Human Body Motion Capture from Multi-Image Video Sequences

Nicola D'Apuzzo
Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland

ABSTRACT

In this paper is presented a method to capture the motion of the human body from multi image video sequences without using markers. The process is composed of five steps: acquisition of video sequences, calibration of the system, surface measurement of the human body for each frame, 3-D surface tracking and tracking of key points. The image acquisition system is currently composed of three synchronized progressive scan CCD cameras and a frame grabber which acquires a sequence of triplet images. Self calibration methods are applied to gain exterior orientation of the cameras, the parameters of internal orientation and the parameters modeling the lens distortion. From the video sequences, two kinds of 3-D information are extracted: a three-dimensional surface measurement of the visible parts of the body for each triplet and 3-D trajectories of points on the body. The approach for surface measurement is based on multi-image matching, using the adaptive least squares method. A full automatic matching process determines a dense set of corresponding points in the triplets. The 3-D coordinates of the matched points are then computed by forward ray intersection using the orientation and calibration data of the cameras. The tracking process is also based on least squares matching techniques. Its basic idea is to track triplets of corresponding points in the three images through the sequence and compute their 3-D trajectories. The spatial correspondences between the three images at the same time and the temporal correspondences between subsequent frames are determined with a least squares matching algorithm. 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 at each time step by forward ray intersection. Velocities and accelerations are also computed. The advantage of this tracking process is twofold: it can track natural points, without using markers; and it can track local surfaces 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). The last step of the process is the definition of selected key points of the human body. A key point is a 3-D region defined in the vector field of trajectories, whose size can vary and whose position is defined by its center of gravity. The key points are tracked in a simple way: the position at the next time step is established by the mean value of the displacement of all the trajectories inside its region. The tracked key points lead to a final result comparable to the conventional motion capture systems: 3-D trajectories of key points which can be afterwards analyzed and used for animation or medical purposes.

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

1. INTRODUCTION

Motion capture systems are mainly used for two applications: in computer animation to increase the level of realism digitizing the desired movements performed by an actor and in biomechanics to precisely measure the movement of joints. The motion capture systems 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 and portability of the system.
The magnetic systems (e.g., Ascension^{TM}, Polhemus^{TM}) use electromagnetic sensors connected to a computer unit which can process the data and produce 3-D data in real time. 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. Wireless systems have also been developed to solve the disadvantage of restricted freedom of

In: El-Hakim, S.F., Gruen, A., Walton, J.S. (Eds.), Videometrics VIII, Proc. of SPIE, Vol. 5013, Santa Clara, USA, 2003.

movement caused by the cabling.
Optical systems (e.g., Motion Analysis™, Vicon™, Qualisys™) are mostly based on photogrammetric methods where the trajectories of signalized target points on the body are measured very accurately.[1,2] They offer complete freedom of movement and interaction of different actors is also possible. In the last years, several improvements have been introduced, such as the use of smart cameras and CMOS sensors to achieve real-time 3–D data acquisition.
Electro-Mechanical systems (e.g., Analogus™) have recently appeared in the market: in this case the person has to wear a special suit with integrated electro-mechanical sensors that register the motion of the different articulations. This method also has the advantage of real-time data transfer from the sensors to the computer without processing; moreover, it is less expensive.
Motion capture can also be achieved by image-based methods. They can be divided into monocular and multi-image systems. Monocular systems use sequences of images acquired by a single camera. To gain three-dimensional information from single video clips, knowledge of human motion must be used. Some systems learn from provided sample training data and apply statistical methods to get the performed 3–D motion.[3–5] Other systems perform the tracking of defined human body models with constraints by sophisticated filtering processes.[6–8]
Multi-image systems use sequences of images acquired simultaneously by two or more cameras. Some systems assume very simple 3-D human models (e.g., articulated objects made of cylinders) whose characteristic sizes and joint angles are determined by comparing the projections of the model into the different images with the extracted silhouettes of the moving person[9,10] or the extracted edges.[11,12] Other systems use image based tracking algorithms to track three-dimensionally different body parts.[13] Mathematical models of human motion can also be used to track directly in 3-D data, which can be trajectories of known key points[14] or dense disparity maps.[15] More sophisticated methods fit generic human body models to extracted 3–D data from video sequences.[16–19]
In this paper, a method to recover from video sequences both 3-D shape and 3-D motion information is presented. The core of the method is the least squares matching tracking algorithm (*LSMTA*) which uses the least squares matching process to establish correspondences between subsequent frames of the same view as well as correspondences between the images of the different views. LSMTA is a process composed of five steps: acquisition of multi-image video sequences, calibration of the system, surface measurement for each frame, surface tracking and tracking key-points.

