In Luc Van Gool (Eds.), Pattern Recognition (DAGM 2002), Lecture Notes in Computer Science 2449, Springer 2002, pp. 50-57, Zurich, Switzerland, September 2002

Human Body Reconstruction from Image Sequences

Fabio Remondino

Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland
E-mail: fabio@geod.baug.ethz.ch

Abstract. The generation of 3-D models from uncalibrated sequences is a challenging problem that has been investigated in many research activities in the last decade. In particular, a topic of great interest is the modeling of real humans. In this paper a method for the 3-D reconstruction of static human body shapes from images acquired with a video-camera is presented. The process includes the orientation and calibration of the sequence, the extraction of correspondences on the body using least squares matching technique and the reconstruction of the 3-D point cloud of the human body.

1. Introduction

The actual interests in 3-D object reconstruction are motivated by a wide spectrum of applications, such as object recognition, city modeling, video games, animations, surveillance and visualization. In the last years, great progress in creating and visualizing 3-D models of man-made and non-rigid objects from images has been made, with particular attention to the visual quality of the results. The existing systems are often built around specialized hardware (laser scanner), resulting in high costs. Other methods based on photogrammetry [12, 19] or computer vision [18], can instead obtain 3-D models of objects with low cost acquisition systems, using photo or video cameras. Since many years, photogrammetry deals with high accuracy measurements from image sequences, including 3-D object tracking [15], deformation measurements or motion analysis [7]; these applications requires very precise calibration, but automated and reliable procedures are available. In the last years many researchers have tried to increase the automation level for the production of 3-D models reducing the requirements and the accuracy of the calibration, with the goal of automatically extract a 3-D model "by freely moving a camera around an object" [18].

Concerning the reconstruction and modeling of human bodies, nowadays the demand for 3-D models has drastically increased. A complete model of human consists of both the shape and the movements of the body. Many available systems consider these two modeling processes as separate even if they are very close. A classical approach to build human shape models uses 3-D scanners [Cyberware]: they are expensive but simple to use and various modeling software are available. Other techniques use structured light methods [2], silhouette extraction [23] and multi-image photogrammetry [6]. These human models can be used for different purposes, like animation, manufacture or medical applications. For animation, only approximative measurements are necessary: the shape can be first defined (e.g. with

3D scanners or with meshsmooth or with volumetric primitives attached to a skeleton) and then animated using motion capture data. Both steps can also be joined fitting general body models to different image measurements [7]. For medical applications or in manufacturing industries, an exact three-dimensional measurement of the body is required [14] and usually performed with scanning devices [Tailor].

In this paper a photogrammetric reconstruction of 3-D shape of human bodies from uncalibrated image sequences is described. The recovered 3-D points can then be modeled with commercial software or used to fit 3-D model of humans. The reconstruction process mainly consists of four parts:

1. Acquisition and analysis of the images (section 2);
2. Calibration and orientation of the sequence (section 3);
3. Matching process on the human body between triplets of images and 3-D point cloud generation by forward intersection of the matched points (section 4).

This work belongs to a project called Characters Animation and Understanding from SEquence of images (CAUSE). Its goal is the extraction of complete 3-D animation models of characters from old movies or video sequences, where no information about the camera and the objects is available.

2. Image acquisition

A still video camera or a standard camcorder can be used for the acquisition of the images. In our case, a Sony DCR-VX700E video camcorder is used. The acquisition lasted ca 45 seconds and requires no movements of the person: this could be considered a limit of the system, but faster acquisition can be realized. Nevertheless, the available commercial systems (3D scanners) require ca 20 seconds for the acquisition. For the process 12 frames are selected out of the sequence around the standing person (figure.1). The resolution of the acquired images is 576x720 pixel.

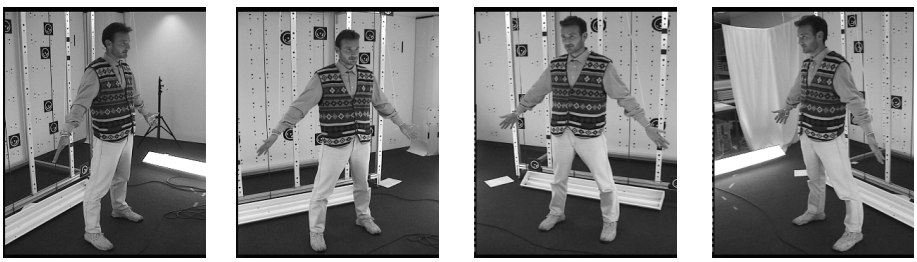

Fig. 1. Four frames (1, 5, 8, 12) of the twelve used for the reconstruction process

The artifacts created by the interlace effect during the digitization process are removed deleting one field of the frame and interpolating the remaining lines. A less smoothly sequence is obtained and the resolution in vertical direction is reduced by 50 per cent. Another possible approach would be to remove lines just in portions of the video where interlacing artifacts are present (adaptive deinterlacing).

