Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER Computer Vision and Image Understanding 93 (2004) 65–85

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

3-D reconstruction of static human body shape from image sequence

Fabio Remondino*

Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology, ETH-Hoenggerberg, 8093 Zurich, Switzerland

Received 10 July 2002; accepted 20 August 2003

Abstract

The generation of 3-D models from uncalibrated image 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 realistic humans, for animation, manufacture or medicine purposes. Nowadays the common approaches try to reconstruct the human body using specialized hardware (laser scanners) resulting in high costs. In this contribution a different method for the three-dimensional reconstruction of static human body shape from monocular image sequence is presented. The core of the presented work describes the calibration and orientation of the images, mostly based on photogrammetric techniques. Then the process includes also the extraction of correspondences on the body using a least squares matching algorithm and the reconstruction of the 3-D body model in point cloud form.

Keywords: Camera calibration; Image orientation; Least squares matching; Human shape reconstruction

1. Introduction

In the last years, great progress in creating and visualizing 3-D models from images has been made, with particular attention to the visual quality of the results. The interests in 3-D object reconstruction are motivated by a wide spectrum of

* Fax: +41-1-633-1101.
E-mail address: fabio@geod.baug.ethz.ch.
URL: http://www.photogrammetry.ethz.ch/.

1077-3142/$ - see front matter
doi:10.1016/j.cviu.2003.08.006

applications, such as object recognition, city modeling, video games, animations, surveillance, and visualization. The mostly used systems are often built around specialized hardware (e.g., laser scanner), resulting in high costs. Other methods based on photogrammetry [13,23] or computer vision [22], can instead obtain 3-D models of objects with low cost acquisition systems, using photos or video cameras. Since many years, photogrammetry deals with high accuracy measurements from image sequences, including 3-D object tracking, deformation measurements or motion analysis [5]; even if these applications require very precise calibration, automated, and reliable procedures are available (e.g. [37]).

Concerning the reconstruction and modeling of static human bodies, nowadays the demand for 3-D models has drastically increased. A complete model of a human consists of both the shape and the movements of the body. These two modeling processes are often considered as separate even if they are very close. The issues involved in creating virtual humans are the acquisition of body shape data (Fig. 1), the modeling of the data and the acquisition of the information for the animation. The common approaches used to recover the (static) shapes are:

- 3-D scanners (e.g. [17,33,44]): they are expensive but simple to use and related software are available to edit and model the obtained point clouds. Body scanners usually capture the shape of the entire human body in ca. 20 s. They use the triangulation principle, with laser light or pattern projection method and a CCD camera(s). Their resolution is approximately of 2 mm and they can also acquire color texture. The results are precise 3-D data of the subject that are then modeled with reverse engineering software. An overview of some 3-D body scanner is presented in Table 1.
- Silhouette extraction [16,30]: they use different images acquired around a static person and usually fit a pre-defined 3-D human model to the extracted image data (Fig. 1E).
- Computer animation software [31,34–36]: these splines-based packages allow the reconstruction of 3-D models without any measurements and 3-D meshes are created smoothing simple polygonal elements (Figs. 1B–D).

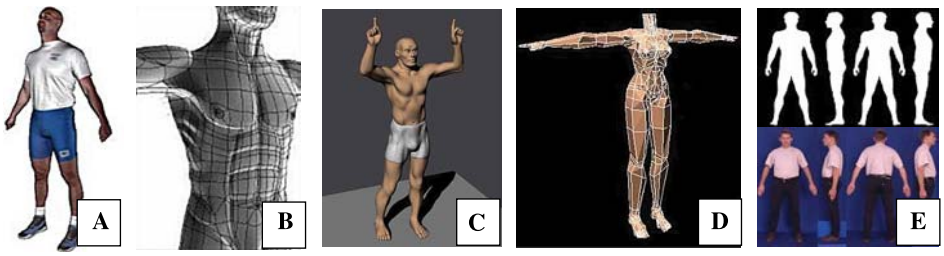

Fig. 1. Different examples of generated human models. (A) Results obtained with a 3-D body scanner (Cyberware [33]). (B) Results obtained using NURBS (Non-uniform Rational B-Splines) on polygons and points in Lightwave [35]. (C) NURBS results in SulpLand [41]. (D) Human model created using 3-D splines ('meshsmooth') in 3-D Studio Max [31]. (E) Silhouette extraction from image sequence for 3-D model reconstruction [16].

Table 1
Some 3-D body scanners with their main characteristics

	Cyberware	TC2 image twin	Vitronics	Inspeck	Hamamatsu	Wicks & Wilson
Product	WB4, WBX	2T4s	Vitus	3-D Full Body	64 BL Scanner	TriForm BS
Time (s)	~17	~12	<20		<16	~12
Accuracy (mm)	~1	~1	~1	~1.5	~1	~2
Technology	Laser line	Structured light	Laser line	Structured light	Structured light	Structured light
Point density (mm)	3×3	2.8×2.5			$\sim 5 \times 5$	

The recovered human body models can be used in different fields, like animation, surveillance, manufacturing, medicine or for ergonomic applications. For animation purpose, only approximative reconstructions are necessary. The shape of static human is first defined (e.g., with image measurements, scanners or animation packages) and then animated with animation packages or using motion capture data [32]. On the other hand, medical applications or manufacture industries, required digital surfaces for metric body measurements and for clothes design [19]; therefore exact 3-D models of the body are needed and usually performed with scanning devices [33,43].

In this work, a photogrammetric approach for the reconstruction of 3-D shapes of static humans using a monocular uncalibrated image sequence is described. The process consists of four parts:

(a) Acquisition and analysis of the image sequence (Section 2);
(b) Calibration and orientation of the images (Section 3);
(c) Matching process on the human body surface (Section 4)
(d) Point cloud generation and modeling (Section 5).

