LEAST SQUARES MATCHING TRACKING ALGORITHM FOR HUMAN BODY MODELING

Nicola D'APUZZO[*], Ralf PLAENKERS[**], Pascal FUA[**]

[*]Swiss Federal Institute of Technology, Zurich, Switzerland
Institute of Geodesy and Photogrammetry (IGP)
nicola@geod.baug.ethz.ch

[**]Swiss Federal Institute of Technology, Lausanne, Switzerland
Computer Graphics Lab (LIG)
Ralf.Plaenkers@epfl.ch, Pascal.Fua@epfl.ch

Commission V Special Interest Working Group on "Animation"

KEYWORDS: Object Tracking, Image Sequences, CCD, Modeling, Animation, 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. The 3-D information extracted from the video sequences can be used to reconstruct the animation model of the original sequence.

1 INTRODUCTION

The approach to human body modeling is usually split into two different cases: the static 3-D model of the body and the 3-D model of the motion. For pure animation purposes or definition of virtualized worlds, where the shape of the human body is first defined and then animated (Badler 2000, Badler et al. 1999, Boulic et al. 1997, Gravila et al. 1996), only an approximative measurement is required. An exact 3-D measurement of the body is instead required in medical applications (Bhatia et al. 1994, Commean et al. 1994, Yumei 1994) or in manufacturing of objects which have to be fitted to a specific person or group of persons; as for example in the space and aircraft industry for the design of seats and suits (McKenna 1999, Boeing Human Modeling System) or more generally in clothes or car industry (Certain et al. 1999, Bradtmiller et al. 1999, Jones et al. 1993, CyberDressForms). Recently, anthropometric databases have been defined (Pauqet et al. 1999, Robinette et al. 1999). Besides the shape information, they contain also other records of the person, which can be used for commercial or research purposes (McKenna 1999). In the last years, the demand for 3-D models of human bodies has drastically increased in all these applications. The currently used approaches for building such models are laser scanner (Daanen et al. 1997, Cyberware), structured light methods (Bhatia et al. 1994, Youmei 1994), infrared light scanner (Horiguchi 1998) and photogrammetry (Vedula et al. 1998). Laser scanners are quite standard in human body modeling, because of their simplicity in the use, the acquired expertise (Brunsman et al. 1997) and the related market of modeling software (Burnsides 1997). Structured light methods are well known and used for industrial measurement to capture the shape of parts of objects with high accuracy (Wolf 1996, GOM). The acquisition time of both laser scanner and structured light systems ranges from a couple of seconds to half minute. In case of human body

modeling, this can pose accuracy problems caused by the need for a person to remain immobile for several seconds. Photogrammetry methods (D'Apuzzo 1998) can instead acquire all data in less than one second.
The second component of the human body modeling process consists in the capture of the motion (Dyer et al. 1995). The different systems can be divided into groups depending on which characteristic is used for classification, e.g. accuracy, time to process the data, method used, price of the system, portability of the system. Photogrammetric systems measure very accurately the trajectories of signalized target points on the body (Boulic et al. 1998, Vicon, Qualisys, Northen Digital); some of them compute the data in real-time. Other systems use electromagnetic sensors which are connected to a computer unit which can process the data and produce 3-D data in real time (Ascension, Polhemus). There are also mechanical systems, where the person has to wear special suits with mechanical sensors which register the movement of the different articulations (Analogus). Motorized video theodolites in combination with a digital video camera have also been used for human motion analysis (Anai et al. 1999). A different approach is used by the image-based methods where image sequences are taken from different positions and then processed to recover the 3-D motion of the body (Gravila et al. 1996).
The common characteristic of these systems is the separated consideration of the two modeling aims: shape and motion are modeled in two different steps. In this paper, we present instead a method to solve the two problems simultaneously, recovering from one data set 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). It 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. Least squares matching has been chosen among others methods for its adaptivity.

2 EXTRACTION OF 3-D DATA FROM VIDEO SEQUENCES

In this section, we will first describe the system for data acquisition and the method used for its calibration. We then depict our methods for the extraction of 3-D data from the multi-image video sequence. The extracted information is of two different types: 3-D points clouds of the visible parts of the human body for each time step and a 3-D vector field of trajectories. The LSMTA can be also used in 2-D mode; we give an example of this possible use in tracking of facial expressions.

