seed points, the stereo matcher automatically determines a dense set of correspondences in the three images. The central image is used as template and the other two (left and right) are used as search images. The matcher searches the corresponding points in the two search images independently and at the end of the process, the data sets are merged to become triplets of matched points. The matcher uses the following strategy: the process starts from one seed point, shifts horizontally in the template and in the search images and applies the least squares matching algorithm in the shifted location. If the quality of the match is good, the shift process continues horizontally until it reaches the region boundaries. The covering of the entire polygonal region of a seed point is achieved by sequential horizontal and vertical shifts (Figure 3).

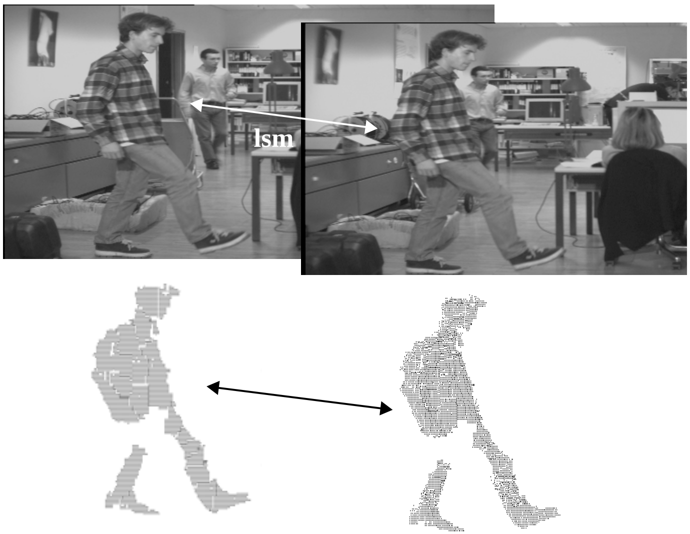

Figure 4: Image pair (top), and matched points (bottom)

To evaluate the quality of the result, different indicators are used (a posteriori standard deviation of the least squares adjustment, standard deviation of the shift in x and y directions, displacement from the start position in x and y direction). Thresholds for these values can be defined for different cases, according to the level of texture in image and to the type of template. If the quality of the match is not satisfactory (quality indicators are bigger than the thresholds), the algorithm computes again the matching process changing some parameters (e.g. smaller shift from the neighbor, bigger size of the patch). The search process is repeated for each polygonal region until the whole image is covered. At the end of the process, holes of areas not analyzed can remain in the set of matched points. The algorithm closes these holes by searching from all directions around them. In the case of poor natural texture, local contrast enhancement of the images is required for the least squares matching. Figure 4 shows and example using a pair of images.

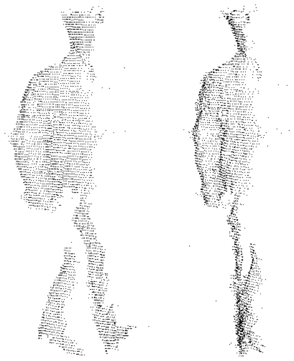

Figure 5: 3-D point cloud

The 3-D coordinates of the matched points are then computed by forward ray intersection using the orientation and calibration data of the cameras. To reduce remaining noise in the 3-D data and to get a more uniform density of the point cloud, a second filter is applied to the data. It divides the object space in voxels of variable dimensions and replace the points contained in each voxel by the center of gravity. The 3-D data resulting after this filtering process have a more uniform density and the noise is reduced. Figure 5 shows the 3-D point cloud derived from the images of figure 4.

Due to the poor natural texture of the shown example, the matching process produces a 3-D point cloud with relatively low density and high noise. In the future, it is planned to integrate in the matching process new functionalities such as geometric constraints and neighborhood constraints. This will improve the results in quality and density.

The process of surface measurement is performed for each frame of the acquired sequence and the obtained data are used by the surface tracking process.

2.3. Tracking Process

2.3.1. Tracking single points. The basic idea is to track triplets of corresponding points through the sequence in the three images, allowing the computation of their 3-D trajectories.