In the process, no a priori information on camera internal and external parameters are assumed; all the required parameters are recovered from the images. The presented 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 video sequences or old movies, where no information of the cameras and the objects is available.

The main aims of the presented work are (1) to present an automatic and reliable approach to orient and calibrate self-acquired images or frames extracted from video sequences and (2) to reconstruct 3-D model of static person from monocular image sequences.

2. Image acquisition

The images can be acquired with a still-video camera or with a camcorder. A complete reconstruction of the human body requires a 360° azimuth coverage, even if in

Fig. 2. Five images (out of twelve) used for the reconstruction process.

Section 6 some examples of 3-D shapes reconstructed from frames acquired only in front of the body are also presented. The acquisition lasts ca. 40 s (less time is necessary if a camcorder is used) and requires no movements of the person. This could be considered a limit of the procedure (if we consider that body scanners require less than 20 s) but a possible solution is presented in Section 5.

Fig. 2 shows five images (out of twelve) of a sequence acquired with a Sony DSC-S70, with a resolution of 1200×1600 pixels. During the acquisition, the camera constant is kept fixed not to deal with varying internal parameters.

3. Calibration and orientation of the images

All applications that deal with the extraction of precise 3-D information from imagery require accurate calibration and orientation procedures as prerequisites of reliable results. The early theories and formulations of orientation procedures were developed in the first half of the 19th century and today a great number of procedures and algorithms are available. A fundamental criterion for grouping the orientation procedures is based on the used camera model:

- perspective camera model: camera models based on perspective collineation have high stability, require a minimum of three corresponding points per image and a stable optics; they often contain non-linear relations, requiring initial approximations of the unknowns;
- projective camera model: projective approaches can deal with variable focal length, but need more parameters, a minimum of six corresponding points and are quite instable (equations need normalization); they are often linear (e.g., Direct Linear Transformation) and easy to handle.

The choice of the camera model is often related to the final application and the required accuracy. As photogrammetry deals with precise measurements from images, accurate calibration of the used sensor is one of the major goals. In case of single camera sensor, the geometrical model for processing is a perspective projection and the associate procedure for the adjustment of the image measurements and the estimation of the camera parameters is the bundle method. The bundle method is considered the most flexible, general, and accurate sensor model, widely used in close-range applications.

The following paragraphs present the calibration and orientation process, which is the core of the work and is based on a photogrammetric bundle-adjustment (Section 3.3): the required tie points (image correspondences) are found automatically as described in Section 3.1, while the approximations of the camera parameters are obtained as reported in Section 3.2.

3.1. Finding image correspondences

The tie points required for the sequence orientation are extracted with a process similar to [9,24], but additional changes and extensions to these approaches are presented and discussed. Moreover the process has been tested also on wide base-line sequences, producing a good number of correspondences.

3.1.1. Interest points

The first step finds a set of interest points or corners in each image of the sequence. Interest points are geometrically stable under different transformations and have high information content. Many algorithms are available, but, in our case, Harris corners detector and Förstner operator turned out to be the most reliable and with the best results. Harris [15] computes a matrix related to the auto-correlation function of the image. The squared first derivatives of the signal are averaged over a window and the eigenvalues of the resulting matrix are the principal curvatures of the auto-correlation function. An interest point is detected if the found two curvatures are high. Förstner operator [10] uses also the auto-correlation function to classify the pixels into categories (interest points, edges or region). Further statistics performed locally allow estimating automatically the thresholds for the classification.

In our applications, 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.

3.1.2. First matching process

Using the extracted corners, at first cross-correlation is performed between image pairs and then the results are refined using adaptive least squares matching (ALSM) [11]. 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 matching process. The least squares matching considers the two image regions as discrete two-dimensional functions, $f(x,y)$ and $g(x,y)$; they can be defined as conjugate regions of a stereo-pair in the left and the right image. $f(x,y)$ is the template, $g(x,y)$ is the patch in the other image. The matching process establishes a correspondence if $f(x,y) = g(x,y)$. Because of random effects (noise) in both images, the previous equation is not consistent. Therefore, a noise vector $e(x,y)$ is added, resulting in $f(x,y) - e(x,y) = g(x,y)$. The location of the function values $g(x,y)$ must be determined in order to provide for the correct match point. This is achieved by minimizing a goal function, which measures the distances between the gray levels in the

template and in the patch. The goal function to be minimized is usually a L_2-norm of the residuals of least squares estimation. The location is described by shift parameters which are estimated with respect to the initial position of $g(x, y)$. In order to account for a variety of systematic image deformations and to obtain a better match, image shaping parameters (affine image shaping), and radiometric corrections are usually also introduced beside the shift parameters.

Often, in sequence analysis, the strength of the candidate matches is only measured with the correlation coefficient while a least squares method is a stronger and widely accepted technique for subpixel accuracy that increases the reliability of the found correspondences. The process returns the best match in the second image for each interest point in the first image. The final number of possible matches depends on the threshold parameters of the ALSM and on the disparity between image pairs; usually it is around 40% of the extracted points.

3.1.3. Filtering process to remove outliers

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 [18]. If $\mathbf{P}_{\mathrm{LEFT}}$ and $\mathbf{P}_{\mathrm{RIGHT}}$ as well as $\mathbf{Q}_{\mathrm{LEFT}}$ and $\mathbf{Q}_{\mathrm{RIGHT}}$ are corresponding points in the left and right image (Fig. 3), the disparity gradient of two points ($\mathbf{P}$, $\mathbf{Q}$) is the *vector* $\mathbf{G}$ defined as:

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

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 $\mathbf{P}$, e.g., the pixel distance of $\mathbf{P}$ between the 2 images; $D(\mathbf{Q}) = (\mathbf{Q}_{\mathrm{LEFT,X}} - \mathbf{Q}_{\mathrm{RIGHT,X}}, \mathbf{Q}_{\mathrm{LEFT,Y}} - \mathbf{Q}_{\mathrm{RIGHT,Y}})$ is the parallax of $\mathbf{Q}$, e.g., the pixel distance of $\mathbf{Q}$ between the 2 images; D_{CS} is the cyclopean separator, e.g., the distance between the two midpoints P' and Q' of the straight segments connecting a point in the left image to the corresponding in the right one.