2.1 Data Acquisition and Calibration

Three synchronized CCD cameras in a linear arrangement (left, center, right) 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.

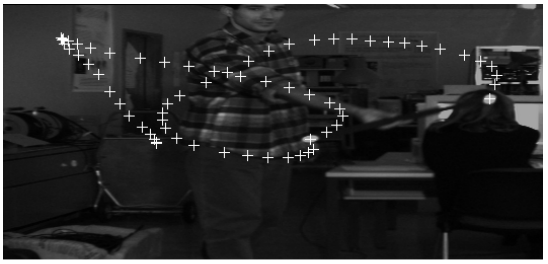

Figure 1. Automatically measured image coordinates of the two points on the reference bar

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 (Figure 1).
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 result of the calibration process are the exterior orientation of the three cameras (position and rotations: 6 parameters), parameters of the interior orientation of the cameras (camera constant, principle point, sensor size, pixel size: 7 parameters), 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

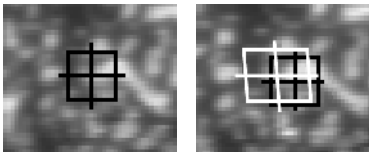

Figure 2. Least squares matching algorithm (LSM). Left: template image, right: search image

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 neighbourhood 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.

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 closest (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 a template image and the other two (left and right) are used as search images. The matcher searches the corresponding points in the two search images independently. 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).

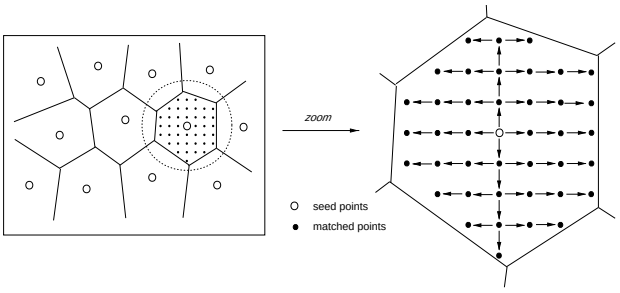

Figure 3: Search strategy for the establishment of correspondences between images

To evaluate the quality of the result, different indicators are used (resulted a posteriori standard deviation of the least squares adjustment, resulted 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 (level of texture in image, 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 neighbour, 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 appear in the set of matched points. The algorithm tries to close these holes by searching from all directions around. In case of poor natural texture, local contrast enhancement of the images is required for the least squares matching. Figure 4 shows the original images taken by the three cameras, the results after contrast enhancement and the matched points which result from the matching process.

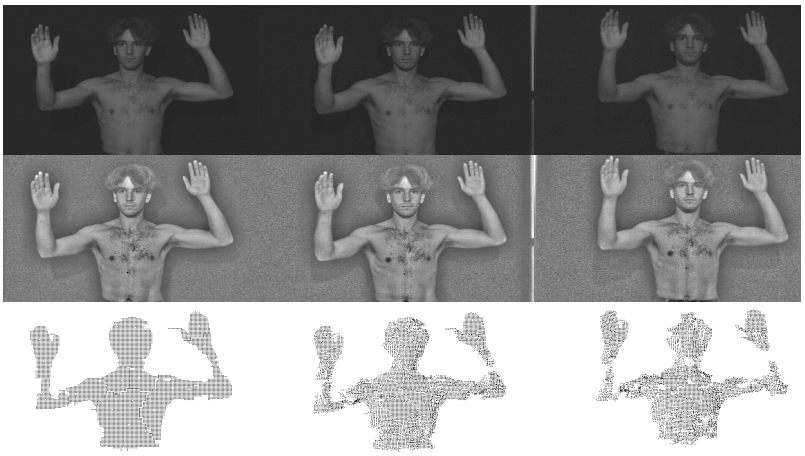

Figure 4. Original triplet (first row), enhanced images (second row) and matched points (third row); the first image on the left is the template

Before computing the 3-D coordinates of the matched points, the data pass through a neighborhood filter. It checks the data for neighbor similarity of the matched points comparing each point with the local mean values of the affine transformation parameters of the matching results. A matching process is repeated after filtering to measure the removed points.