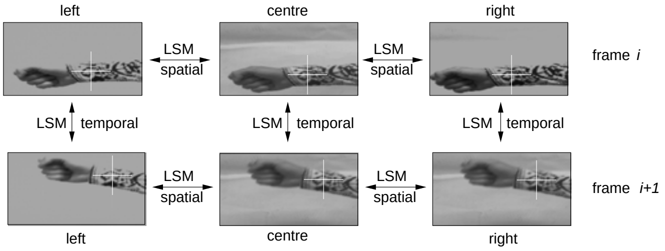

Figure 6: Temporal and spatial LSM

The tracking process is based on least squares matching techniques (figure 6). The spatial correspondences between the three images from the different cameras at the same time step (*spatial LSM*) and the temporal correspondences between subsequent frames of each camera (*temporal LSM*) are computed using the same least squares matching algorithm mentioned before.

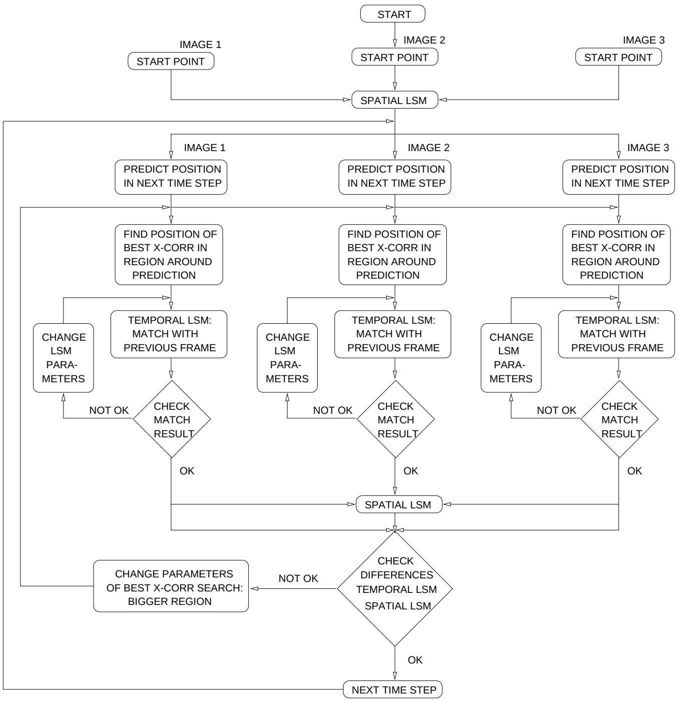
START
IMAGE 1
START POINT
IMAGE 2
START POINT
IMAGE 3
START POINT
SPATIAL LSM
IMAGE 1
PREDICT POSITION IN NEXT TIME STEP
IMAGE 2
PREDICT POSITION IN NEXT TIME STEP
IMAGE 3
PREDICT POSITION IN NEXT TIME STEP
FIND POSITION OF BEST X-CORR IN REGION AROUND PREDICTION
FIND POSITION OF BEST X-CORR IN REGION AROUND PREDICTION
FIND POSITION OF BEST X-CORR IN REGION AROUND PREDICTION
CHANGE LSM PARA-METERS
TEMPORAL LSM: MATCH WITH PREVIOUS FRAME
CHANGE LSM PARA-METERS
TEMPORAL LSM: MATCH WITH PREVIOUS FRAME
CHANGE LSM PARA-METERS
TEMPORAL LSM: MATCH WITH PREVIOUS FRAME
NOT OK
CHECK MATCH RESULT
NOT OK
CHECK MATCH RESULT
NOT OK
CHECK MATCH RESULT
OK
OK
OK
SPATIAL LSM
CHANGE PARAMETERS OF BEST X-CORR SEARCH: BIGGER REGION
NOT OK
CHECK DIFFERENCES TEMPORAL LSM SPATIAL LSM
OK
NEXT TIME STEP

Figure 7: Flowchart of the LSM tracking process

The flowchart of figure 7 shows the basic operations of the tracking process. To start the process a triplet of corresponding points in the three images is required. This is achieved with the least squares matching algorithm (*spatial LSM*). The process can then enter the tracking loop. The fundamental operations of the tracking process are three: (*1*) predict the position in the next frame, (*2*) search the position with the highest cross correlation value and (*3*) establish the point in the next frames using least squares matching (*temporal LSM*). These three steps are computed in parallel for the three images. Figure 8 shows graphically the process. For the frame at time *i+1*, a linear prediction of the position of the tracked point from the two previous frames is determined (step *1*). A search box is defined around this predicted position in the frame at time *i+1*. This box is scanned for searching the position which has the higher cross correlation between the image of frame at time *i* and the image of frame at time *i+1* (step *2*). The least squares matching algorithm is then applied at that position and the result can be considered as the exact position of the tracked point in the new frame (step *3*).