If $\mathbf{P}$ and $\mathbf{Q}$ are close together in both images, they should have a similar parallax (e.g., a small numerator in Eq. (1)). Therefore, the smaller the disparity gradient $\mathbf{G}$ is, the more the two correspondences are in agreement. The performance of the filtering are improved if the process is performed locally and not on the whole image, because the algorithm can achieve incorrect results due to very different disparity values and

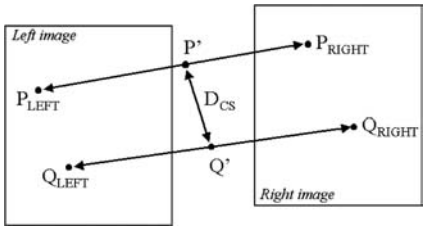

Fig. 3. The disparity gradient between two correspondences ($\mathbf{P}$ and $\mathbf{Q}$) in image left and right.

in presence of translation, rotation, shearing, and scale between consecutive images. The image is divided in patches (usually 6 or 8, according to the image size); then for each patch, the sum $\mathbf{G}_{\mathrm{SUM}}$ of all disparity gradients $\mathbf{G}$ of each matched point relative to all other neighborhood matches inside the patch is computed. Those points that have a disparity gradient sum $\mathbf{G}_{\mathrm{SUM}}$ greater than the median of the sums are rejected. The process removes ca. 80% of the false correspondences. This simple test on the local consistency of the matched points is very useful to remove bad matches at low computational time. Other possible algorithms used to remove false correspondences are described in [21] and [29]. The process for the identification of image correspondences described in [21] could be also included in our method, weighting the proximity matrix with the sigma of our ALSM process (instead of the used cross-correlation coefficient).

3.1.4. Relative orientation between image pairs

The aim is to perform a pairwise relative orientation and an outliers rejection using those matches that pass the previous filtering step. Based on the coplanarity condition, the process computes the projective singular correlation between two images [20], 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) [7]. The fundamental matrix $\mathbf{M_{12}}$ is defined by the equation:

$$\mathbf{p}_1^{\mathrm{T}}\ \mathbf{M}_{12}\ \mathbf{p}_2 = 0 \quad \text{with} \quad \mathbf{p}_i = [\,x_i \quad y_i \quad 1\,]^{\mathrm{T}} \tag{2}$$

for every pair of matching points $\mathbf{p_1}$, $\mathbf{p_2}$ (homogeneous vectors) in image 1 and 2. The epipoles of the images are defined as the right and left null-space of $\mathbf{M_{12}}$ and can be computed with singular value decomposition of $\mathbf{M_{12}}$. A point $\mathbf{p_2}$ in the second image lies on the epipolar line $\mathbf{l_2}$ defined as $\mathbf{l_2} = \mathbf{M_{12}p_1}$ and must satisfy the relation $\mathbf{p_2 l_2} = 0$. Similarly, $\mathbf{l_1} = \mathbf{M_{12}^T p_2}$ represents the epipolar line in the first image corresponding to $\mathbf{p_2}$ in the second image. The 3×3 singular matrix $\mathbf{M}$ can be computed just from image points and at least 7 correspondences are needed to compute it. Many solutions have been published to compute $\mathbf{M}$, but to cope with possible blunders, a robust method of estimation is required. In general RANSAC-like algorithms [8] and least median estimators are very powerful in presence of outliers; therefore the least median of the squares (LMedS) method is used to achieve a robust computation of the epipolar geometry and to reject possible outliers [25,29]. LMedS estimators solve non-linear minimization problems and yield the smallest value for the median of the squared residuals computed for the data set. For this reason they are very robust in case of false matches or outliers due to false localization.

3.1.5. Guided matching process

The computed epipolar geometry is then used to refine the matching process, which is now performed as guided matching along the epipolar lines. This geometric constraint restricts the searching area and allows a higher threshold for the matching process. A maximal distance from the epipolar line is also 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 other possible blunders. After these processes, the number of matched points between the image pair is around 60% of the extracted interest points.

3.1.6. Relative orientation between triplet of images

While the computed epipolar geometry between image pairs can be correct, not every correspondence that supports the relative orientation is necessarily valid. This because we are considering just 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, as shown in Fig. 4, right). 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** [26]; it is represented by a set of three 3×3 matrices and is computed only with image correspondences without knowledge of the motion or calibration of the cameras. For every triplet of views (Fig. 4, left), if $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ are corresponding points in the images, then for every line $\mathbf{l}_2$ through $\mathbf{p}_2$ in image 2 and for every line $\mathbf{l}_3$ through $\mathbf{p}_3$ in image 3, the fundamental trifocal constraint states:

$$\mathbf{l}_2^{\mathrm{T}}[\mathbf{T}\mathbf{p}_1]\mathbf{l}_3 = 0 \quad \text{with} \quad [\mathbf{T}\mathbf{p}_1]_{ij} = T_1^{ij}x_1 + T_2^{ij}y_1 + T_3^{ij} \tag{3}$$

If we consider only the corresponding points, each triplet $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ must satisfy the matrix equation:

$$[\mathbf{p}_2]_x[\mathbf{T}\mathbf{p}_1][\mathbf{p}_3]_x = 0 \tag{4}$$

with $[\mathbf{p}]_x$ a skew-symmetric matrix of the homogeneous vector $\mathbf{p}$.

If a triplet of points $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ satisfy Eq. (4), it means that the corresponding points support the tensor T_{123}.

