3D Reconstruction of Static Human Body with a Digital Camera

Fabio Remondino

Institute of Geodesy and Photogrammetry
Swiss Federal Institute of Technology
ETH Hoenggerberg
8093 Zurich, Switzerland
E-mail: fabio@geod.baug.ethz.ch
Web: http://www.photogrammetry.ethz.ch/

ABSTRACT
Nowadays the interest in 3D reconstruction and modeling of real humans is one of the most challenging problems and a topic of great interest. The human models are used for movies, video games or ergonomics applications and they are usually created with scanner devices. In this paper a new method to reconstruct the shape of a static human is presented. Our approach is based on photogrammetric techniques and uses a sequence of images acquired around a standing person with a digital still video camera or with a camcorder. First the images are calibrated and orientated using a bundle adjustment and automatically recover the required tie points and initial approximations of the unknowns. After the establishment of a stable adjusted image block, an image matching process is performed between consecutive triplets of images. Finally the 3D coordinates of the matched points are computed with a mean accuracy of ca 2mm by forward ray intersection using the restored camera parameters. The obtained point cloud can then be triangulated to generate a surface model of the body or a virtual human model can be fitted to the recovered 3D data. Results of the 3D human point cloud with pixel color information are presented.

Key words: Calibration, Vanishing Point, Orientation, Epipolar Geometry, Matching, 3D Reconstruction

1. INTRODUCTION

The generation of 3D models from uncalibrated sequences is a challenging problem that has been investigated within different research activities in the last decade. In particular, a topic of great interest is the modeling of real humans, starting from image measurements or range data. Nowadays the demand for 3D models of humans has drastically increased. As result, there are now available many systems that are optimised for extracting accurate measurements of the body and model the whole surface almost automatically. The 3D models are used in movies, video games, virtual environments, e-commerce or ergonomics applications and they are usually created with scanner devices. A complete model of human usually consists of the shape and the movements of the body. The available systems consider these two modeling processes as separate even if they are very close. The issues involved in creating virtual humans are the acquisition of body shape data, the modeling of the data and the acquisition of the information for the animation. A classical approach (Figure 1) to build human shape models uses 3D scanners [e.g. 4, 8, 15, 25, 26]: they are quite expensive but simple to use and various software are available to model the result measurements. They work according to different technologies (laser line, structured light) and provide million of points with often related color information. An overview of some 3D body scanner is presented in Table 1. Other techniques are based on silhouette extraction [27] or multi-image photogrammetry [10]. Moreover animation and modeling software allow to generate virtual human model just subdividing and smoothing polygon boxes and without any measurements [e.g. 1, 2, 17]. The generated human models (Figure 2) can be used for different purposes, like animation, manufacture or medical applications and the quality and accuracy of the results is related to the final application. For animation, only approximative measurements are necessary: the shape can be first defined (e.g. with 3D scanners or smoothing 3D meshes with splines or attaching generalized cylinders to a skeleton or using NURBS on polygons and points) and then animated using motion capture data. For medical applications an exact three-dimensional measurement of the body is required for analysis, representation, measurements or diagnosis and usually performed with 3D scanners. Manufacturing and clothing industries are instead adopting systems and technologies that enable the customers to visualize

themselves in a garment before buying it [4]. Moreover standard sizes can be calculated based on scanned data from hundreds of people and clothes can then be developed accordingly.

	Cyberware	**TC2 Image Twin**	**Vitronics**	**Inspeck**	**Hamamatsu**	**Wicks&Wilson**
Product	WB4, WBX	2T4s	Vitus	3D Full Body	64 BL Scanner	TriForm BS
Time	~17 sec	~12 sec	< 20 sec		<16 sec	~12 sec
resolution/ accuracy	~ 1 mm	~1 mm	~1 mm	~1.5 mm	~1 mm	~2 mm
measurements/ technology	laser line	structured light	laser line	structured light	structured light	structured light
point grid density	3 x 3 mm	2.8 x 2.5 mm			~ 5 x 5 mm	

Table 1: Some 3D Body Scanners with their main characteristics

In this paper a photogrammetric approach for the reconstruction of 3D shapes of static humans from uncalibrated images is described. The process consists of three parts:
1) Acquisition of the images around the static human with a digital camera;
2) Calibration and orientation of the images;
3) Matching process on the human body surface and point cloud generation.
For the complete reconstruction of a human model, a full 360 degree azimuth coverage is required. The image acquisition can last 40 seconds and this could be considered a limit of the process, as no movement of the person are required.
This work belongs to a project called Characters Animation and Understanding from SEquence of images (CAUSE). Its goal is the extraction of complete 3D animation models of characters from old movies or video sequences, where no information about the camera and the objects is available.