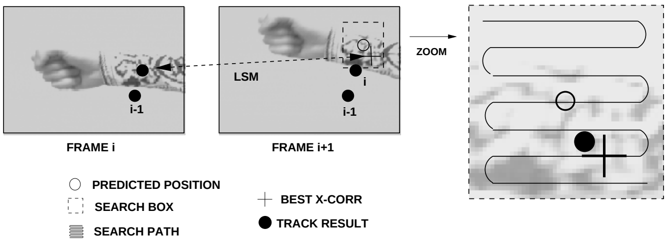

Figure 8: Tracking in image space: temporal LSM is applied at the position of best cross correlation

Like explained before, this process is performed in parallel for the three images of the different views. To test the individual results in the three images, a *spatial LSM* is then executed at the positions resulting from the *temporal LSM*s (see flowchart of figure 7) and if no significant differences occur between the two matches, the point is considered tracked and the process can continue to the next time step. On the other hand, if the differences are too large, the process goes back to step (*2*) by searching the value of best cross correlation in a bigger region around the predicted position. If the result is rejected again, the tracking process stops.

The results of the tracking process are the coordinates of a point in the three images through the sequence, thus the 3-D trajectory is determined by computing the 3-D coordinates of the point for each time step by forward ray intersection (figure 9). Velocities and accelerations are also computed.

This way of tracking points may produce errors which cannot be easily detected. In fact, the only control of the tracking result is the test executed between the *spatial LSM* results and the *temporal LSM* results. There is no 3-D control of the trajectories. Thus, false trajectories can be generated even if the tracking results seems good. A new test has to be integrated in the process to detect the false trajectories.

This can be achieved by tracking part of surfaces and not only single points. In this case, the result of the tracking process can be considered as a vector field of trajectories, which can be checked for consistency and local uniformity. Indeed, since the human body can be considered as an articulated moving object, the resulting vector field of trajectories must be locally uniform, i.e. the velocity vector must be nearly constant in sufficiently small regions at a particular time and filters can therefore be defined to check these properties. The next paragraph will describe the approach.

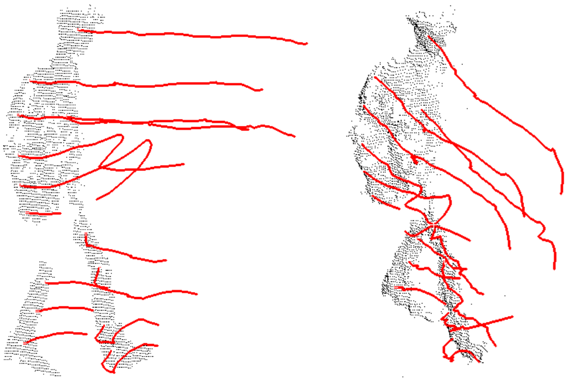

Figure 9: 3-D trajectories

2.3.2. Surface tracking. Tracking surface parts means track simultaneously points belonging to a common surface. Practically, the tracking process described in the previous paragraph, is applied to all the points matched on the surface of the first frames. With this approach, a new problem has to be considered: during the sequence, some surface parts can be lost by occlusion and new parts of the surface can appear (e.g. the legs which occlude each other during a walk sequence). For this reason, a new functionality has to be integrated in the tracking process. With a defined frequency (which can be for example every 2 frames), the data resulting from the tracking process is checked for density before proceeding to the next time step (see flowchart in figure 10). In the regions of low density, new points previously computed (surface measurement of the body for each frame) are integrated in the process, so that new appearing surface parts are also tracked.