The testfield in the background contains many similar targets (repeated pattern) but they are used just as features for the processing. No 3-D information is available.

3. Calibration and orientation of the image sequence

Camera calibration and image orientation are prerequisites for accurate and reliable results for all those applications that rely on the extraction of precise 3-D information from imagery. With the calibration procedure, the geometric deviations of the physical reality from the ideal pinhole camera system are determined. The early theories and formulations of orientation procedures were developed more than 70 years ago and today there is a great number of procedures and algorithms available. A fundamental criterion for grouping these procedures is the used camera model, i.e. the projective camera or the perspective camera model. Camera models based on perspective collineation require stable optics, a minimum of 3 image correspondences and have high stability. On the other hand, projective approaches can deal with variable focal length, but are quite instables, need more parameters and a minimum of 6 image correspondences.

The calibration and orientation process used in this work is based on a photogrammetric bundle-adjustment (section 3.3); the required tie points (image correspondences) are found automatically (section 3.1) with the following steps:

- interest points extraction from each image;
- matching of potential feature pairs between adjacent images;
- false matches clearing using local filtering;
- epipolar geometry computation to refine the matching and remove outliers;
- correspondences tracking in all the image sequence.

In the following section these steps are described. The process is completely automatic; it is similar to [9] and [20], but some additional changes and extensions to these algorithms are presented and discussed.

3.1 Determination of image correspondences

The first step is to find a set of interest points or corners in each image of the sequence. Harris corners detector or Foerstner interest operator are used. The threshold on the number of corners extracted is based on the image size. A good point distribution is assured by subdividing the images in small patches and keeping only the points with the highest interest value in those patches.

The next step is to match points between adjacent images. At first cross-correlation is used and then the results are refined using adaptive least squares matching (ALSM) [10]. 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 ALS matching process. The process returns the best match in the second image for each interest point in the first image. The final number of possible matches between image pairs is usually around 40% of the extracted points.

The found matched pairs always contain outliers, due to the unguided matching process. Therefore a filtering of false correspondences has to be performed. A process based on disparity gradient concept is used [13]. If $\mathbf{P}_{\text{LEFT}}$ and $\mathbf{P}_{\text{RIGHT}}$ as well as $\mathbf{Q}_{\text{LEFT}}$ and $\mathbf{Q}_{\text{RIGHT}}$ are corresponding points in the left and right image, the disparity gradient of the pair ($\mathbf{P}$,$\mathbf{Q}$) is the vector G defined as:

$$G = \frac{|\, D(\mathrm{P}) - D(\mathrm{Q}) \,|}{D_{CS}(\mathrm{P}, \mathrm{Q})} \qquad (1)$$

where

$D(\mathbf{P}) = (\mathbf{P}_{\mathrm{LEFT,X}} - \mathbf{P}_{\mathrm{RIGHT,X}}, \mathbf{P}_{\mathrm{LEFT,Y}} - \mathbf{P}_{\mathrm{RIGHT,Y}})$ is the parallax of **P** between the 2 images, also called disparity of **P;**

$D(\mathbf{Q}) = (\mathbf{Q}_{\mathrm{LEFT,X}} - \mathbf{Q}_{\mathrm{RIGHT,X}}, \mathbf{Q}_{\mathrm{LEFT,Y}} - \mathbf{Q}_{\mathrm{RIGHT,Y}})$ is the parallax of **Q** between the 2 images, also called disparity of **Q**;

$D_{CS} = [(\mathbf{P}_{\mathrm{LEFT}} + \mathbf{P}_{\mathrm{RIGHT}})/2, (\mathbf{Q}_{\mathrm{LEFT}} + \mathbf{Q}_{\mathrm{RIGHT}})/2]$ is the cyclopean separator, e.g. the difference vector between the two midpoints of the straight line segment connecting a point in the left image to the corresponding in the right one.

