Markerless Motion Capture from Single or Multi-Camera Video Sequence

F. Remondino
IGP - ETH Zurich, CH
fabio@geod.baug.ethz.ch
www.photogrammetry.ethz..ch

N. D'Apuzzo
Homometrica Consulting
nda@homometrica.ch
www.homometrica.ch

G. Schrotter, A. Roditakis
IGP - ETH Zurich, CH
gerhard@geod.baug.ethz.ch
roditak@geod.baug.ethz.ch

Abstract

The modeling of human characters consists of two distinguished processes, namely the definition of 3D shape and 3D movement of the body. To achieve realism, both shape and motion information should be acquired from real persons. The two procedures are usually considered separately, by using full body scanners and motion capture systems. In this paper, we report about deterministic methods developed at IGP to perform the two modeling processes by using a unique set of data. Monocular or multi-camera videogrammetry in fact allow for human body modeling and motion reconstruction at the same time. The two reliable and accurate photogrammetric procedures are introduced and examples are presented.

Keywords: photogrammetry, matching, capturing techniques, virtual human

1. Introduction

The realistic modeling of human characters from video sequences is a challenging problem that has been investigated a lot in the last decade. Recently the demand of 3D human models is drastically increased for applications like movies, video games, ergonomic, e-commerce, virtual environments and medicine. A complete human model consists of both 3D shape and movements of the body: most of the available systems consider these two modeling procedures as separate even if they are very closed. A standard approach to capture the static 3D shape (and colour) of an entire human body uses laser scanner technology [1][2]: it is quite expensive but it can generate a whole body model in about 20 seconds. Afterwards the 3D shape can be animated (articulating the model and changing its posture) [3][4][5] or dressed for garment models [6][7]. On the other hand, precise information related to character movements is generally acquired with motion capture techniques: they involve electro-magnetic sensors [8] or a network of cameras [9][10] and prove an effective and successfully mean to replicate human movements. In between, single- or multi-stations videogrammetry offers an attractive alternative technique, requiring cheap sensors, allowing markerless tracking and providing, at the same time, for 3D shapes and movements information. Model-based approaches are very common, in particular with monocular video streams [11][12], while deterministic approaches are almost neglected, often due to the difficulties in recovering the camera parameters and because of human limbs occlusions. Generally computer vision techniques, image cues, background segmentation, prior knowledge about human motion, pre-defined articulated body models and no camera model are used to recover motions and 3D information from monocular sequences [13][14]. A linear combination of 3D basis shapes [15][16] has shown to achieve quite good results, even if it was not applied to the full body tracking and reconstruction.

On the other hand, multi-cameras approaches [17][18][19] are employed to increase reliability, accuracy and avoid problems with self-occlusions. New solutions were recently presented employing multiple synchronized video cameras and structured light projectors [27]. They assure an accurate dynamic surface measurement. However, to date, the technology limits the acquisition to small areas of the human body, as for example the face.

In this paper we report about human body modeling and motion reconstruction from

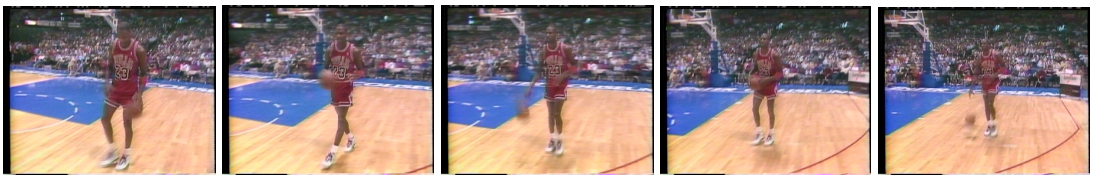

Figure 1: Some frames (720x576 pixel) of a moving character extracted from an old video-tape.

uncalibrated monocular videos as well as from multi-camera image sequences. The approaches presented here combine many ideas and algorithms that have been developed in the recent years at the Institute of Geodesy and Photogrammetry (ETH Zurich). The goal is to bring them together and show reliable and accurate photogrammetric procedures to recover 3D data and generate virtual characters from videos for visualization and animation purposes. The reality-based virtual humans can be used in areas like film production, entertainment, fashion design and augment reality while the recovered 3D positions could serve as basis for the analysis of human movements or medical studies. The analysis of existing monocular videos can furthermore allow the generation of 3D models of characters who may be long dead or unavailable for common modeling techniques.