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. The first filter was based on the matching results space, the second filter is instead applied to the 3-D data. It divides the object space in voxels (whose dimensions can vary) and the 3-D points contained in each voxel are replaced by its 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.

Figure 5. 3-D point cloud after passing filtering

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.

2.3 Tracking Process

2.3.1 Tracking single points. The basic idea of the tracking process is to track triplets of corresponding points through the sequence in the three images. Therefore, at the end of the process it is possible to compute their 3-D trajectories.

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

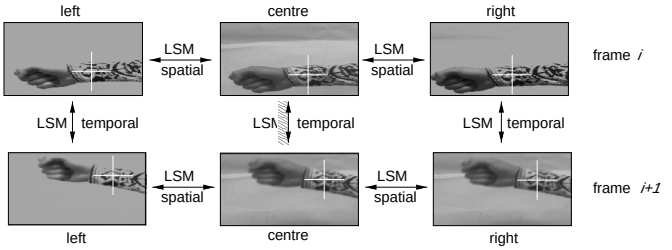

Figure 6. Temporal and spatial LSM

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 needed. 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.

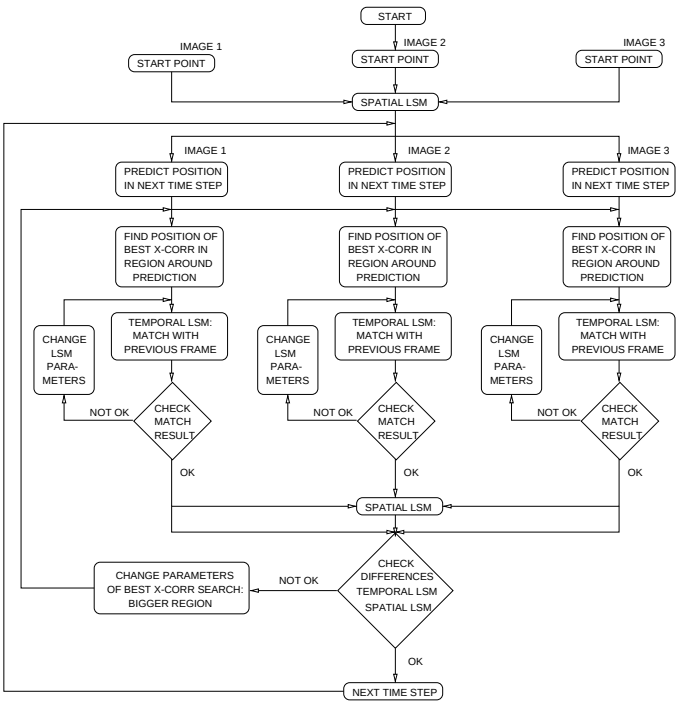

Figure 7. Flowchart of the LSM tracking process

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*). This position is considered an approximation of the exact position of the point to be tracked.

The least squares matching algorithm is then applied at that position and the result can be considered the exact position of the tracked point in the new frame (step *3*).
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 in 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. If instead 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, then the tracking process stops.

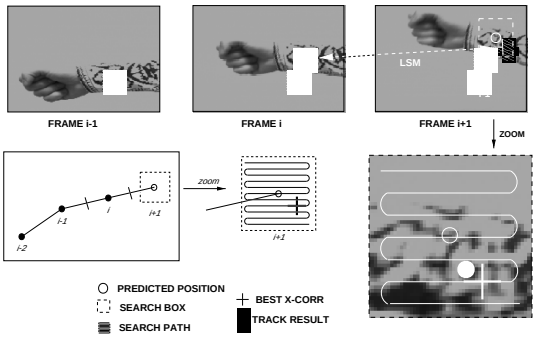

Figure 8. Tracking in image space: temporal LSM is applied at the position of best cross correlation

The result 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. 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. Therefore, filters can be defined to check these properties. The next paragraph describes the approach.

2.3.2 Surface tracking. Tracking surface parts means track simultaneously points belonging to a common surface. Practically, the tracking process depicted 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 get lost by occlusion and new parts of 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. Before proceeding to the next time step, the data resulting from the tracking process is checked for density (see flowchart in Figure 9). This operation is executed with a defined frequency (which can be for example every two frames). In the regions of low density (determined by a threshold), new points are integrated in the process, so that new appearing surface parts are also tracked. The new points come from data previously computed (surface measurement of the body for each frame).