If **P** and **Q** are close together in one image, they should have a similar disparity (and a small numerator in equation 1). Therefore, the smaller the disparity gradient G is, the more the two correspondences are in agreement. This filtering process is performed locally and not on the whole image: in fact, because of the presence of translation, rotation, shearing and scale in consecutive images, the algorithm achieves incorrect results due to very different disparity values. The sum of all disparity gradients of each match relative to all other neighborhood matches is computed. Then the median of this sum of disparity gradients is found, and those matches that have a disparity gradient sum greater than this median sum are removed. The process removes ca. 80% of the false correspondences. Other possible approaches to remove false matches are described in [17] and [22].

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 process computes the projective singular correlation between two images [16], also called epipolar transformation (because it transforms an image point from the first image to an epipolar line in the second image) or fundamental matrix (in case the interior orientation parameters of both images are the same) [8]. The singular correlation condition between homologous image points of two images is:

$$\mathbf{x}_1^{\mathrm{T}}\ \mathbf{M}\ \mathbf{x}_2 = 0 \quad \text{with} \quad \mathbf{x}_1^{\mathrm{T}} = [\mathrm{x}_1 \quad \mathrm{y}_1 \quad 1], \quad \mathbf{x}_2 = [\mathrm{x}_2 \quad \mathrm{y}_2 \quad 1]^{\mathrm{T}} \qquad (2)$$

Many solutions have been published to compute the 3x3 singular matrix **M**, but to cope with possible blunders, a robust method of estimation is required. In general least median estimators are very powerful in presence of outliers; so the Least Median of the Squares (LMedS) method is used to achieve a robust computation of the epipolar geometry and to reject possible outliers [21].

The computed epipolar geometry is then used to refine the matching process, which is now performed as guided matching along the epipolar lines. A maximal distance from the epipolar line is set as threshold to accept a point as potential match or as outlier. Then the filtering process and the relative orientation are performed again to get rid of possible blunders. However, while the computed epipolar geometry can be correct, not every correspondence that supports the 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. An example can be a repeated pattern that is aligned with the epipolar line (fig.2, left). These kinds of ambiguity and blunders can be removed considering the epipolar geometry between three consecutive images (fig.2, right). A linear representation for

the relative orientation of three images is represented by the trilinear tensor. For every triplet, a tensor is computed with a RANSAC algorithm using the correspondences that support two adjacent images and their epipolar geometry.

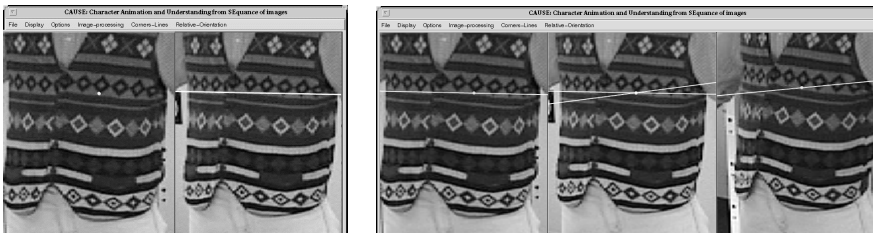

Fig. 2. Left: epipolar line aligns with a repeated pattern.
Right: epipolar geometry between triplet of images used to remove ambiguities and outliers

As result, for each triplet of images, a set of corresponding points, supporting a trilinear tensor, is available. Then we consider all the overlapping tensors (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_b, 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 is equal to (x'_b,y'_b, x'_c,y'_c) in the successive tensor, 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 in a photogrammetric bundle-adjustment.

3.2 Initial approximations for the unknown orientation parameters

Because of its non-linearity, the bundle-adjustment (section 3.3) needs initial approximations for the unknown interior and exterior parameters.

An approach based on vanishing point is used to compute the interior parameters of the camera (principal point and focal length). The vanishing point is the intersection of parallel lines in object space transformed to image space by a perspective transformation of the camera. Man-made objects are often present in the images; therefore geometric information of the captured scene can be derived from these features. The semi-automatic process to determine the approximations of the interior parameters consist of:

- edge extraction with Canny operator and merging of short segments taking into account segments slope and distance from the center of the image;
- interactive identification of three mutually orthogonal directions;
- classification of the extracted and aggregated lines according to their directions;
- computation of the vanishing point for each direction [5];
- determination of the principal point and the focal length of the camera [4].

The approximations of the exterior orientation are instead computed using spatial resection. In photogrammetry, spatial resection is defined as the process where the spatial position and orientation of an image is determined, based on image measurements and ground control points. 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, 4 object points

measured on the human body are used to compute the approximations of the external orientation of the cameras.