2. DATA ACQUISITION AND CALIBRATION

The term *multi-image* refers to multiple images acquired from different positions in the space describing the same scene and *multi-image sequence* refer to multi-images acquired during a time interval. To record a scene with movement, the multiple images have to be acquired simultaneously. The precision of the synchronization of the multiple imaging devices plays an essential role for the accuracy potential of the measurement achieved using the images.
Various methods can be used to acquire multi-image sequences. For the sequence presented in this paper, three synchronized progressive scan CCD cameras in a triangular arrangement are used. A sequence of triplet images is acquired with a frame grabber and the images are stored with 640x480 pixels at 8 bit quantization. Figure 1 shows a frame of the acquired sequence. In order to increase the contrast, the images of this sequence are enhanced using Wallis filtering,[20] the result is shown on the right of the figure.

Figure 1. A frame of the acquired sequence. Left: original image, right: after Wallis filter contrast enhancement.

To calibrate the system, the reference bar method[21] 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: six parameters), the parameters of the interior orientation of the cameras (camera constant, principle point, sensor size, pixel size: seven parameters), the parameters for the radial and decentering distortion of the lenses and optic systems (five parameters) and two additional parameters modeling other effects as differential scaling and shearing.[22,23] A thorough determination of these parameters modeling distortions and other effects is required to achieve high accuracy.

3. SURFACE MEASUREMENT

The approach for surface measurement is based on multi-image photogrammetry using images acquired simultaneously. A dense set of corresponding points in the images is established by a matching process and the 3–D coordinates of the matched points are computed. The measured surface results therefore as a 3–D point cloud. The multi-image matching process[24] is based on the adaptive least squares method[25] with the additional geometrical constraint of the matched point to lie on the epipolar line. It produces automatically a dense and robust set of corresponding points starting from few seed points. Depending on the case, the seed points may be generated fully automatically or selected manually in one image. The full automatic mode is useful for dynamic surface measurement and tracking processes, where the number of multi-image sets to be processed can be very large. In this case, Foerstner interest operator[26] is applied on the template image to determine automatically marking points where the matching process may perform robustly; the corresponding points in the other images are then established by searching for the best matching results along the epipolar line.
After the definition of the seed points, the template image is divided into polygonal regions according to which of the seed point is the closest (Voronoi tessellation). Starting from the seed points, the automatic matcher produces a dense set of corresponding points in each region by sequential horizontal and vertical shifts (see figure 2).

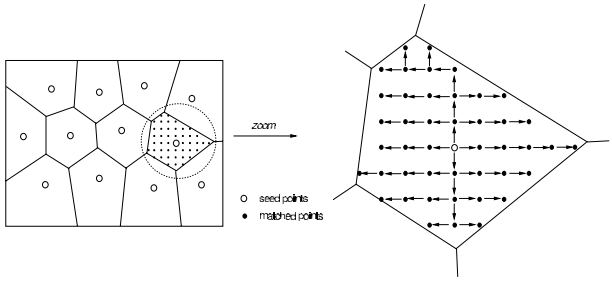

Figure 2. Search strategy for the matching process. Left: Voronoi tessellation. Right: starting from the seed points, each region is covered by sequential horizontal and vertical shifts.