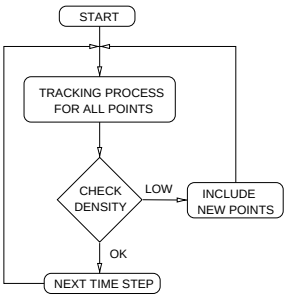

Figure 9. Flowchart of the surface tracking process

Figure 10 shows 6 frames of a walking sequence and the results of the surface tracking process.

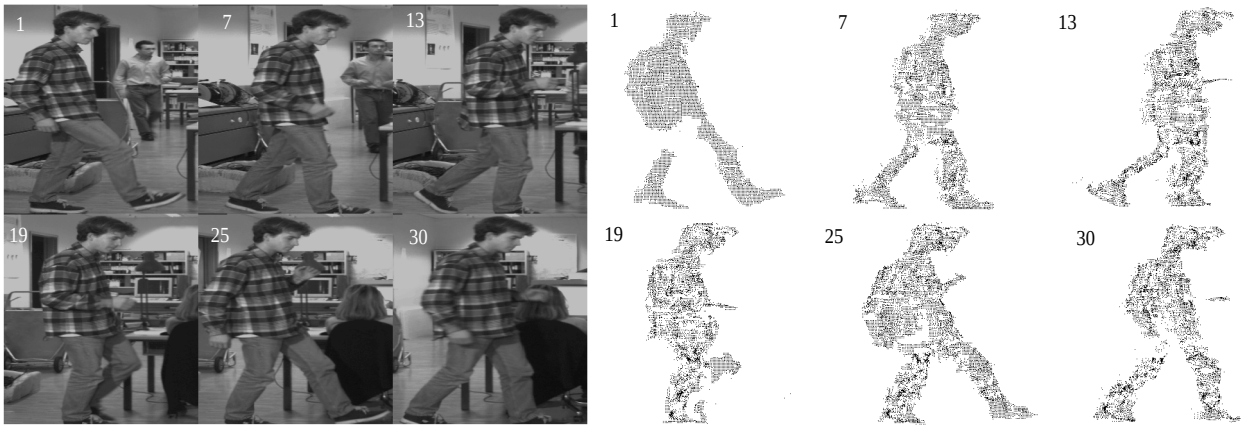

Figure 10. Left: 6 frames of a walk sequence (upper left to lower right)
Right: Tracked points displayed in image space for the 6 frames

As explained before, the tracking process can produce false trajectories. This is clearly shown in the Figure 11, where the computed 3-D trajectories for 30 frames are displayed (for the walking sequence of Figure 10). The vector field of trajectories (position, velocity and acceleration) can now 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 consists of thresholds for the velocity and acceleration (Figure 12, left). The second filter checks for the local uniformity of the motion, both in space and time (Figure 12, right). To check this property, the space is divided in voxels, for each voxel at each time step a mean value of the velocity vector is computed. The single trajectories are compared to local (in space and time) mean values of the velocity vector. If the differences are too large, the trajectory is considered to be false and it is truncated or removed.

As it can be seen comparing Figure 13 with Figure 11, the majority of the false trajectories are removed or truncated by the two filters. Still, some false trajectories remain in the data after filtering.

Figure 11. 3-D trajectories of the tracked points.
Left: frontal view, right: lateral view

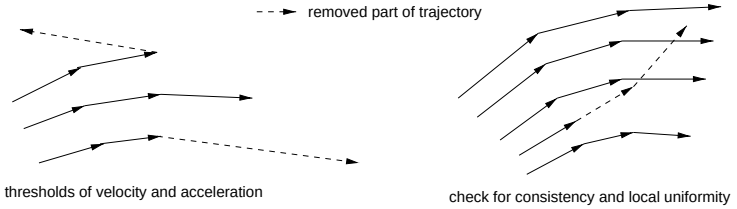

Figure 12. Filter to remove or truncate false trajectories.
Left: threshold filter, right: consistency and uniformity filter

Figure 13. 3-D trajectories after filtering.
Left: frontal view, right: lateral view

2.3.3 LSMTA in 2-D mode: The LSMTA is a flexible tool and can also be used in 2-D mode. In that case, the sequence of a single camera, e.g. a camcorder, is processed. The use of a single image sequence cannot obviously produce 3-D data but for some cases the 3-D information is not required. The Figure 14 shows a simple example of tracking facial expressions, where some key points are tracked through the sequence. The images were indeed acquired with a video camcorder. This example underlines the flexibility of the LSMTA which can produce in this case simple animation, tracking key points on the face without using markers.