Relation (4) can be used to verify whether image points (or lines) are correct corresponding features between different views. Moreover, with constraint (4), it is possible to *transfer* points, e.g., compute the image coordinates of a point in the third view, given the corresponding image positions in the first two images and the related T tensor. This transfer is very useful when in one view are not found many correspondences. The point transfer can be solved also using the fundamental matrix,

Fig. 4. Three views geometry: correspondences $\mathbf{p}_i$ and $\mathbf{l}_i$ corresponding to point P and line L (left). Relative orientation between triplet of images (right).

but the trifocal constraint can avoid ambiguities and remove blunders. Moreover, from the tensor it is possible to derive the fundamental matrices between the first and the third view; e.g., given 3 images, $\mathbf{M_{13}}$ between image 1 and 3 is given by:

$$\mathbf{M_{31}} = [\mathbf{e}_3]_x [\, T_1, \quad T_2, \quad T_3 \,]\mathbf{e}_2, \tag{5}$$

where $\mathbf{e}_i$ is the epipole of image i and $[\mathbf{e}_i]_x$ is the skew-symmetric matrix formed with $\mathbf{e}_i$.

Therefore, the transfer of $\mathbf{p}_3$ can be expressed as:

$$\mathbf{p}_3 = (\mathbf{M_{13}}\mathbf{p}_1) \times (\mathbf{M_{23}}\mathbf{p}_2), \tag{6}$$

e.g., the intersection of two epipolar lines in the third view.

The 27 unknowns of the tensor $\mathbf{T}$, defined up to a scale factor, are computed with a RANSAC algorithm [8] using the correspondences that support two adjacent pair of images and their epipolar geometry. Once $\mathbf{T}$ has been computed, a projective transfer is performed in image area where the density of matched points (number of found correspondences) is lower than a predefined threshold value.

3.1.7. Tracking of the found correspondences in all the sequence

When the matching process between image pairs and triplets is completed, we consider all the overlapping tensors ($T_{123}, T_{234}, T_{345}, \ldots$) and we look for those correspondences which support 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 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 for the successive bundle-adjustment (Section 3.3).

3.2. Initial approximation of the unknowns

Because of its non-linearity, the bundle-adjustment (Section 3.3) needs initial approximations for the unknown interior and exterior orientations. A new fully automated approach to recover the interior parameters of the camera is presented while a standard routine is applied to compute the initial camera poses.

3.2.1. Interior orientation

An approach based on vanishing point is used to compute the interior parameters of the camera (i.e., 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. 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

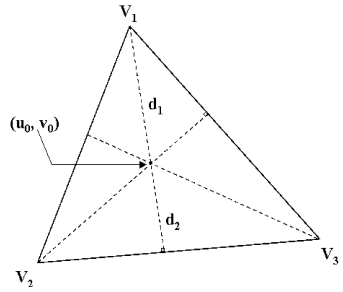

Fig. 5. The principal point (u_0, v_0) of the camera identified as the orthocenter of the triangle with vertices the three vanishing point V_i. The focal length is defined as the square root of the product of the distances d_1 and d_2.

camera [3]. 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 (Fig. 5). The process used to determine the approximations of the interior parameters is based on:

1. Straight lines extraction with Canny operator;
2. merging short segments taking into account segments slope and distance from the center of the image;
3. identification of three mutually orthogonal directions and classification of the extracted and aggregated lines according to their directions;
4. computation of the three vanishing points for each direction [4]. Each line $\mathbf{l}_i$ is represented by its homogeneous coordinates (a_i, b_i, c_i); if there are only two lines, the cross product of them gives the coordinates of the vanishing point; if n lines $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n$ are involved, we get the "best fit" vanishing point forming the matrix $\mathbf{L}$ as:

$$\mathbf{L} = \sum_{i=1}^{n} \begin{bmatrix} a_i a_i & a_i b_i & a_i c_i \\ a_i b_i & b_i b_i & b_i c_i \\ a_i c_i & b_i c_i & c_i c_i \end{bmatrix} \tag{7}$$

and computing the vanishing point as the eigenvector associated with the smallest eigenvalue.

5. determination of the principal point and the focal length of the camera [3].

Step 3 can be realized with two different approaches. In a *fully-automated mode*:

3.1. for each line, compute its slope and its orthogonal distance from the image center;
3.2. plot the 2 entities and identify 3 groups, related to the orthogonal directions of the extracted lines (e.g. Fig. 6);
3.3. classify the lines into the related directions according to some threshold values on the slopes.

In a *semi-automated mode*:

3.1. select two lines to identify one direction;
3.2. intersect the selected lines to compute the associated vanishing point;

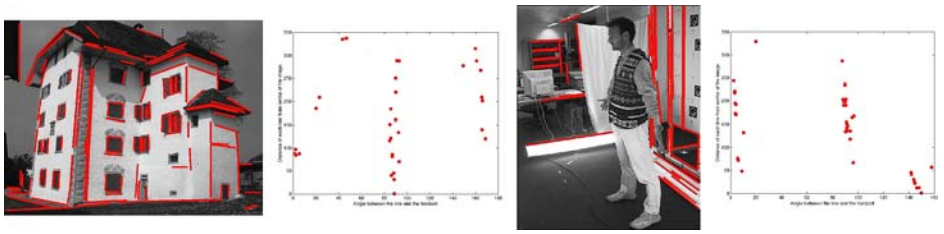

Fig. 6. Automatic classification of the extracted lines according to their orientation. Two examples of different scenes are presented. (Left) An external image with only a castle. (Right) An internal image with different objects. In both cases the lines are corrected divided in three groups and each group is associated to one (orthogonal) direction.