3.3 Self-calibration with bundle-adjustment

A versatile and accurate perspective calibration technique is the photogrammetric bundle adjustment with self-calibration [11]. It is a global minimization of the reprojection error, developed in the 50's and extended in the 70's. The mathematical basis of the bundle adjustment is the collinearity model, e.g. a point in object space, its corresponding point in the image plane and the projective center of the camera lie on a straight line. The standard form of the collinearity equations is:

$$x - x_0 = -c \cdot \frac{r_{11}(X - X_0) + r_{21}(Y - Y_0) + r_{31}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{U}{W} \qquad (3)$$

$$y - y_0 = -c \cdot \frac{r_{12}(X - X_0) + r_{22}(Y - Y_0) + r_{32}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{V}{W}$$

where:

x, y are the point image coordinates;

x_0, y_0 are the image coordinates of the principal point PP;

c is the camera constant;

X, Y, Z are the point object coordinates;

X_0, Y_0, Z_0 are the coordinates in object space of the perspective center;

r_{ij} are the elements of the orthogonal rotation matrix R between image and object coordinate systems. R is a function of the three rotation angles of the camera.

The collinearity model needs to be extended in order to take into account systematic errors that may occur; these errors are described by correction terms for the image coordinates, which are functions of some additional parameters (APs). Usually a set of 10 APs is used [1,3] to model symmetric, radial and decentering distortion. Solving a self-calibrating bundle adjustment means to estimate the cameras exterior and interior parameters, the object coordinates of the tie points and the APs, starting from a set of observed correspondences in the images (and possible control points).

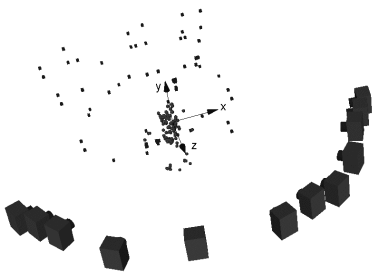

Fig. 3. Recovered cameras poses and object points

The tie points found with the process of section 3.1 are imported in the bundle. Two collinearity equations as in (3) are formed for each image point. Combining all equations of all points in all the images, a system of equations is built. These equations are non-linear with respect to the unknowns and, in order to solve them with a least squares method, must be linearized, thus requiring initial approximations (section 3.2). The resulting exterior orientation of the cameras and the used tie points are shown in figure 3.

4. Matching process and 3-D reconstruction of the human body

After the establishment of an adjusted image block, an automated matching process is performed [6], in order to produce a dense set of corresponding image points. It establishes correspondences between triplet of images starting from few seed points and is based on the adaptive least squares method. One image is used as template and the others as search image. 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. For the process, all consecutive triplets are used. The 3-D coordinates of each matched triplet are then computed by forward intersection using the orientation parameters achieved in phototriangulation (section 3.3). At the end, all the points are joined together to create a unique point cloud of the human body. To reduce the noise in the 3-D data and get a more uniform density of the point cloud, a spatial filter is applied. After the filtering process, a uniform 3-D point cloud is obtained, as shown in figure 4 (left and central). For realistic visualization, each point of the cloud is back projected onto the central image of the sequence to get the related pixel color. The result is presented in figure 4 (right).

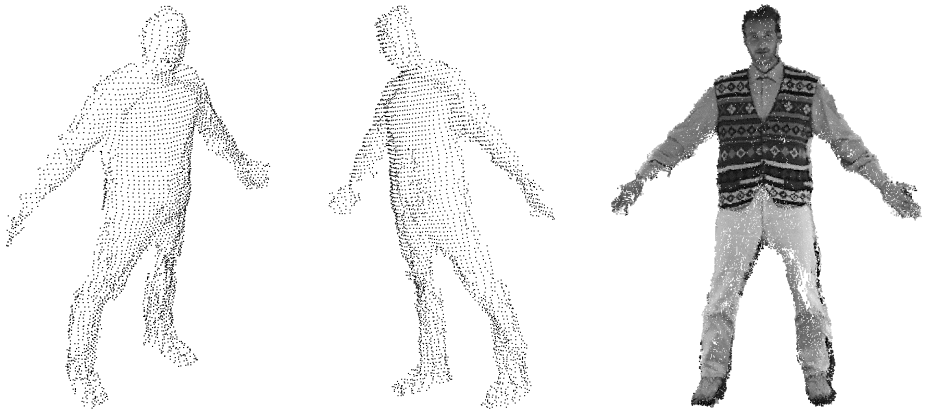

Fig. 4. Recovered 3-D point cloud of the human body and visualization with pixel intensity

5. Conclusions

In this paper a low cost system to create 3-D models of static human bodies from uncalibrated image sequence has been presented. The process is automatic; the obtained 3-D point cloud can be imported in commercial software to easily create a surface model of the person or 3-D human model can be fitted to the recovered data. The processes in [9] and [20] have been extended with ALSM and local filtering while LMedS has been used in relative orientation to remove outliers. As future work, the reconstruction of the body shape will also be extended to the back part of it. The process for the identification of image correspondences described in [17] could be included in our method, weighting the proximity matrix with the sigma of the ALSM

process. Moreover, sequences where the camera is still and the person is moving or both camera and person are moving will be investigated.