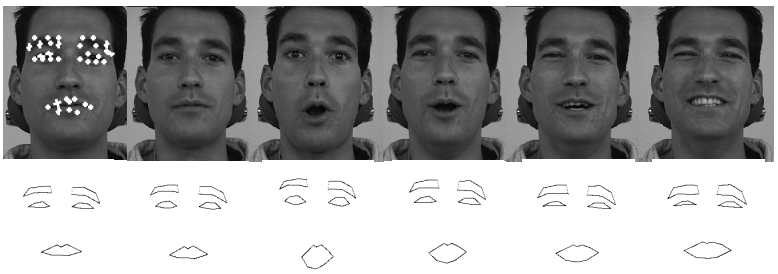

Figure 14. Some frames of a single camera image sequence
(the crosses on the first frame show the tracked points).
Bottom: basic animation created joining the tracked points with lines

3 USE OF 3-D DATA FOR HUMAN BODY MODELING

From the multi-image sequence, the process described extracts data in form of a 3-D point cloud of the visible body surface at each time step and a vector field of 3-D trajectories. Figure 16 shows the results achieved by a 2-D contour tracking algorithm using the 3-D trajectories. The algorithm is based on the snake technique (Kass et al. 1988). Given an extracted contour in one frame, the trajectory information of surrounding 3-D points, projected onto the image plane, is used to predict the position of the contour in the next frame. The silhouette information and the measured 3-D points for each frame are used to fit a complete animation model to the data. The results of the fitting process are shown in Figure 17. For the detailed explanation of the process we refer to the related publication (Plaenkers et al. 1999).

Figure 16. Results of the silhouette tracking process

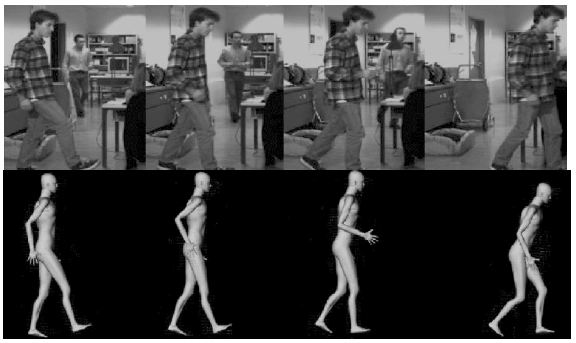

Figure 17. Results of the fitting process

4 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). Initially, the two different types of data are very noisy, therefore adequate filters have been developed and applied to the data.

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.

ACKNOWLEDGEMENTS

The work reported here was funded in part by the Swiss National Science Foundation.

REFERENCES

Anai T., Chikatsu H., 1999. Development of Human Motion Analysis and Visualisation System. Int. Archives of Photogrammetry and Remote Sensing, Onuma, Japan, 32 (5-3W12), pp. 141-144

Badler N. I., 2000. Animation 2000++. IEEE Computer Graphics and Applications, 20(1), pp. 28-29

Badler N. I., Palmer O., Bindiganavale R., 1999. Animation Control for Real Time Virtual Humans. Comm. ACM, 42 (8), pp. 65-73

Bhatia G., Smith K. E., Commean P. K., Whitestone J., Vannier M. W., 1994. Design of a Multisensor Optical Surface Scanner. Sensor Fusion VII, SPIE Proc., Boston, USA, Vol. 2355, pp. 262-273

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

Boulic R., Thalmann D., Molet T., Huang Z., 1997. An animation interface designed for motion capture. Computer Animation '97 Conference, IEEE Press, pp. 77-85

Bradtmiller B., Gross M., 1999. 3D Whole Body Scans: Measurement Extraction Software Validation. "Digital Human Modeling for Design and Engineering" Conference and Exposition, May 1999, Hague, The Netherlands