2. Monocular Videos

A moving character imaged with one moving camera (Figure 1) represents the most difficult case for the deterministic 3D reconstruction of its poses. To avoid copyright problems, existing sport videos are considered. They are usually acquired with a stationary but freely rotating camera and with a very short baseline between the frames.

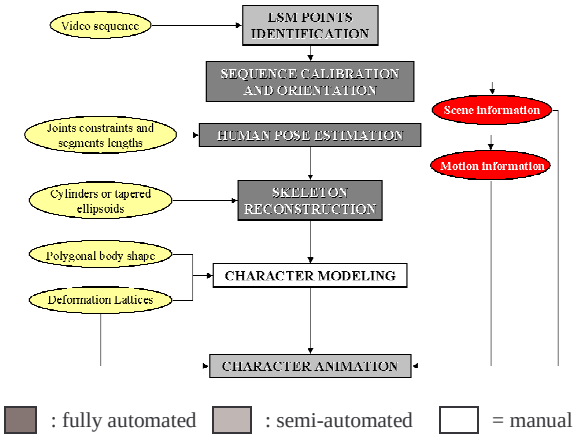

Figure 2: Workflow for character modeling and animation from monocular videos.

The full reconstruction and modeling process (Figure 2) consists of (1) calibration and orientation of the images, (2) pose estimation and human skeleton reconstruction and (3) character modeling and animation. The photogrammetric calibration and orientation of the video is performed with a photogrammetric bundle adjustment [20]. The procedure is required to achieve the camera parameters necessary for the determination of the scene's metric information and for the human's poses estimation. Tie points are measured semi-automatically in the images by means of template Least Squares Matching (LSM) [21] and imported in the adjustment as weighted observations. All the unknown parameters are treated as stochastic variables while significance tests are applied for the determinability of the camera parameters.

Once the sequence is oriented, some key-frames where the 3D poses will be recovered, are selected (while in the other intermediate frames the 3D positions will be automatically interpolated). In each key-frame the human body is firstly reconstructed in a skeleton form, with a series of joints connected with segments of known relative lengths. The human joints are measured semi-automatically through the frames with LSM. Then a scaled orthographic projection, together with constraints on joints depth and segment's perpendicularity, is applied to obtained accurate and reliable 3D models [22].

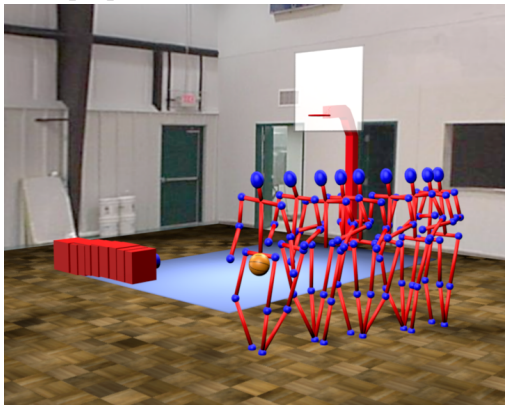

Figure 3: Recovered camera poses and reconstructed moving character in skeleton form.

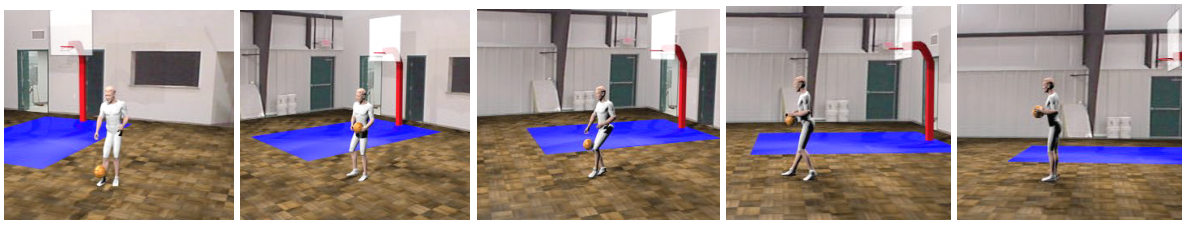

Figure 4: Some frames of the animation created fitting an H-ANIM model onto the recovered 3D poses.