The process works adaptively at each shift, changing some parameters (e.g. smaller shift, bigger size of the patch) if the quality of the match is not satisfactory. Several indicators are used to define the match quality: a posteriori standard deviation of the least squares adjustment, standard deviation in x and y directions, displacement from the start position in x and y direction and distance to the epipolar lines. Thresholds for these values are defined according to the texture and the type of the images.
The coverage of the entire image is achieved by repeating the process for all the polygonal seed points regions. At the end of the process, holes of not analyzed areas can occur in the set of matched points. The process tries to match the missing points by searching from all directions around the holes.
With the proposed strategy, the time required by the matching process is short; to give an example, on a Pentium III 600 MHz machine, about 25,000 points are matched in a triplet in less than 10 minutes.
The 3–D coordinates of the matched points are then computed by forward ray intersection using the orientation

and calibration data of the cameras.
The result of the surface measurement process are 3–D point clouds for each time step; this data will then be used in the successive surface tracking process. Figure 3 shows an example of the matching results achieved on an image triplet of the sequence and the computed 3–D point cloud.

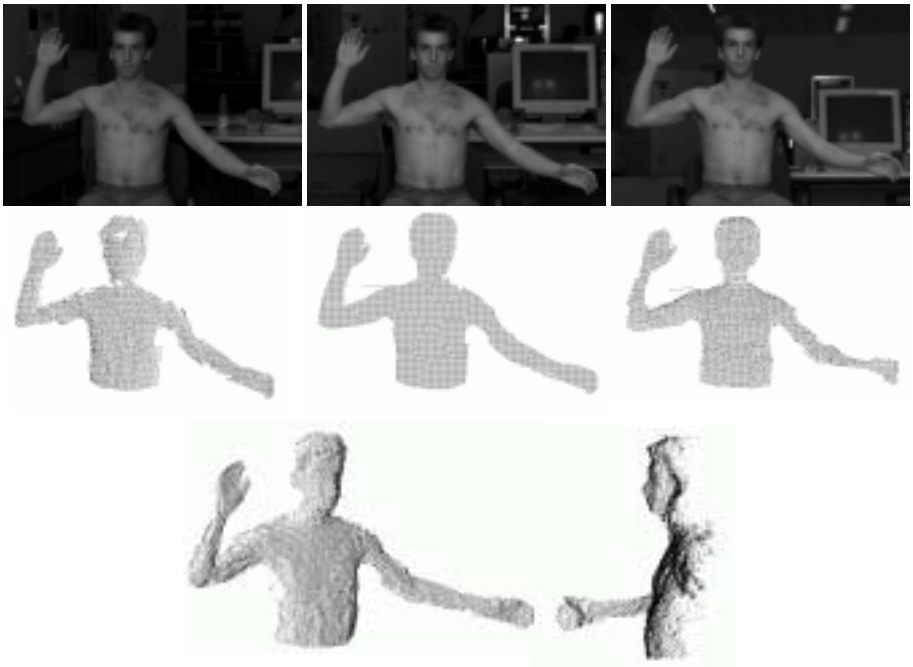

Figure 3. Matching process on a triplet. Top: image triplet. Center: matched points. Bottom: computed 3–D point cloud.

4. SURFACE TRACKING

4.1. Least Squares Matching Tracking Algorithm

The basic idea is to track corresponding points in the multi-images through the sequence and compute their 3–D trajectories. The process is based on least squares matching techniques (*LSM*): the *spatial* correspondences between the images of the different views and the *temporal* correspondences between subsequent frames are computed using the least squares matching algorithm (see figure 4).

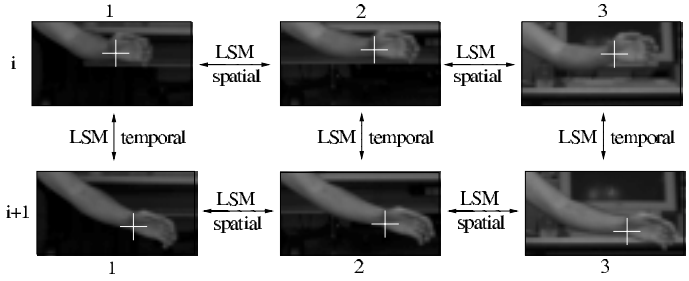

Figure 4. LSM tracking: temporal and spatial correspondences are established with LSM.

The process starts matching corresponding points in the multi-images and continues with the tracking loop: (1) predict the position in the next frame, (2) search the position with the best cross correlation value and (3) establish the point in the next frames using least squares matching (*temporal LSM*).