Figure 1: Two examples of 3D body scanner: Cyberware [8] (left) and Vitronics Vitus [25] (center). The triangulation process between projector camera and target subject.

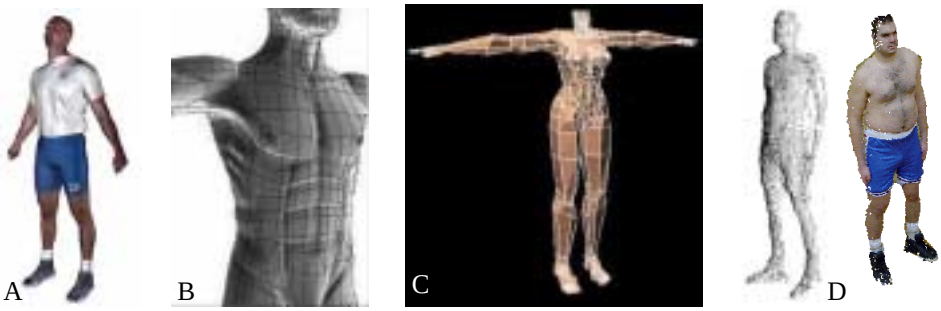

Figure 2: Different examples of generated human model. A: Results obtained with a 3D body laser scanner (Cyberware [8]). B: Results obtained using NURBS (Nonuniform Rational B-Splines) on polygons and points in Lightwave [17]. C: Results of human model created using 3D splines (meshsmooth) in 3D Studio Max [2]. D: Point cloud and texturized points of human model obtained from an image sequence [20]

2. IMAGE SEQUENCE CALIBRATION AND ORIENTATION

The calibration and orientation of the images have to be performed in order to extract precise 3D information of the scene. Our approach is based on a photogrammetric bundle adjustment (section 2.4); the required image correspondences (section 2.1) are found with an automatic process already presented in [20]. Because of its non-linearity, the bundle-adjustment needs initial approximations for the unknown interior and exterior orientation parameters, which are found as described in section 2.2 and 2.3. To make the full orientation process more general and usable for any kind of sequence, no a priori known 3D information is used neither the known parameters of the used camera. Only the focal length of the camera is kept fix not to deal with varying camera constant.

2.1. Automatic tie points identification

The orientation procedure needs a set of corresponding image points between the images in order to recover cameras poses and 3D information of the scene. These correspondences (tie points) are extracted automatically, without manual intervention. The limitation of the approach is that the baseline between the frames must be relatively short (not too small to avoid poorly estimation of the 3D structure) and the rotation around the optical axis should be limited, otherwise the adaptive least square template matching used to find the correspondences will fail.

2.1.1. Interest points

The first step is to detect a set of interest points or corners in each image of the sequence. Harris corner detector is used using a threshold on the number of corners extracted based on the image size. A good point distribution is assured by subdividing the images in small patches (9x9 pixel on an image of 1200x1600) and keeping only the points with the highest interest value in those patches.

2.1.2. Correspondences matching

The extracted features between adjacent images have to be match. At first cross-correlation is used and then the results are refined using adaptive least square matching (ALSM) [12]. The cross-correlation process uses a small window around each point in the first image and tries to correlate it against all points that are inside a bigger window in the adjacent image. The point with biggest correlation coefficient is used as approximation for the template matching process. The process returns the best match in the second image for each interest point in the first image.

2.1.3. Filtering false correspondences

Because of the unguided matching process, the found matched pairs always contain outliers. Therefore a filtering of the incorrect matches has to be performed. To evaluate the quality of the match, the disparity gradient between the correspondences is computed [16, 20]. The smaller is the disparity gradient, the more the two correspondences are in agreement. The sum of all disparity gradients of each matched point relative to all other neighbourhood matches is computed. Those matches that have a disparity gradient sum greater than the median of the sums are removed. In case of big baselines or in presence (at the same time) of translation, rotation, shearing and scale between consecutive images, the algorithm can achieve incorrect results: therefore the filtering process has to be performed locally and not on the whole image.