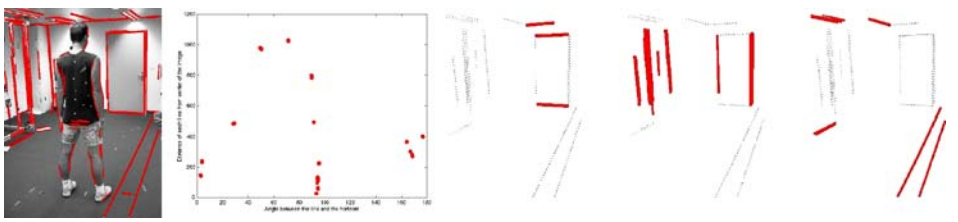

Fig. 7. Main lines extracted with Canny operator (left). Plot of the longer and aggregated lines according to their slopes and distance from the image center (middle). Lines automatically classified in the three orthogonal directions (right).

3.3. for all the extracted 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.

Usually semi-automated approaches are used and other possible methods that use the image itself and the geometrical properties of imaged objects to calibrate a camera are described in [14,27,28].

In Fig. 7 the results of the described process, applied on the sequence of Fig. 2, are shown. The longer lines are correctly classified with the fully-automated approach into the three directions.

3.2.2. Exterior orientation

The required initial approximations of the exterior parameters (3 positions of the projection center and 3 rotation angles) are instead computed using spatial resection. In photogrammetry, spatial resection is defined as a process where the spatial position and orientation of an image is determined, based on image measurements and ground control points, through collinearity condition. 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. Photogrammetric bundle adjustment

A versatile and accurate perspective orientation procedure is the photogrammetric bundle adjustment with self-calibration [12]. It is a global minimization of the reprojection error, developed in the 1950s and extended in the 1970s. 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:

$$\begin{aligned} x - x_0 &= -c \cdot \frac{r_{11}(X - X_0) + r_{21}(Y - Y_0) + r_{31}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{U}{W}, \\ y - y_0 &= -c \cdot \frac{r_{12}(X - X_0) + r_{22}(Y - Y_0) + r_{32}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{V}{W}, \end{aligned} \tag{8}$$

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 (i.e., self-calibration by additional parameters); these errors are described by correction terms for the image coordinates, which are functions of some additional parameters (APs). A set of additional parameters widely used in photogrammetry [1,2] consists of the parameters of interior orientation $(\Delta x_P, \Delta y_P, \Delta c)$, a scale factor for the uncertainty in pixel spacing (S_X), a shear factor (A) modeling a non-orthogonality of the image coordinate system, the parameters describing symmetrical radial lens distortion (K_1, K_2, K_3), and parameters of decentering lens distortion (P_1, P_2). The extended collinearity equations have the following form:

$$\begin{aligned} x - x_0 &= -c \cdot \frac{U}{W} + \Delta x, \\ y - y_0 &= -c \cdot \frac{V}{W} + \Delta y, \end{aligned} \tag{9}$$

where:

$$\begin{aligned} \Delta x &= -x_0 - \frac{\overline{x}}{c}\Delta c - \overline{x}S_x + \overline{y}A + \overline{x}r^2K_1 + \overline{x}r^4K_2 + \overline{x}r^6K_3 + (2\overline{x}^2 + r^2)P_1 + 2P_2\overline{xy}, \\ \Delta y &= -y_0 - \frac{\overline{y}}{c}\Delta c + \overline{x}A + \overline{y}r^2K_1 + \overline{y}r^4K_2 + \overline{y}r^6K_3 + +2P_1\overline{xy} + (2\overline{y}^2 + r^2)P_2 \end{aligned} \tag{10}$$

with:
- K_i: parameters of symmetrical radial lens distortion
- P_i : parameters of decentering lens distortion
- $\overline{x} = x - x_0;\ \overline{y} = y - y_0$
- $r = \sqrt{\overline{x}^2 + \overline{y}^2}$

The functions in Eq. (10) are called 'physical model,' as all the components (APs) can be attributed to physical error sources.

Solving a (self-calibrating) bundle adjustment means to estimate the additional parameters in Eq. (10) as well as position and orientation of the camera(s) and object coordinates using the correspondences in the images (and possible control points, to have a consistent reference system). Two collinearity equations can be formed for each image point. Combining all equations of all points in all the images leads to a system of equations to be solved.

Eq. (8) (or the extended version in (9)) can be described as:

$$l = f(x), \tag{11}$$

i.e., a function that relates the image observations to the parameters in the right side, where:
- $x = [\Delta X, \Delta Y, \Delta Z, \Delta X_0, \Delta Y_0, \Delta Z_0, \Delta\omega, \Delta\phi, \Delta\kappa, AP_i]$ is the vector of the unknowns;
- ΔX, ΔY, ΔZ are the changes to approximations of the object coordinates of a point
- $\Delta X_0 \ldots \Delta\kappa$ are the changes to approximations of exterior orientation elements
- AP_i additional parameters;

For the estimation of **x**, the Gauss–Markov model of least squares is normally used. The formed equations are non-linear with respect to the unknowns parameters and, in order to solve them with a least squares method, they must be linearized, thus requiring approximations. A first order Taylor expansion is used for the linearization and introducing a true error vector e, Eq. (11) becomes:

$$-e = A \cdot x - l, \tag{12}$$

where:
- e is the true error;
- **A** is the design ($n \times u$) matrix, of rank u, containing the partial derivatives of the n Eq. (8) with respect to the u unknowns, evaluated with the approximations;
- l is the absolute vector of the observations (observed minus approximated).

The system (12) is solved with the least squares solution:

$$\hat{x} = (A^{\mathrm{T}} P A)^{-1} A^{\mathrm{T}} P l \tag{13}$$

with P the weight matrix of the observations.

Due to the non-linear characteristic of the problem, iterations need to be performed. The residuals v of the observations and the a posteriori variance factor σ_0 are computed as shown in (14), with r the redundancy or degree of freedom (e.g., the difference between number of equations and number of unknowns):

$$v = A \cdot \hat{x} - l; \quad \hat{\sigma}_0 = \sqrt{\frac{v^{\mathrm{T}} P v}{r}}. \tag{14}$$

To invert the matrix $A^{T}PA$ in (13) an external 'datum' of the network is required; this is given by fixing the 7 parameters of a spatial similarity transformation of the network. These information is introduced with some control points (7 fixed coordinates values) or fixing 7 elements of the exterior orientation of the images.