For each key-frame, the recovered 3D human skeleton is afterwards transformed to the absolute reference system with a 3D-conformal transformation and the 3D coordinates are refined within a bundle adjustment using the recovered camera parameters (Figure 3). Finally, to improve the visual quality and realism of the model, an H-ANIM virtual character [23] or a laser scanner polygonal model [1] can be fitted onto the recovered 3D data. The fitting and animation processes are performed with the animation features of Maya 5.0 software [24] (Figures 4 and 5).

The recovered virtual character can be used for augmented reality applications, persons identification or to generate new scenes involving models of characters who are dead or unavailable for common modeling systems.

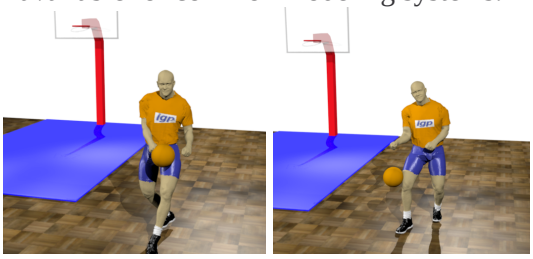

Figure 5: Two views showing the results of the fitting process with a polygonal model.

3. Multi-camera Videos

Multi-stations videogrammetry are usually employed to avoid limbs occlusions and increase reconstruction's reliability. Our method is composed of five steps: (1) acquisition of video sequences, (2) calibration of the system, (3) surface measurement of the human body in each frame, (4) 3D surface tracking and filtering, (5) tracking of key points. Multiple synchronized cameras acquire simultaneously sequences of a scene from different direction (*multi-image sequence*). In the presented example, three synchronized progressive scan CCD cameras (640x480) in a triangular arrangement were used. The imaged person does not necessarily need to be naked. A photogrammetric self-calibration method [25] is used to determine very accurately the exterior and interior camera parameters as well as some additional parameters modelling the distortion caused by the lenses. Afterwards, a surface measurement, based on multi-image LSM [21] with the additional geometrical constraint of the matched point to lie on the epipolar line, is performed. The automatic matching process [26] determines a dense and robust set of corresponding points in the images starting from few seed points automatically selected in the region of interest (*spatial matching*). The 3D coordinates of the matched points are then computed by forward ray intersection using the orientation and calibration data of the cameras (Figure 6).

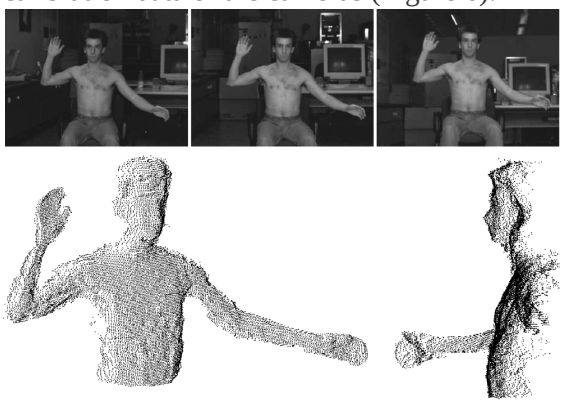

Figure 6: Surface measurement. Top: image triplet. Bottom: computed 3D point cloud.

This process is performed for each image triplet of the multi-image sequence, resulting in a dynamic measurement. Figure 7 shows the determined 3D point clouds in some frames.

As next step, a tracking process, also based on LSM technique, is applied. Its basic idea is to track the corresponding points of each triplet through the sequence and compute their 3D trajectories (*temporal matching*). The spatial correspondences between the triplets acquired at the same time are therefore matched with the subsequent frames (see Figure 8).

Figure 7: Dynamic surface measurement: 3D point clouds for some frame of the sequence.