2.1.4. Epipolar geometry between image pairs

The next step performs a pairwise relative orientation and an outlier rejection using those matches that pass the filtering process. Based on the coplanarity condition, the fundamental matrix is computed with the Least Median of the Squares (LMedS) method; LMedS estimators solve non-linear minimization problems and yield the smallest value for the median of the squared residuals computed for the data set. Therefore they are very robust in case of false matches or outliers due to false localisation. The computed epipolar geometry is then used to refine the matching process, which is now performed as guided matching along the epipolar line.

2.1.5. Epipolar geometry between image triplets

Even if the computed epipolar geometry is correct, not every correspondence that supports this relative orientation is necessarily valid. This because we are considering just the epipolar geometry between couple of images and a pair of correspondences can support the epipolar geometry by chance (e.g. a repeated pattern aligned with the epipolar line). These

kinds of ambiguities and blunders can be reduced considering the epipolar geometry between three consecutive images. A linear representation for the relative orientation of three views is represented by the trifocal tensor T [22]. T is represented by a set of three 3x3 matrices and is computed only with image correspondences without knowledge of the motion or calibration of the cameras. The tensor T is defined up to a scale factor and can be computed from at least 7 correspondences. In our process, for each triplet of images, the tensor is computed with a RANSAC algorithm [11] using the correspondences that support two adjacent pair of images and their epipolar geometry. The RANSAC is a robust estimator, which fits a model (tensor) to a data set (triplet of correspondences) starting from a minimal subset of the data. The found tensor is used to verify whether image points are correct corresponding features between different views. Moreover it is possible to 'transfer points', i.e. compute the image coordinates of a point in the third view, given the corresponding image positions in the first two images and the related tensor [13]. This transfer is very useful when in one view are not found many correspondences. As result of this step, for each triplet of images, a set of corresponding points, supporting the related epipolar geometry is available.

2.1.6. Tracking image correspondences through the sequence

After the computation of a T tensor for every consecutive triplet of images, we consider all the overlapping tensors (e.g. T_{123}, T_{234}, T_{345},...) and we look for those correspondences which are present in consecutive tensors. That is, given two adjacent tensors T_{abc} and T_{bcd} with supporting points (x_a,y_a, x_b,y_b, x_c,y_c) and (x'_b,y'_b, x'_c,y'_c, x'_d,y'_d), if (x_b,y_b, x_c,y_c) in the first tensor T_{abc} is equal to (x'_b,y'_b, x'_c,y'_c) in the successive tensor T_{bcd}, this means that the point in images a, b, c and d is the same and therefore this point must have the same identifier. Each point is tracked as long as possible in the sequence. The obtained correspondences are used as tie points for the successive bundle-adjustment (section 2.4).

2.2. Approximation for the internal parameters

A first approximated value of principal point and focal length of the camera is computed using the vanishing points of the images. Man-made objects are often present in the images, therefore features like straight lines and angles can be used to retrieve information about the used camera or the 3D structure of the captured scene. In particular, a set of parallel lines in object space is transformed by the perspective transformation of the camera into a set of lines that meet in image space in a common point: the vanishing point. Usually in the images, three main lines orientations associated with the three directions of the cartesian axis are visible. Each direction identifies a vanishing point. The orthocenter of the triangle formed from the three vanishing points of the three mutually orthogonal directions identifies the principal point of the camera [6]. The focal length can then be computed as the square root of the product of the distances from the principal point to any of the vertices and the opposite side (Figure 3).

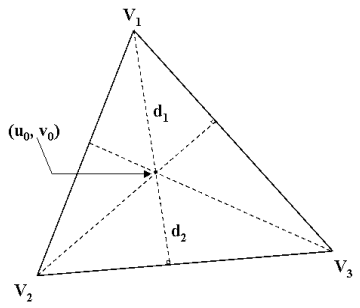

Figure 3: The pricipal point (u_0, v_0) of the camera identified as the orthocenter of the triangle with vertices the three vanishing point V_i. The focal lenght is defined as the square root of the product of the distances d_1 and d_2

The majority of the vanishing points detection methods rely on line segments detected in the images [e.g. 14, 21]. Our approach is also based on line segments clustering and works according to the following steps:

1. straight lines extraction with Canny operator;
2. aggregation of short segments taking into account the segments slope and the distance between segments;
3. identification of three mutually orthogonal directions and classification of the merged lines according to their directions;
4. computation of the three vanishing points for each direction with a SVD approach [7, 20];
5. determination of the principal point and the focal length of the camera [6].