In general, all unknown parameters of the bundle are treated as stochastic variables, allowing to consider and include a priori information about them. Not all APs can necessarily be determined from a given arrangement of images and object points. Moreover, non-determinable parameters (over-parameterization) can lead to a degradation of the results. For example, if not enough accurate control points are used, not all APs are correctly determinable or if no-rotated images are available, the position of the principal point of the camera is not reliably computed.

3.4. Results

In our application, using the process described in Section 3.1 and the images presented in Fig. 2, 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 and to recover metric results. The automatic tie points identification worked quite well even if there is an almost wide baseline between the images (ca. 18°). The initial approximations of the interior parameters are computed as described in Section 3.2 and then imported in the adjustment. After 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 (K_1) 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 $\sigma_x = 5.2$ mm, $\sigma_y = 5.4$ mm, and $\sigma_z = 6.2$ mm. The final exterior orientation of the images as well as the 3-D coordinates of the tie points are shown in Fig. 8.

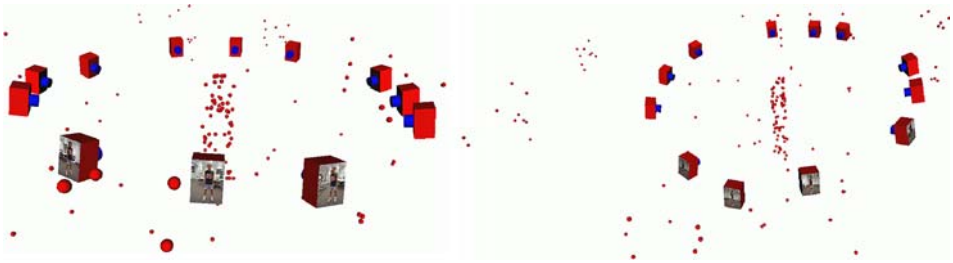

Fig. 8. The recovered metric poses of the 12 images after the adjustment and the 3-D positions of the used tie points.

Table 2
Results of the matching process: the obtained correspondences for each triplet and the number of 3-D points generated with a forward intersection

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

4. Matching process on the human body

In order to produce a dense and robust set of corresponding image points (then used to reconstruct the human shape), an automated matching process similar to [6] is used. It establishes correspondences between three images starting from few seed points and its is based on the adaptive least squares method [11]. One image is used as template and the other two as search images. The patches in the search image are modified by an affine transformation (translation, rotation, shearing, and scaling) to find the correct position of the matched point. The algorithm matches corresponding points in the neighborhood of a selected point in the search image (approximation point) by minimizing the sum of the squares differences of the gray value between the two image patches. Starting from few seed points manually selected in the three images, the matching process automatically determines a dense set of correspondences in the triplet. 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 2-D points. If the orientation parameters of the cameras are available, the geometric constraints between the images can also be used in the matching process (multi-photo geometrically constrained matching). The matching is applied to all consecutive triplets of Fig. 2 and the results are reported in Table 2.

To evaluate the quality of the matching results, different indicators are used: a posteriori standard deviation of the least squares adjustment, standard deviation of the shift in x and y directions and displacement from the start position in x and y direction. Thresholds for these values can be defined for different cases, according to the level of texture in image and to the type of template. The least squares matching process can have some problems if lacks of natural texture are presents or low-resolution images are used. The performance of the process, in case of uncalibrated images, can only be improved with some local contrast enhancement of the images (e.g., Wallis filter).

5. 3-D reconstruction and modeling of the human body shape

The 3-D coordinates of each matched triplet are then computed via forward intersection. Using the collinearity Eq. (8) and the results of the orientation process, the

3-D matched points are determined with a least square solution. If the standard deviation of the adjustment does not satisfy a certain threshold value, the triplet of points is rejected (Table 2: the number of 3-D points in each triplet of images is always smaller then the measured 2-D correspondences). For each triplet of images, a point cloud is computed and then all the points are joined together to create a unique point cloud of the human. Because of small movements of the person, the point cloud of each single triplet could appear misalign respect to the others. To avoid this possible misalignment, a 3-D conformal coordinate transformation is performed with a least squares approach. No weights are used for the coordinates. One triplet is taken as reference and all the others are transformed according to the reference one. Then, in order to reduce the noise in the 3-D data and 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 of each box is computed; the filter can then be used in two different modes:

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

Moreover, in areas with holes, a semi-automatic closure of the gaps is performed, using the curvature and density of the surrounding points.

The visualization of the obtained 3-D shape (Fig. 9, 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 Fig. 9, central.

Concerning the modeling of the recovered point cloud, the following solutions are possible:

- generation of a polygonal surface: from the unorganized 3-D data a non-standard triangulation procedure is required. It can be found in commercial packages, like

Fig. 9. 3-D point cloud of the human body imaged in Fig. 2: pre and after filtering results (ca. 20,000 points) and other views (left). Visualization of the recovered point cloud with pixel intensity (middle). RAMSIS [40] human model which can be used to model the recovered 3-D data (right).

reverse engineering software (e.g. [38,39,42]), but they need very dense point cloud to generate a correct triangulation and surface model. They also allow editing operations, like points holes filling or polygons corrections.

- fitting of predefined 3-D model of human: this procedure usually does not require the generation of a surface model and the recovered point cloud is used as basis for a fitting process [5,40]. In [5], 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 recovered 3-D data. In [40] a CAD human model (Fig. 9, 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.