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

Brunsman M. A., Daanen H., Robinette K. M., 1997. Optimal Postures and Positioning for Human Body Scanning. Proc. of International Conference on Recent Advances in 3-D Digital Imaging and Modeling, IEEE Computer Society Press, Los Alamitos, CA, USA

Burnsides D. B., 1997. Software for Visualization, Analysis and Manipulation of Laser Scan Images. Paper #3023-15, ASC-97-0378, Internal Report Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, USA

Certain A., Stuetzle W., 1999. Automatic Body Measurement for Mass Customisation of Garments. Proc. of Second Int. Conference on 3-D Digital Imaging and Modeling, Ottawa, Canada, pp. 405-412

Commean P., Smith K., Vannier M., 1994. Automated Limb Prothesis Design. Visualisation in Biomedical Computing, Vol. 19, pp. 493-503

Daanen H., Brunsman M. A., Taylor S. E., Nurre J. H., 1997: Accuracy problems in whole body scanning. IS&T/SPIE's Symposium of Electronic Imaging, Vol. 3023.

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

Dyer S., Martin J., Zulauf J., 1995. Motion Capture White Paper.
http://reality.sgi.com/jam_sb/mocap/MoCapWP_v2.0.html (31 March 2000)

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, Hakodate, Japan, 32(B5), pp. 866-873

Gruen A., 1985. Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography, 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

Horiguchi C., 1998. BL (Body Line) Scanner. The development of a new 3D measurement and Recostruction system. Int. Archives of Photogrammetry and Remote Sensing, Hakodate, Japan, 32(B5), pp. 421-429

Jones P. R. M., West G., Brooke-Wavell K., 1993. Interrogation of 3D Body Data for Applications in Manufacturing Industries. Application of Computers to Manufacturing Engineering, Directorate of the Science and Engineering Research Council, Research Conference Proceedings, Sheffield University, USA, pp. 20-25

Kass M., Witkin A., Terzopoulos D., 1988. Snakes: Active Contour Models. IJCV, 1(4), pp. 321-331

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

McKenna P., 1999. Measuring Up. Air Force Airman Magazine
http://www.af.mil/news/airman/0296/measurin.htm (31 March 2000)

Paquet E., Rioux M., 1999. Neferiti: a query by content software for three-dimensional models database management. Image and Vision Computing 17(2), pp. 157-166

Plaenkers R., D'Apuzzo N. 1999. Automated Body Modeling from Video Sequences. Proc. IEEE International Workshop on Modelling People (mPeople), Corfu, Greece, September

Robinette K. M., Daanen H., 1999. The Caesar Project: A 3-D Surface Anthropometry Survey. Proc. of Second Int. Conference on 3-D Digital Imaging and Modeling, Ottawa, Canada, pp. 380-386

Vedula S., Rander P., Saito H., Kanade T., 1998. Modeling, Combining and Rendering Dynamic Real-World Events from Image Sequences. Proc. of 4th Conference on Virtual Systems and Multimedia, Gifu, Japan, Vol. 1, pp. 326-332

Wolf H. G. E., 1996. Structured Lighting for Upgrading 2D-Vision systems to 3D. Proc. of Int. Symposium on Laser, Optics and Vision for Productivity and Manufacturing I, Besancon, France, pp. 10-14

Youmei G., 1994. Non-contact 3D biological shape measurement from multiple views. M.Sc. Thesis, University of Western Australia

WEB addresses:

Analogus: http://www.analogus.com (31 March 2000)
Ascension Technology Corporation: http://www.ascension-tech.com (31 March 2000)
Boeing Human Modeling System: http://www.boeing.com/assocproducts/hms (31 March 2000)
CyberDressForms Inc.: http://www.cyberdressforms.com (31 March 2000)
Cyberware: http://www.cyberware.com (31 March 2000)
GOM Optical Measurement Techniques: http://www.gom.com (31 March 2000)
Northern Digital Inc.: http://www.ndigital.com (31 March 2000)
Polhemus: http://www.polhemus.com (31 March 2000)
Qualisys: http://www.qualisys.com (31 March 2000)
Vicon: http://www.vicon.com (31 March 2000)