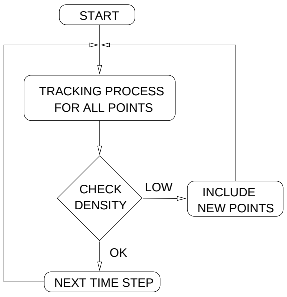

Figure 10: Flowchart of the surface tracking process

As explained before, the tracking process may produce false trajectories. This is clearly shown in figure 11, where the computed 3-D trajectories for 50 frames are displayed for a surface part of the arm. The false trajectories can easily be recognized because they don't follow the uniform movement of the majority. The vector field of trajectories can indeed be checked for consistency and local uniformity of the movement.

Two filters are applied to the results to remove or truncate false trajectories. The first filter removes the largest errors using thresholds of the velocity and acceleration. Depending on the movement performed (e.g. a running sequence would have larger values than a walking sequence) the two thresholds of maximal velocity and acceleration are defined at the begin of the process and remain constant during the sequence. The second filter checks for the local uniformity of the motion, both in space and time. To check this property, the space is divided in voxels and for each voxel a mean value of the velocity vector is computed at each time step. The single trajectories are then compared to the local mean values of the velocity vector and if the differences are too large, the trajectory is considered false and it is truncated or removed. The level of difference is computed as percentage of difference of the magnitude and angle to the local mean velocity vector.

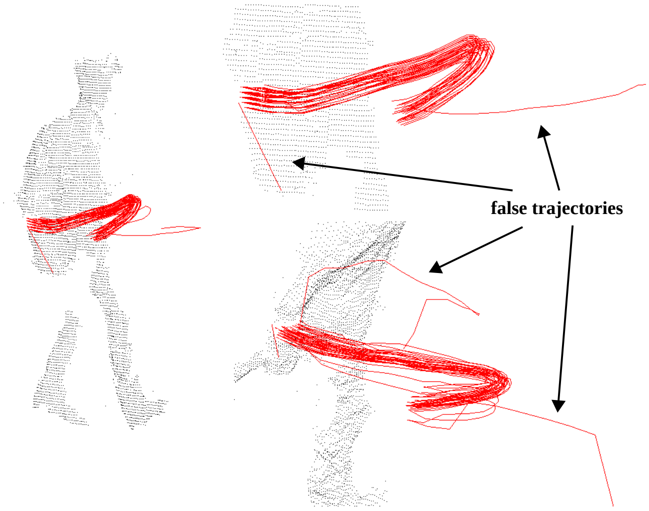

Figure 11: Surface tracking: the false trajectories don't follow the uniform movement of the majority

2.3.3. Tracking e-points. After filtering the trajectories, a problem still remains. An ideal trajectory starts from the begin of the sequence and lasts till the end (figure 12, left), but this is not the normal case since the length of the trajectories are varying from a minimum of 3 frame. The results are therefore a set of broken trajectories (figure 12, right). This effect is caused by occlusion, lack of texture, lost of track points, appearance of new points and cannot be removed. In order to solve this problem we have introduced the concept of e-point: extended point. The e-point is a 3-D region whose size can vary and its center define the position. The e-points are interactively defined in a graphical enhanced user interface. They can be easily placed and moved in the 3-D space. Since the vectorfield of trajectories represent points on the surface of the body, the e-points too have to be considered lying on the body.
They are tracked in a simple way: the position in the next time step is established by the mean value of the displacement of all the trajectories inside the region defined by the e-point. The size of the region to be tracked acts an important role: a larger size will provide lower accuracy but higher robustness. The most favorable compromise should be chosen.

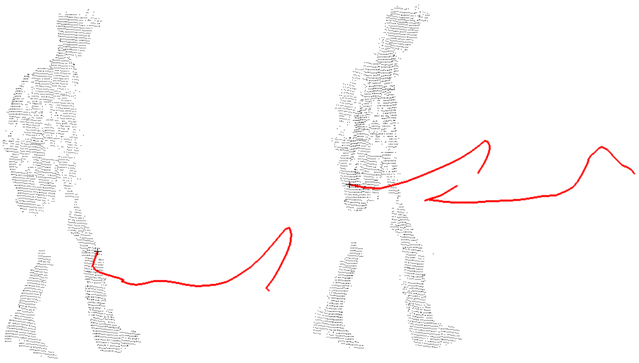

Figure 12: Ideal trajectory (left), broken trajectory (right)