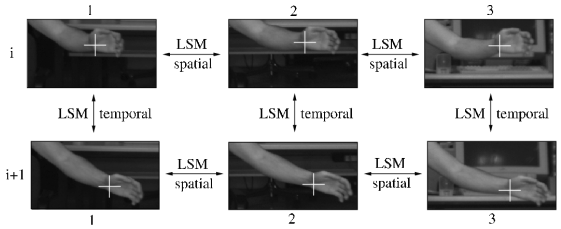

Figure 8: LSM tracking: temporal and spatial correspondences are established with LSM.

The results of the automated tracking process are the coordinates of a point in the three images through the sequence, thus its 3D trajectory is determined 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).

To extract general motion information from the dense set trajectories, the *key-point*s are introduced. A key-point is a 3D region manually defined in the vector field of trajectories, whose size can vary and whose position is defined by its centre of gravity. 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. Key-points can be placed in such a way that their trajectories can describe complex movements. An example is shown in Figure 9.

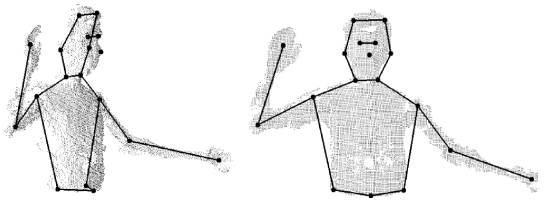

Figure 9: Example of key-points tracked on the human body.

The key-points represent in this case an approximation for the joint trajectories. The final result of the tracking process is shown in Figure 11 together with the corresponding frames; two view of the 3D trajectories of the key-points are displayed in Figure 10.

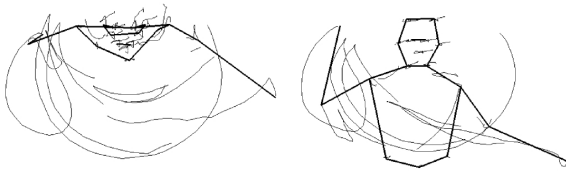

Figure 10: View from the top (left) and from the front (right) of the 3D trajectories of the key-points.

The key-points can afterwards be used for further visualization purposes. Using an H-ANIM human model and the freeware raytracer Povray (Persistence of Vision Ray-Tracer™), the different limbs of an H-ANIM character can be automatically posed in the correct position described by the recovered key-points (see Figure 12).

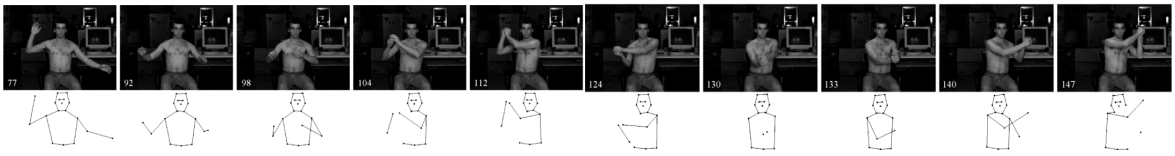

Figure 11: Tracking key-points: some frames of the sequence and view of the tracked key-points.

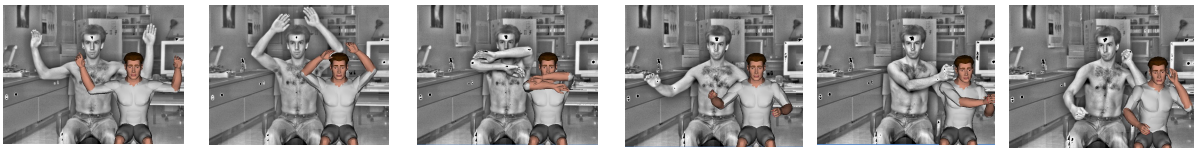

Figure 12: Some frames showing the fitting of an H-ANIM character onto the recovered 3D data using the extracted key-points.

4. Conclusions

In this paper two methods to perform both human body modeling and motion reconstruction from uncalibrated monocular videos as well as from multi-camera image sequences were discussed. The methods are based on reliable and accurate photogrammetric procedures that recover the 3D data from the images through a camera model. Videogrammetry is therefore a powerful and reliable solution for these kinds of applications. The obtained photogrammetric data can be used for gait analysis, biomechanics studies or as motion input for avatars in virtual environments.