Step 3 can be realized with two different approaches. In a *fully-automatic mode*:
3.1. for each line, compute its slope and its orthogonal distance from the image center;
3.2. plot the 2 entities (e.g. Figure 4) and identify 3 groups, related to the orthogonal directions;
3.3. classify the lines into the related directions according to some threshold values on the slopes.
In a *semi-automatic mode*:
3.1. select two lines to identify one direction;
3.2. intersect the selected lines to compute the associated vanishing point;

3.3. for all the lines compute the orthogonal distances from each vanishing point;
3.4. classify each line into the direction associated with the minimal distance from the vanishing point.

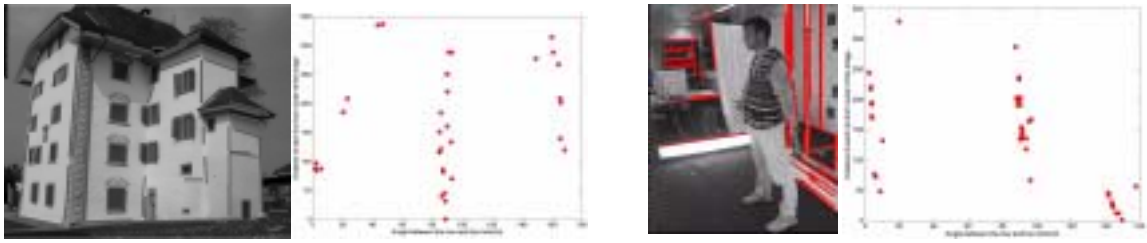

Figure 4:Automatic classification of the extracted lines according to their orientation. Two example of different scenes are presented. Left: an outside image with only a castle. Right: an internal image with different objects. In both cases the lines are corrected divided in three groups and each group is associated to one (orthogonal) direction

2.3. Approximation for the external parameters

The approximations of the exterior orientation are computed using spatial resection. The resection is defined as the process where the spatial position and orientation of an image is determined, based on image measurements and object points information. If at least 3 object points are available, the exterior parameters can be determined without iterations; when a fourth point exists, a unique solution based on least squares can be achieved. In our case, some points measured on the human body are used as reference to compute an approximations of the positions of the cameras.

2.4. Photogrammetric bundle adjustment

The mathematical basis of the bundle adjustment is the collinearity model and it performs a global minimization of the reprojected error. Usually the collinearity equations need to be extended in order to meet the physical reality, by introducing some systematic errors; these errors are compensated with correction terms for the image coordinates, which are functions of a set of additional parameters (AP) [3, 5]. Solving a bundle adjustment means to estimate the additional parameters as well as position and orientation of the camera(s) and object coordinates starting from a set of correspondences in the images (and possible control points).

2.5. Image acquisition and results of the orientation process

For the reconstruction, a set of 12 images (Figure 5) are acquired with a Sony Cybershot digital still video camera. The acquisition process lasted ca 40 seconds and required no movements of the person: this could be considered a limit of the system, but faster acquisition can be realized (e.g. with a standard video camera). The resolution of the acquired images is 1200x1600 pixel.

Figure 5: Six (out of twelve) images used for the 3D reconstruction of the human body

The automatic tie points identification (section 2.1) worked quite well even if a small baseline could allow a bigger number of correspondences. At the end of the process, 150 points are found and imported in the bundle as well as four control points (measured manually on the body) used for the space resection process.
The initial approximations of the interior parameters are computed as described in section 2.2 and the result of the automatic extraction and classification of the lines are reported in Figure 6.