2.4. Tracking key e-points of human body

We are interested in the capture of the movement of the human body, therefore we select key e-points which can define the motion. Depending on the complexity of the movement we want to capture, a minimal number of key e-points have to be defined. For this test sequence, we have chosen a small set of e-points placed on the body near the joints: feet, knees, hips, hands, elbows, shoulders, neck, head, bust (figure 13). Since we use only 3 CCD cameras which acquires video sequences frontally, we cannot track the complete motion performed by the person. Other cameras would be required to get informations at the back side.
The result of the e-point tracking process is shown in figure 14. Like a conventional motion capture system, the results are trajectories off key points on the body, which can be used to reconstruct the motion performed by the person. To get the joints of the human body, offset values of the key e-points on the body have to be defined.

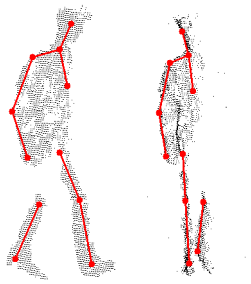

Fig. 13: Key e-points on the human body

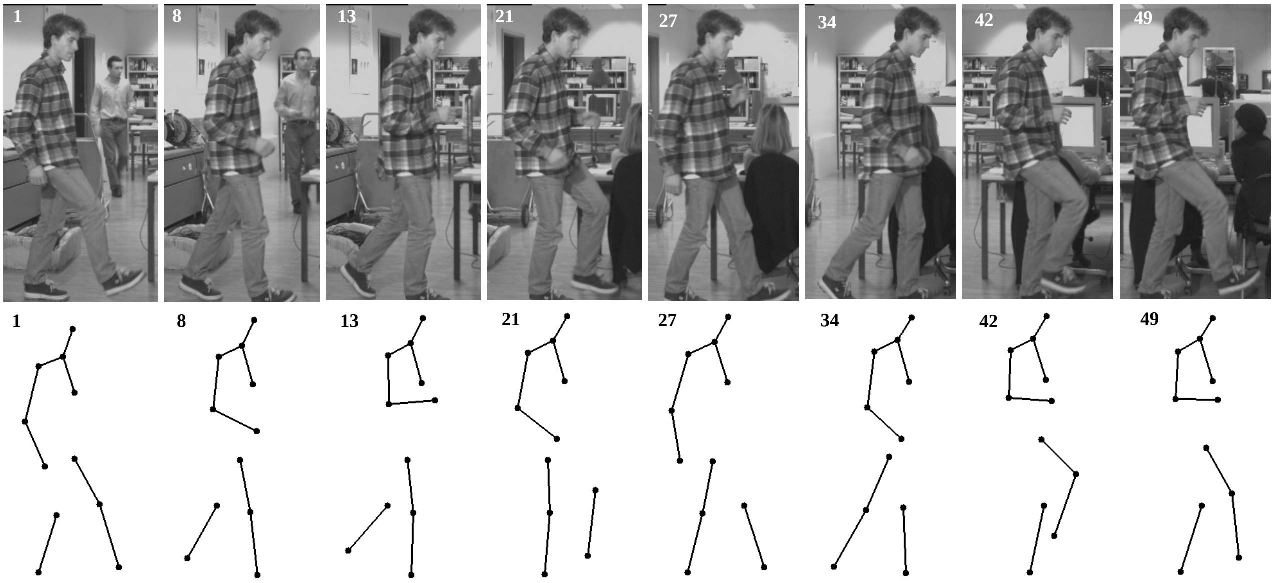

Fig. 14: Tracking key e-points of the human body, the results are 3-D trajectories of key points on the body

3. CONCLUSIONS AND FUTURE WORK

A process for an automated extraction of 3-D data from multi-image sequences has been presented. The extracted 3-D data is composed of two parts: measurement of the body surface at each time step of the sequence and a vector field of 3-D trajectories (position, velocity and acceleration). Adequate filters have been developed and applied to the data in order to reduce the noise. The extracted data are used to track in 3-D key points on the human body and the results are comparable with the conventional motion capture systems. The main advantage of our methods are: (1) its flexibility, a number of key points can be introduced and their position can be freely defined and (2) the non requirement of using markers.