The position of the tracked point at time *i+2* is linearly predicted from the two previous times *i+1* and *i* (step 1). A search box is defined around this predicted position in the frame at time *i+2* and is scanned for the position which has the best value of cross correlation between the image of frame at time *i+1* and the image of frame at time *i+2* (step 2). The least squares matching algorithm is applied at that position and the result can be considered the exact position of the tracked point in the new frame (step 3). Figure 5 shows graphically the tracking process in image space.

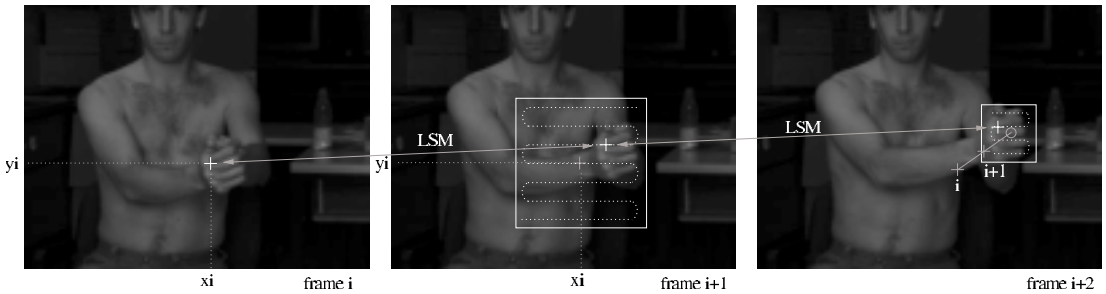

Figure 5. Tracking in image space: at the beginning of the process (when the prediction is not yet possible) the location in the next time step is searched in a larger area centered in the location of the previous frame (image coordinates (xi, yi)). Once the point is tracked at least in two frames, the prediction is possible and the search area is therefore smaller and centered in the predicted position (circle in frame *i+2*).

This process is performed in parallel for the different images. To test the individual results, *spatial LSM* is computed at the positions resulting from the *temporal LSM*s 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 step (2) of the process is repeated by searching the value of best cross correlation in a bigger region around the predicted position. If the new result is also rejected, the tracking process stops.
The result of the multi-image tracking process applied on a single point are its coordinates in the multi-images through the sequences, thus the 3–D coordinates of the point for each time step can be computed by forward ray intersection resulting in its 3–D trajectory. Its velocity and acceleration can then also be determined for each time step.
The proposed process can be used to track well defined points on the human body surface. Trajectories of single points are however not sufficient to understand and record the motion and movement of a human or the changes of a human body surface part. The next two sections propose a method to extract more complex information from the image sequence.

4.2. Tracking surface parts

Requirement for the surface tracking process is the measurement of the surface parts of interest for each multi-image of the sequence. The tracking process of section 4.1 is then applied to all the points measured on the surface, resulting in a vector field of trajectories.
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 the surface can appear. For this reason, a new functionality has to be integrated in the tracking loop: before passing to the next time step, the density of the data resulted from the tracking process is checked with a threshold. In the regions of low density, new points are imported from the previously computed data (surface measurement for each time step). In this way, new appearing surface parts or lost points are integrated in the tracking process.
The results of the surface tracking process is a dense set of trajectories, figure 6 shows an example.
In case of poor texture of the surface, the tracking process can produce false trajectories, that can easily be recognized because they do not follow the common movement of the majority. The vector field of trajectories can indeed be checked for local uniformity of the movement. Two filters are applied to remove or truncate false trajectories. The first one removes larger errors using thresholds of the velocity and acceleration. Depending on the movement performed, the two thresholds are defined at the begin of the process and remain constant during the sequence. The second filter checks for the local uniformity of the movement both in space and time.