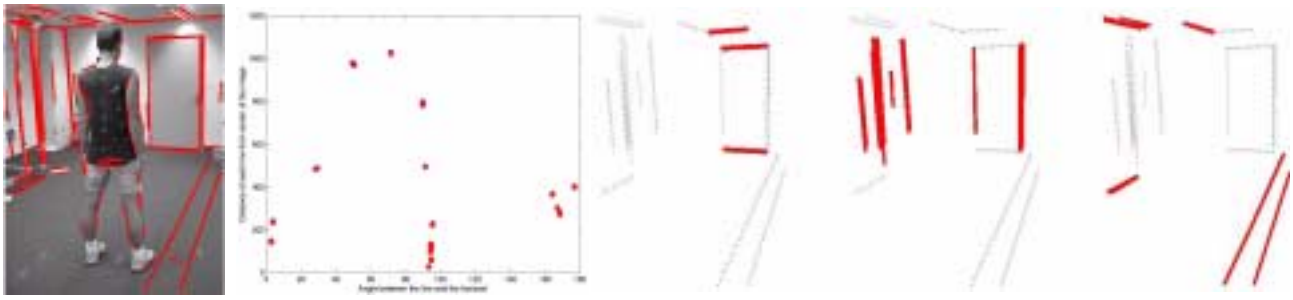

Figure 6: All extracted lines with Canny operator (left). Plot of the longer and aggregated lines according to their slopes(central plot). Lines automatically classified in the three orthogonal directions (right)

At the end of the bundle adjustment, a camera constant of 8.4 mm was estimated while the correct position of the principal point could not be determined because no camera roll diversity was present: therefore it was kept fix in the middle of the images. Concerning the lens distortion parameters, only the first parameter of radial distortion (K1) turned out to be significant while the others were not estimated, as an over-parameterization could lead to a degradation of the results. The average standard deviation of the computed object points coordinates located on the human figure are σ_x= 5.2 mm, σ_y= 5.4 mm and σ_z= 6.2 mm. The final exterior orientation of the images as well as the 3D coordinates of the tie points are shown in Figure 7.

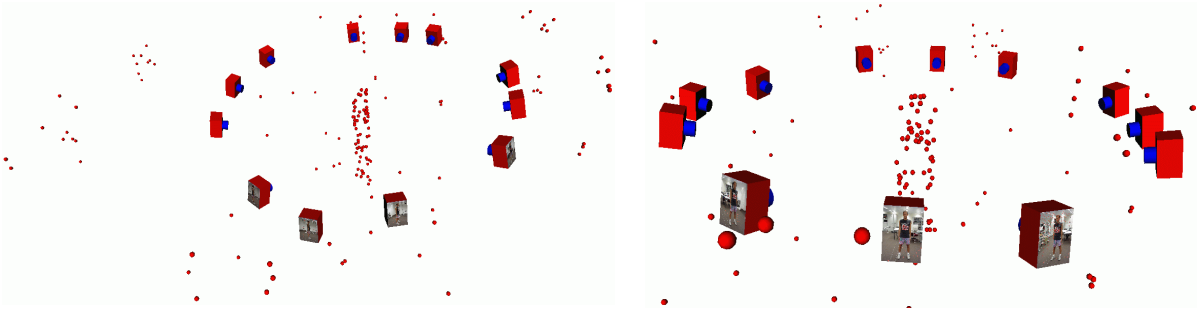

Figure 7: The recovered poses of the 12 images after the adjustment with the 3D coordinated of the used tie points

3. MATCHING PROCESS ON THE HUMAN BODY

In order to recover the 3D structure of the human figure, a dense set of corresponding image points is extracted with an automated matching process [10]. The matching establishes correspondences between triplet of images starting from some seed points selected manually and distributed on the region of interest. 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 matching can fail if lacks of natural texture are presents (e.g. uniform color); therefore the performance of the process can be improved with some local contrast enhancement of the images (Wallis filter).

4. 3D RECONSTRUCTION AND MODELING

The three-dimensional coordinates of the matched points are then computed by forward intersection using the results of the orientation process. In the obtained 3D point cloud some outliers can be present. Therefore, to reduce the noise in the 3D data and to get a more uniform density of the point cloud, a spatial filter is applied. The object space is divided into boxes and the center of gravity is computed for each box. The filter can then be used in two different modes:

- to reduce the density: the points contained in each box are replaced by its center of gravity;
- to remove big outliers: points with big distances from the center of gravity are rejected.

In the generated point cloud some holes could be present if the matching process fails. Therefore a post-processing filling is necessary: the gaps are closed inserting new points based on the density and curvature of the surrounding points.