References

[1] Cyberware: http://www.cyberware.com [Accessed November 2004]

[2] Vitus: www.vitus.de/english/home_en.html [Accessed November 2004]

[3] B. Allen, B. Curless and Z. Popovic. Articulated body deformation from range scan data. *Proc. SIGGRAPH*, pp.612-619, 2002

[4] H. Seo and N. Magnenat-Thalmann. An Automatic Modeling of Human Bodies from Sizing Parameters", *ACM SIGGRAPH* Symposium on Interactive 3D Graphics, pp. 19-26, 2003

[5] X. Ju, N. Werghi and J.P. Siebert. Automatic Segmentation of 3D Human Body Scans. *International Conference on Computer Graphics and Imaging*, pp 239-244, 2000

[6] F. Cordier, N. Magnemat-Thalman. Real-time animation of dressed virtual humans. *Computer Graphics Forum*, Blackwell, Vol.21(3), 2002

[7] F. Cordier, H. Seo and N. Magnenat-Thalmann. Made-to-Measure technologies for online clothing store. IEEE *CG&A* special issue on 'Web Graphics', pp.38-48, 2003

[8] Ascension: http://www.ascension-tech.com/ [Accessed November 2004]

[9] Vicon: http://www.vicon.com [Accessed November 2004]

[10] Motion Analysis: http://www.motionanalysis.com/ [Accessed November 2004]

[11] C. Sminchisescu. Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences *Ph.D. Dissertation*, INRIA Grenoble, 2002

[12] H. Sidenbladh, M. Black, and D. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. *ECCV02*, Springer Verlag, LNCS 1843, pp. 702-718, 2000

[13] M. Leventon, W. Freeman. Bayesian estimation of 3D human motion from an image sequence. TR-98-06, MERL, 1998

[14] M. Yamamoto, A. Sato, S. Kawada, T. Kondo and Y. Osaki. Incremental Tracking of Human Actions from Multiple Views. IEEE *CVPR Proc*., 1998

[15] C. Bregler, A. Hertzmann, H. Biermann. Recovering Non-Rigid 3D Shape from Image Streams. *Proc. IEEE CVPR* 2000

[16] L. Torresani, D. Yang, E. Alexander, C. Bregler. Tracking and Modeling non-rigid objects with Rank Constraints. *Proc. IEEE CVPR* 2001

[17] N. D'Apuzzo, R. Plänkers, P. Fua, A. Gruen and D. Thalmann. Modeling human bodies from video sequences. In El-Hakim/Gruen (Eds.), Videometrics VI, *Proc. of SPIE*, Vol. 3461, pp. 36-47, 1999

[18] D.M. Gavrila and L. Davis. 3D model-based tracking of humans in action: a multi-view approach. IEEE *CVPR Proc*. pp. 73-80, 1996

[19] S. Vedula and S. Baker. Three Dimensional Scene Flow. *ICCV* '99, Vol. 2, pp. 722-729.

[20] F. Remondino and N. Boerlin. Photogrammetric calibration of sequences acquired with a rotating camera. *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXIV, part 5/W16, 2004

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

[22] F. Remondino and A. Roditakis. Human Figures Reconstruction and Modeling from Single images or Monocular Video Sequences. *IEEE International 3DIM Conference*, pp. 116-123, 2003

[23] H-Anim: http://www.h-anim.org [Accessed November 2004]

[24] Maya: http://www.aliaswavefront.com/ [Accessed November 2004]

[25] H.G. Maas. Image sequence based automatic multi-camera system calibration techniques. *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXII, part 5, pp. 763-768, 1998

[26] N. D'Apuzzo. Measurement and modeling of human faces from multi images. *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXV, part 5, pp. 241-246, 2002

[27] L. Zhang, N. Snavely, B. Curless and S.M. Seitz. Spacetime faces: high-resolution capture for modeling and animation. *ACM SIGGRAPH Proceeding*, 2004