6. Other examples

The presented approach to recover human body shapes from generic image sequences has been tested also on other data sets. In Fig. 10, three frames of a 6-images sequence acquired with a Sony Cybershot still digital camera are shown. A total of 148 correspondences are found in the images and then imported in the bundle adjustment. The theoretical precision of the extracted tie points on the human figure is $\sigma_X = 4.5\,\text{mm}$, $\sigma_Y = 4.8\,\text{mm}$, $\sigma_Z = 6.2\,\text{mm}$ while the standard deviation of unit weight a posteriori is $1.8\,\mu\text{m}$ (1/4 of the pixel size). The computed camera poses and 3-D coordinates of the tie points are shown in Fig. 10, right.

The matching process (Section 4) is performed with two triplets of images. The following forward intersection generated quite uniform 3-D data and after the filtering, a point cloud of ca. 8500 points is obtained (Fig. 11).

Another reconstruction is performed using a sequence acquired with a video camera handycam Sony DCR-VX700E. The artifacts created by the interlacing during the digitization process have to be removed: therefore one field of the video is deleted and then the remaining lines are interpolated. A less smooth sequence is obtained as

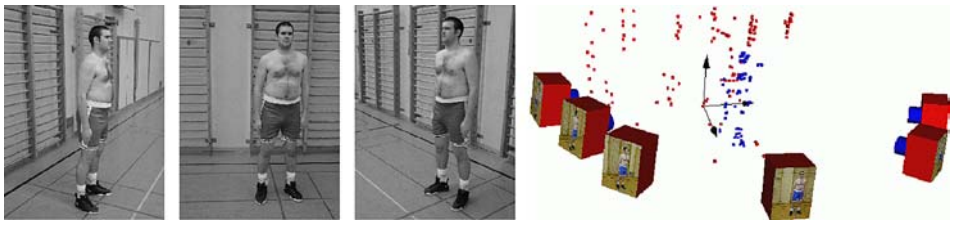

Fig. 10. Three images (out of six) used for the 3-D shape reconstruction process (left). The recovered camera poses after the bundle adjustment, performed with 148 tie points automatically extracted from the sequence (right). The slabs on the background of the person are well visible as well as the position of the person.

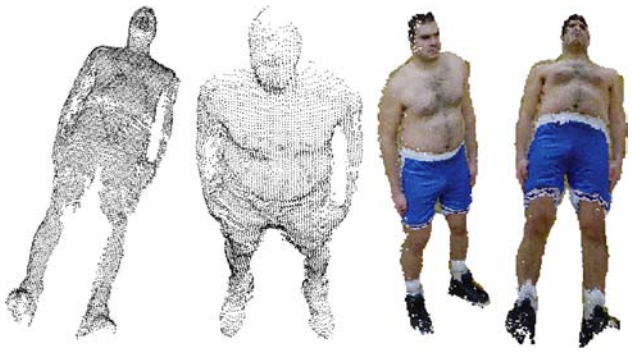

Fig. 11. Point cloud (8523 points) of the human body (left) and visualization of the results with related pixel intensity (right).

the resolution in vertical direction is reduced by 50%. Instead of removing all the odd (even) lines, another possible approach could be to remove lines just in portions of the video where interlace artifacts are present.

For the process 12 frames are selected out of a 30 s sequence acquired in front of the standing man (Fig. 12).

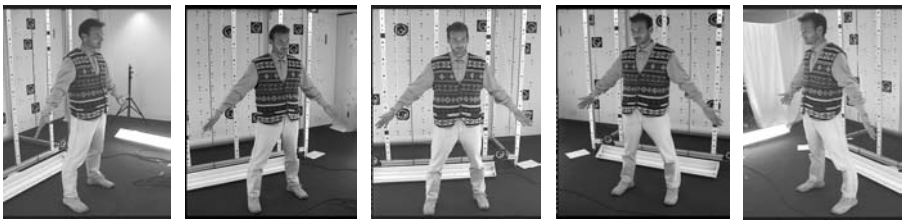

Fig. 12. Five frames (out of twelve) of the acquired video sequence. The images have a resolution of 576×720 pixel.

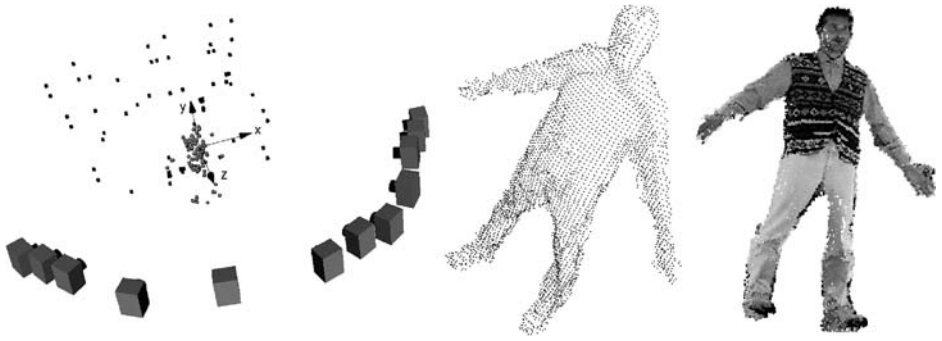

Fig. 13. Recovered camera positions and object coordinates (left). 3-D point cloud of the human body (central) and relative visualization with pixel intensity (right).

The testfield in the background contains many similar targets (repeated pattern) and they are used just as features for the processing. Moreover a repeated pattern is very good object to test the algorithm presented in Section 3.1 and verify its reliability. The parameters of the video-camera are recovered as described in Section 3. Four control points measured on the body are used for the bundle. The final configuration of the system after the adjustment is shown in Fig. 13, left.

Then the matching procedure is performed with all consecutive triplets of images and the obtained and filtered point cloud (5604 points) is shown in Fig. 13, central and right.