Moreover, if small movements are occurred during the acquisition, the point cloud of each single triplet could appear misaligned respect to the others. Therefore a 3D conformal transformation can be applied: one triplet is taken as reference and all the others are transformed according to the reference one.
Concerning the modeling of the recovered point cloud, the following solutions are possible:

1. generation of a polygonal surface: from the unorganized 3D data a non-standard triangulation procedure is required. It can be found in commercial packages [e.g. 19, 23, 24], but they need very dense point cloud to generate a correct triangulation and surface model. They allow post-processing operations, like points holes filling or polygons editing.
2. fitting of 3D model of human: this procedure usually does not require the generation of a surface model and the recovered point cloud are used as basis for a fitting process [9, 18]. In [9], a layered human model is used: first a skeleton is defined, then metaballs are used to simulate muscles and skin on the skeleton and finally an adjustment is performed on the parameters of metaballs and skeleton to make the model correspond to the data. In [18] a CAD human model (Figure 8, right) is manually fitted directly on the body measurements. The RAMSIS human model consists of an internal part, the skeleton, and an external part, the body surface; it was developed as a highly efficient CAD tool for the ergonomic design of vehicle interiors.

4.1. Results of the 3D reconstruction

Triplet	2D correspondences	3D points
A (front)	12273	11813
B (lateral1)	5703	5249
C (back)	11576	11450
D (lateral2)	6380	5485

Table 2: Results of the matching: the obtained correspondences and the number of 3D points

The matching algorithm described in section 3 is applied on 4 triplets of images (Table 2): the front one, the two laterals and the triplet on the back part of the person.
Because of some lacks of texture and uniform color of the dress, some holes are present in the resulted matching correspondences. Afterwards, the measured 2D correspondences are converted in an unique 3D point cloud. In those areas with holes, an automatic closure of the gaps is performed. Then after the filtering process, a reduced point cloud of ca 20 000 points is obtained. The visualization of the obtained 3D shape (Figure 8, left) does not respect the quality of the results because the cloud is not enough dense and in the plotting there is overlapping between upper and lower layer of points. For realistic visualization of the results, each point of the cloud is also back-projected onto one image (according to the direction of visualization) to get the related pixel color. The results are presented in Figure 8, central.

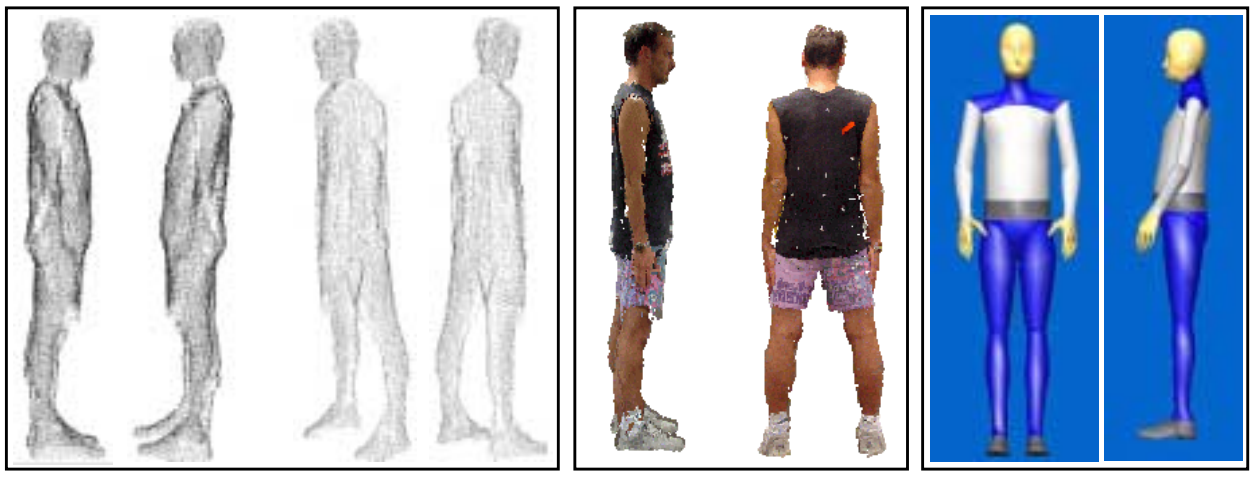

Figure 8: 3D point cloud of the human of figure 1 before and after filtering (left). Visualization of the 3D points with related pixel color (central). RAMSIS [18] human model which can be used to model the recovered 3D data (right)

5. CONCLUSIONS

In this paper we have described a technique for automatic image orientation and calibration (only the space resection step is not fully automated as the control points have to be manually identify in the images). Moreover an approach to recover

3D shape of humans has been presented: it does not require any body suits neither projection of pattern nor particular device for the acquisition. The recovered 3D point cloud of the person are computed with a mean accuracy in x-y of 2.3 mm and in z direction of 3.3 mm. The 3D data can be imported in commercial packages for modeling or used in a fitting process for any visualization or animation purpose by the graphic community.
As future work, the orientation process will be tested on sequences acquired from old videos where people are moving and the 3D reconstruction of these characters will be investigated.