A lot of work still remains for the future to improve the quality of the extracted 3-D data. For the surface measurement, the most important feature which has to be integrated in the process is the definition of geometric and neighborhood constraints in the least squares matching algorithm. The consideration of neighborhood information should be also integrated in the tracking process to achieve more reliable results.

In addition, the gain in robustness and level of automation should be also considered, since the final goal of the project is the development of a fully automated and robust process.

REFERENCES

Anai T., Chikatsu H., 2000. Dynamic Analysis of Human Motion Using Hybrid Video Theodolite. Int. Archives of Photogrammetry and Remote Sensing, Vol. 33, Part B5/1, Amsterdam, The Netherlands, pp. 25-29

Boulic R., Fua P., Herda L., Silaghi M., Monzani J.-S., Nedel L., Thalmann D., 1998. An Anatomic Human Body for Motion Capture. Proc. EMMSEC’98, Bordeaux, France

Brown D. C., 1971. Close-Range Camera Calibration. Photogrammetric Engineering and Remote Sensing, Vol. 37(8), pp. 855-866,

Cham T.-J. And Rehg J. M., 1999. A Multiple Hypothesis Approach to Figure Tracking. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA, Vol. 2, pp. 239-245

Cheung G. K. M. et al., 2000. A Real Time System for Robust 3D Voxel Reconstruction of Human Motions. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, Vol. II, pp. 714-720

D’Apuzzo N., 1998. Automated photogrammetric measurement of human faces. International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part B5, Hakodate, Japan, pp. 402-407

D’Apuzzo N., Plänkers R., Fua P., 2000: Least Squares Matching Tracking Algorithm for Human Body Modeling, International Archives of Photogrammetry and Remote Sensing, Vol. 33, Part B5/1, Amsterdam, The Netherlands, pp. 164-171

Delamarre Q. and Faugeras O., 1999. 3D Articulated Models and Multi-View Tracking with Silhouettes. Proc. of 7th IEEE International Conference on Computer Vision, Kerkyra, Greece, Vol. I, pp. 716-721

Deutscher J. et al., 2000. Articulated Body Motion Capture by Annealed Particle Filtering. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, Vol. II, pp. 126-133

Fua P., Gruen A., Plänkers R., D’Apuzzo N. and D. Thalmann, 1998. Human body modeling and motion analysis from video sequences. International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part. B5, Hakodate, Japan, pp. 866-873

Gruen A., 1985. Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography, Vol. 14(3), pp. 175-187

Gavrila D. M., Davis. L., 1996. 3-D model-based tracking of humans in action: a multi-view approach. Proc. Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, pp. 73-80

Iwai Y. et al., 1999. Posture Estimation using Structure and Motion Models. Proc. of 7th IEEE International Conference on Computer Vision, Kerkyra, Greece, Vol. I, pp. 214-219

Jojic N. et al., 1999. Tracking Self-Occluding Objects in Dense Disparity Maps. Proc. of 7th IEEE International Conference on Computer Vision, Kerkyra, Greece, Vol. I, pp. 123-130

Maas H.-G., 1998. Image sequence based automatic multi-camera system calibration techniques. International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part B5, Hakodate, Japan, pp. 763-768

Mahoney D. P., 2000. A New Track for Modeling Human Motion. Computer Graphics World, May 2000, pp. 18-20

Ohno H. and Yamamoto M., 1999. Gesture Recognition using Character Recognition Techniques on Two-Dimensional Eigenspace. Proc. of 7th IEEE International Conference on Computer Vision, Kerkyra, Greece, Vol. I, pp. 151-156

Rosales R. and Sclaroff S., 2000. Inferring Body Pose without Tracking Body Parts. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, Vol. II, pp. 721-727

Segawa H. and Totsuka T., 1999. Torque-based Recursive Filtering Approach to the Recovery of 3D Articulated Motion from Image Sequences. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA, Vol. 2, pp. 340-345

Song Y. et al., 2000. Towards Detection of Human Motion. Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, Vol. I, pp.810-817

WEB addresses (accessed at 1/nov/2000):

Analogus: http://www.analogus.com

Ascension Technology Corporation: http://www.ascension-tech.com

Northern Digital Inc.: http://www.ndigital.com

Polhemus: http://www.polhemus.com

Qualisys: http://www.qualisys.com

Vicon: http://www.vicon.com