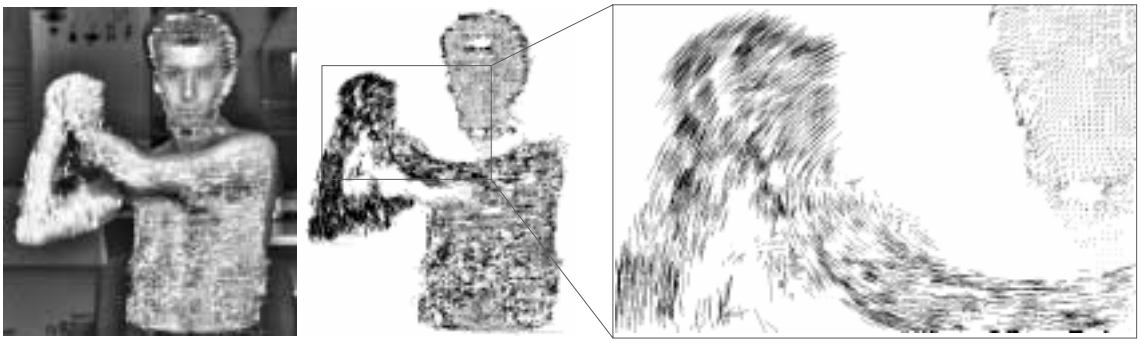

Figure 6. Surface tracking results. The trajectories are drawn as displacement from the previous time step. Left: image and tracked points, right: tracked points and detail.

To check this property, the space is divided in regular voxels and local mean values of the velocity vectors are computed in each voxel for each time step. The single trajectories are then compared to the local mean values and truncated or removed if the differences in magnitude and direction are too large.
After filtering the trajectories, a problem can still remain. Ideal trajectories start from the beginning of the sequence and last till the end. However, in some cases, depending on the quality of the image sequence and on the type of surface, the result of the tracking process is a set of broken trajectories with a varying length. This effect, particularly strong when measuring full human movements, is mainly caused by occlusion and lack of texture. To solve this problem, the concept of a *key-point* is introduced.

4.3. Tracking key-points

A key-point is a 3–D region defined in the vector field of trajectories, whose size can vary and whose position is defined by its center of gravity. Key-points are interactively defined in a graphical user interface. They can be easily placed and moved in 3–D space. The key-points 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 its region. The size of the region to be tracked acts an important role: if a small size is chosen, the key-point can be considered with good approximation to lie on the surface and to represent fictitious markers on the human body. If, on the other hand, larger sizes are chosen, the key-point represents a more approximative position of e.g. joints of the human body.
For the sequence presented in this paper, the key-points are placed in such a way that their trajectories can describe the complex movement performed by the person. A small set is placed near the joints: hands, elbows, shoulders, sides of neck; three are placed on the bottom of the upper body part; three on the face (nose and eyes) and four on the head (see figure 7).

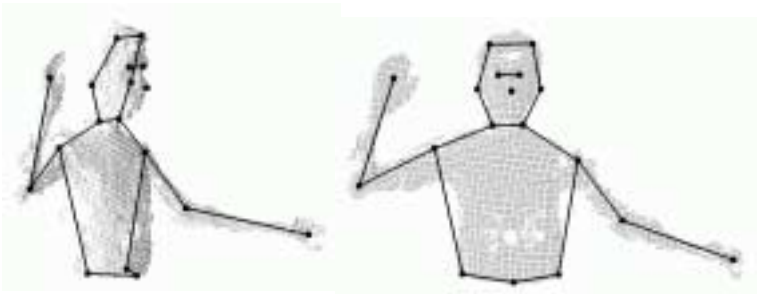

Figure 7. Example of key-points tracked on the human body.

Because of the complexity of the performed movement of this sequence, the size of the region defining the key-points is chosen large enough to guarantee robustness. The key-points represent in this case an approximation for the joint trajectories. The final result of the tracking process is shown in figure 8 together with the corresponding