ACKNOWLEDGMENT

I would like to thank Daniela Poli of my Institute for helping me with the forward intersection program.

REFERENCES

1. 3D Render, http://www.3drender.com/ [October 2002]
2. 3D Studio Max, http://www.3dmax.com/ [October 2002]
3. Beyer, H.A., *Geometric and Radiometric Analysis of CCD-Cameras. Based Photogrammetric Close-Range system*. Ph.D. thesis 51, IGP ETH Zurich, 1992
4. BodySkanner, http://www.scansewsuccess.com/ [October 2002]
5. Brown, D.C., *Close-range Camera Calibration*, PE&RS, Vol.37, No.8, pp. 855-866, 1971
6. Caprile B., Torre, V., *Using vanishing point for camera calibration*, International Journal of Computer Vision, Vol.4, No.2, pp. 127-139, 1990
7. Collins, R.T., *Model acquisition using stochastic projective geometry*, PhD thesis, Computer Science Department, Massachussetts University, 1993
8. Cyberware: http://www.cyberware.com [October 2002]
9. D'Apuzzo, N., Plänkers, R., Fua, P., Gruen, A., Thalmann, D., *Modeling human bodies from video sequences*, Videometrics VI, Proc. of SPIE, Vol. 3461, San Jose, USA, pp. 36-47, 1999
10. D'Apuzzo, N., *Modeling human faces with multi-image photogrammetry*. 3-Dimensional Image Capture and Applications V, SPIE Proc., Vol. 4661, pp. 191-197, 2002
11. Fischler, M, Bolles, R., *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*, Comm. Assoc. Comp. Mach., Vol. 24 (6), pp. 381-395, 1981
12. Grün A., *Adaptive least squares correlation: a powerful image matching technique*, South African Journal of Photogrammetry, Remote Sensing and Cartography 14(3), pp. 175-187, 1985
13. Hartley, R., Zissermann, A., *Multi-View Geometry in Computer Vision*, Cambridge University Press, pp. 607, 2000
14. F.A. van den Heuvel, *Vanishing point detection for architectural photogrammetry*, International Archives of Photogrammetry and Remote Sensing, Vol. 32, part 5, pp.652-659,1998
15. Horiguchi C., *Body Line Scanner. The development of a new 3-D measurement and Reconstruction system*, Int. Archives of P&RS Vol.32, part B5, pp.421-429, 1998
16. Klette R., Schlüns, K., Koschan, A., *Computer Vision: Three-dimensional data from images*, Springer Press, 1998
17. Lightwave, http://www.lightwave3d.com/ [October 2002]
18. Ramsis, http://www.ramsis.de [October 2002]
19. Rapid Form, RSI GmbH, http://www.rsi.gmbh.de/rapidform2000_e.htm [October 2002]
20. Remondino, F., *Image Sequence Analysis for Human Body Reconstruction*, Int. Archives of P&RS, Vol. 34, part 5, pp. 590-595, Corfu (Greece), 2002
21. Rother, C., *A new approach for vanishing point detection in architectural environments*, 11th British Machine Vision Conference, Bristol, UK, pp 382-291, 2000
22. Shashua, A., *Trilinearity in visual recognition by aligment*, ECCV, Lectures Notes in Computer Science, Vol.800, Springer-Verlag, pp.479-484, 1994
23. Spider, Alias Wavefront
24. StudioMagic, Raindrop Geomagic Studio, http://www.geomagic.com [October 2002]
25. Vitus: http://www.vitus.de/english/ [October 2002]
26. Wolf, H.G, *Structured lighting for upgrading 2D-vision system to 3D*, Proc. Int.Symposium on Laser, Optics and Vision for Productivity and Manufacturing I, pp. 10-14, 1996
27. Zheng, J.Y., *Acquiring 3D models from sequences of contours*, IEEE Transaction on PAMI, 16(2), pp 163-178, 1994