References

1. Beyer, H.: Geometric and Radiometric Analysis of CCD-Cameras. Based Photogrammetric Close-Range system. Ph.D. thesis 51, IGP ETH Zurich (1992)
2. Bhatia G., Smith K. E., et al.: Design of a Multisensor Optical Surface Scanner. Sensor Fusion VII, SPIE Proc. 2355 (1994) 262-273
3. Brown, D.C.: Close-range Camera Calibration. PE&RS, Vol.37, No.8 (1971) 855-866
4. Caprile B., Torre, V.: Using vanishing point for camera calibration. International Journal of Computer Vision, Vol.4, No.2 (1990) 127-139
5. Collins, R.T.: Model acquisition using stochastic projective geometry. PhD thesis, Computer Science Dep., University of Massachusetts, 1993
6. D'Apuzzo, N.: Modeling human faces with multi-image photogrammetry. 3-Dimensional Image Capture and Applications V, SPIE Proc., Vol. 4661 (2002) 191-197
7. D'Apuzzo N., Plänkers R.: Human Body Modeling from Video Sequences. Int. Archives of Photogrammetry and Remote Sensing, Vol.32 (1999) 133-140
8. Faugeras O., Luong Q.T., et al.: Camera Self-calibration: Theory and Experiments. Lecture Notes in Computer Science 588, ECCV '92, Springer-Verlag (1992) 321-334
9. Fitzgibbon, A, Zisserman, A.: Automatic 3D model acquisition and generation of new images from video sequences. Proc. of ESP Confernce (1998), pp. 1261-1269
10. Grün A.: Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, RS and Cartography Vol. 14, No. 3 (1985) 175-187
11. Grün A., Beyer, H.: System calibration through self-calibration. In: Grün, Huang (Eds.): Calibration and Orientation of Cameras in Computer Vision, Springer 34 (2001) 163-193
12. Gruen, A., Zhang, L., Visnovcova, J.: Automatic Reconstruction and Visualization of a Complex Buddha Tower of Bayon, Angkor, Cambodia. Proc. 21 DGPF (2001) 289-301
13. Klette R., Schlüns, K., Koschan, A.: Computer Vision: Three-dimensional data from images. Springer (1998)
14. McKenna P.: Measuring Up. Magazine of America's Air Force, Vol. XL, No.2 (1996)
15. Maas, H. G.: Digital Photogrammetry for determination of tracer particle coordinates in turbulent flow research. PE&RS, Vol.57, No.12 (1991), 1593-1597
16. Niini, I.: Relative Orientation of Multiple images using projective singular correlation. Int. Archives of Photogrammetry and Remote Sensing, Vol. 30, part 3/2 (1994), 615-621
17. Pilu, M: Uncalibrated stereo correspondences by singular value decomposition. Technical Report HPL-97-96, (1997), HP Bristol
18. Pollefeys, M.: Tutorial on 3-D modeling from images. Tutorial at ECCV 2000 (2000)
19. Remondino, F.: 3-D reconstruction of articulated objects from uncalibrated images. 3-Dimensional Image Capture and Applications V, SPIE Proc., Vol. 4661 (2002) 148-154
20. Roth, G., Whitehead, A.: Using projective vision to find camera positions in an image sequence. 13th Vision Interface Conference (2000)
21. Scaioni, M.: The use of least median squares for outlier rejection in automatic aerial triangulation. In: Carosio, Kutterer (Eds): Proc. of the 1st Int. Symposium on "Robust Statistics and Fuzzy Techniques in Geodesy and GIS", ETH Zurich (2001) 233-238.
22. Zhang, Z., Deriche, R, et al.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. TR 2273, INRIA (1994)
23. Zheng, J. Y: Acquiring 3D models from sequences contours. IEEE Transaction on PAMI, 16 (2), pp 163-178 (1994)

Cyberware: http://www.cyberware.com; Taylor: http://www.taylor.com [June 2002]