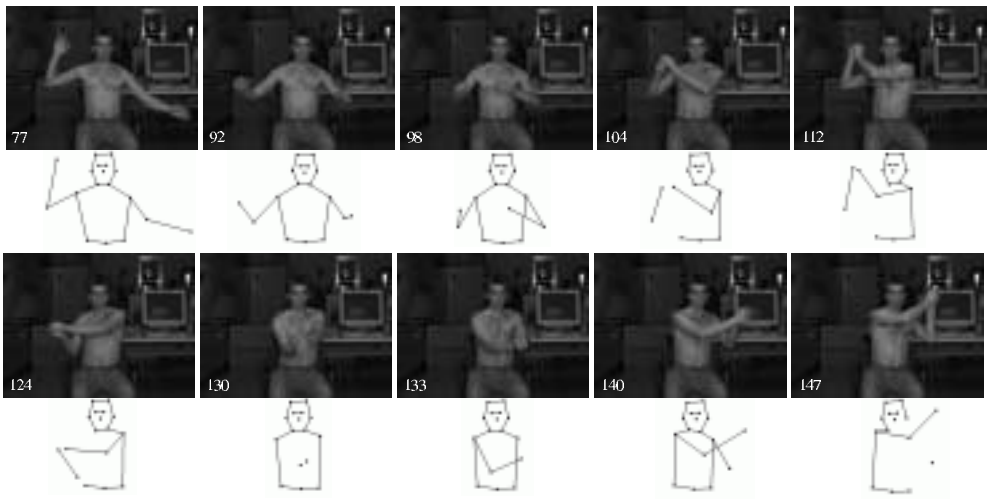

Figure 8. Tracking key-points: some frames of the sequence and frontal view of the position of the tracked key-points.

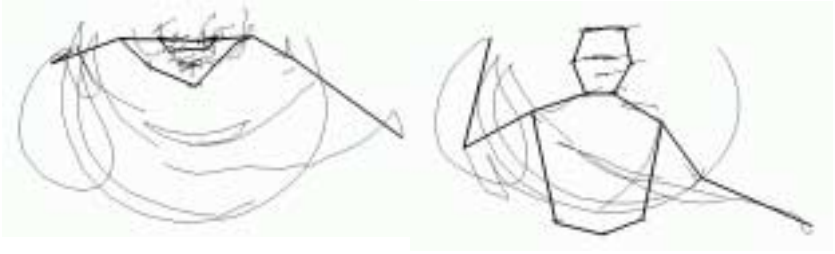

Figure 9. View from the top (left) and from the front (right) of the 3–D trajectories of the key-points.

frames; two view of the 3–D trajectories of the key-points are displayed in figure 9 (note that the connections between key-points are intended only for a better understanding of the images).
As can be seen, e.g., in the frame number 130 of figure 8, some key-point cannot be tracked during the entire sequence because of strong occlusions caused by the complex movement. A possible solution to this problem is the use of more than three cameras, acquiring multi-images all around the person.

5. CONCLUSIONS

A process for an automated extraction of 3–D data from multi-image sequences has been presented. The gained 3–D data can be of two different types: surface measurement of human body parts at each time step of the sequence or surface tracking in form of a vector field of 3–D trajectories (position, velocity and acceleration).
The dynamic surface measurement and tracking procedure are part of a pilot project still under development. Its final goal is a fully automatic system to model most realistically human bodies from video sequences. Work still remains for the future to increase the automation level and to improve the quality of the extracted 3-D data. Moreover, the use of only three CCD cameras limits the measurement and tracking to only one side of the human body. For very complex movements, the body parts have to be imaged both in front as well as sideways and from the back, therefore the acquisition system has to be extended. Other important cues that can be determined from the vector field of trajectories resulted from the tracking process are all the rotations angles, which are required to describe precisely the movement of articulations such as shoulder-elbow-hand.

ACKNOWLEDGMENTS

The work reported here was funded in part by the Swiss National Science Foundation. I am grateful to the Computer Graphic Lab of the EPF Lausanne for providing me with the multi-images sequences.

REFERENCES

1. M. Tsuruoka, R. Shibasaki, E. Box, and S. Murai, "Biomechanical 3-D analysis of a human sit-to-standing sequence using two CCD video cameras," *International Archives of Photogrammetry and Remote Sensing* **30**(5W1), 1995.
2. R. Boulic, P. Fua, L. Herda, M. Silaghi, J.-S. Monzani, L. Nedel, and D. Thalmann, "An anatomic human for motion capture," in *Proc. of EMMSEC'98*, (Bordeaux, France), 1998.
3. D. Mahoney, "A new track for modeling human motion," *Computer Graphics World* , pp. 18–20, May 2000.
4. R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 721–727, (South Carolina, USA), 2000.
5. Y. Song, X. Feng, and P. Perona, "Towards detection of human motion," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **1**, pp. 810–817, (South Carolina, USA), 2000.
6. T.-J. Cham and J. Rehg, "A multiple hypothesis approach to figure tracking," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 239–245, (Fort Collins, USA), 1999.
7. H. Segawa and T. Totsuka, "Torque-based recursive filtering approach to the recovery of 3D articulated motion from image sequences," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 340–435, (Fort Collis, USA), 1999.
8. J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 126–133, (South Carolina, USA), 2000.
9. Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with silhouettes," in *Proc. of the 7th IEEE Int. Conf. on Computer Vision*, **1**, pp. 716–721, (Kerkyra, Greece), 1999.
10. G. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 714–720, (South Carolina, USA), 2000.
11. W. Kinzel and R. Behring, "Initializing the recognition of moving persons," *International Archives of Photogrammetry and Remote Sensing* **30**(5W1), 1995.
12. D. Gravila and L. Davis, "3-D model based tracking of humans in action: a multi-view approach," in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 73–80, (San Franciso, USA), 1996.
13. H. Ohno and M. Yamamoto, "Gesture recognition using character recognition techniques on two-dimensional eigenspace," in *Proc. of 7th IEEE Conf. on Computer Vision*, **1**, pp. 151–156, (Kerkyra, Greece), 1999.
14. Y. Iwai, K. Ogaki, and M. Yachida, "Posture estimation using structure and motion models," in *Proc. of 7th IEEE Int. Conf. on Computer Vision*, **1**, pp. 214–219, (Kerkyra, Greece), 1999.
15. N. Jojic, M. Turk, and T. Huang, "Tracking self-occluding objects in dense disparity maps," in *Proc. of 7th IEEE Int. Conf. on Computer Vision*, **1**, pp. 123–130, (Kerkyra, Greece), 1999.
16. P. Fua, L. Herda, R. Pläkers, and R. Boulic, "Human shape and motion recovery using animation models," *International Archives of Photogrammetry and Remote Sensing* **33**(B5), pp. 253–268, 2000.
17. L. Herda, P. Fua, R. Plaenkers, R. Boulic, and D. Thalmann, "Skeleton-based motion capture for robust reconstruction of human motion," in *Proc. of Computer Animation 2000*, IEEE CS Press, 2000.
18. R. Plänkers, "Tracking and modeling people in video sequences," *Int. Journal of Computer Vision and Image Understanding* **81**(3), pp. 285–302, 2001.
19. P. Fua, A. Gruen, N. D'Apuzzo, and R. Plänkers, "Markerless full body shape and motion capture from video sequences," *International Archives of Photgrammetry and Remote Sensing* **34**(B5), pp. 256–261, 2002.
20. R. Wallis, "An approach to the space variant restoration and enhancement of images," in *Proc. of Symposium on Current Mathematical Problems in Image Science*, pp. 329–340, (Naval Postgraduate School, Monterey, USA), 1976.
21. H.-G. Maas, "Image sequence based automatic multi-camera system calibration techniques," *International Archives of Photogrammetry and Remote Sensing* **32**(B5), pp. 763–768, 1998.
22. H. Beyer, *Geometric and radiometric analysis of a CCD-camera based photogrammetric close-range system*. PhD thesis, Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland, 1992.
23. D. Brown, "Close-range camera calibration," *Photogrammetric Engineering and Remote Sensing* **37**(8), pp. 855–866, 1971.
24. N. D'Apuzzo, "Measurement and modeling of human faces from multi images," *International Archives of Photgrammetry and Remote Sensing* **34**(B5), pp. 241–246, 2002.
25. A. Gruen, "Adaptive least squares correlation: a powerful image matching technique," *South African Journal of Photogrammetry, Remote Sensing and Cartography* **14**(3), pp. 175–187, 1985.
26. W. Foerstner and E. Guelch, "A fast operator for detection and precise location of distincs points, corners and center of circular features," in *Proc. of ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pp. 281–305, (Interlaken, Switzerland), 1987. June 2-4.