7. Conclusion

In this paper a method to analyze uncalibrated image sequences and create 3-D shape models of static human bodies have been presented. The main part of the work describes the calibration and orientation procedure, which can be used to recover 3-D scenes and cameras positions from different image sequences. The procedure is reliable and achieves good results not only with small baseline image sequences. A photogrammetric approach, in particular for the adjustment, was used. The presented bundle adjustment with self-calibration is a powerful tool for calibration and systematic error compensation, not always used in the vision community. In our applications, it provided for accurate orientation and location of the cameras and for accurate reconstruction of the human shape. The automated matching process on the human body recovered point clouds that can be imported in commercial software for editing, modeling or animation purposes. The full process therefore can be used to analyze old video sequences and reconstruct 3-D models of static characters who may be long dead or unavailable for 3-D reconstruction systems like body scanners.

References

[1] H.A. Beyer, Geometric and Radiometric Analysis of CCD-Cameras. Based Photogrammetric Close-Range system, Ph.D. thesis 51, IGP ETH Zurich, 1992.
[2] D.C. Brown, Close-range camera calibration, PE&RS 37 (8) (1971) 855–866.
[3] B. Caprile, V. Torre, Using vanishing point for camera calibration, Int. J. Comput. Vis. 4 (2) (1990) 127–139.
[4] R.T. Collins, Model Acquisition using stochastic projective geometry, PhD Thesis, Computer Science Dep., University of Massachusetts, 1993.
[5] N. D'Apuzzo, R. Plänkers, P. Fua, Least squares matching tracking for human body modeling, Int. Arch. Photogrammetry Rem. Sens. 33 (B5/1) (2000).
[6] N. D'Apuzzo, Modeling human faces with multi-image photogrammetry. 3-Dimensional image capture and applications V, SPIE Proc. 4661 (2002) 191–197.
[7] O. Faugeras, Q.T. Luong, et al., Camera self-calibration: theory and experiments, in: Lecture Notes in Computer Science, ECCV '92, vol. 588, Springer-Verlag, Berlin, 1992, pp. 321–334.
[8] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Comm. Assoc. Comp. Mach. 24 (6) (1981) 381–395.

[9] A. Fitzgibbon, A. Zisserman, Automatic 3D model acquisition and generation of new images from video sequences, Proc. Eur. Signal Process. Conf. (1998) 1261–1269.
[10] W. Förstner, E. Gülch, A fast operator for detection and precise location of distinct points, corners and centres of circular features, in: ISPRS Intercommission Workshop on Fast Processing of Photogrammetric Data, Interlaken, Switzerland, 1987.
[11] A. Grün, Adaptive least squares correlation: a powerful image matching technique, South Afr. J. Photogrammetry, Remote Sens. Cartogr. 14 (3) (1985) 175–187.
[12] A. Grün, H. Beyer, System calibration through self-calibration, in: Grün, Huang (Eds.), Calibration and Orientation of Cameras in Computer Vision, vol. 34, Springer, New York, 2001, pp. 163–193.
[13] A. Grün, F. Remondino, L. Zhang, Reconstruction of the Great Buddha of Bamiyan, Afghanistan, Int. Arch. Photogrammetry Rem. Sens. 34 (5) (2002) 363–368, Corfu (Greece).
[14] R.M. Haralick, L.G. Shapiro, Computer and Robot Vision, Addison-Wesley, Reading, MA, 1993.
[15] C. Harris, M. Stephens, A combined corner and edge detector, Alvey Vision Conference, 1988, pp. 147–151.
[16] A. Hilton, D. Beresfors, T. Gentils, R. Smmith, W. Sun, J. Illingworth, Whole-body modeling of people from multiview images to populate virtual worlds, in: The Visual Computer, vol. 16, Springer-Verlag, Berlin, 2000, pp. 411–436.
[17] C. Horiguchi, Body line scanner. The development of a new 3-D measurement and Reconstruction system, Int. Arch. Photogrammetry Rem. Sens. 32 (B5) (1998) 421–429.
[18] R. Klette, K. Schlüns, A. Koschan, Computer Vision: Three-dimensional data from images, Springer Press, New York, 1998.
[19] P. McKenna, Measuring up, Magazine of America's Air Force XL (2) (1996).
[20] I. Niini, Relative orientation of multiple images using projective singular correlation, Int. Arch. Photogrammetry Rem. Sens. 30 (part 3/2) (1994) 615–621.
[21] M. Pilu, Uncalibrated stereo correspondences by singular value decomposition, TR HPL-97-96, HP Bristol, 1997.
[22] M. Pollefeys, Tutorial on 3-D modeling from images, Tutorial at ECCV 2000, 2000.
[23] F. Remondino, 3-D reconstruction of articulated objects from uncalibrated images. 3-Dimensional image capture and applications V, SPIE Proc. 4661 (2002) 148–154.
[24] G. Roth, A. Whitehead, Using projective vision to find camera positions in an image sequence, 13th Vision Interface Conference, 2000.
[25] M. Scaioni, The use of least median squares for outlier rejection in automatic aerial triangulation, Proc. of 1st Int. Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS, ETH Zurich, 2001, pp. 233–238.
[26] A. Shashua, Trilinearity in visual recognition by aligment, in: J.O. Ecklund (Ed.), ECCV, Lectures Notes in Computer Science, vol. 800, Springer-Verlag, Berlin, 1994, pp. 479–484.
[27] G.B. Thomas, Calculus and Analytic Geometry, fouth ed., Addison-Wesley, Reading, MA, 1969.
[28] F.A. van den Heuvel, Vanishing point detection for architectural photogrammetry, Int. Arch. Photogrammetry Rem. Sens. 32 (part 5) (1998) 652–659.
[29] Z. Zhang, R. Deriche, et al., A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, TR 2273, INRIA, 1994.
[30] J.Y. Zheng, Acquiring 3D models from sequences of contours, IEEE Trans. Patt. Anal. Mach. Intell. 16 (2) (1994) 163–178, Web Site of System and Software:.
[31] D Studio Max 3.1. Available from http://www.discreet.com/products/3dsmax/ and related plugging http://max3d.3dluvr.com/main.html [August 2003].
[32] Animation Lab. Available from http://www.cc.gatech.edu/gvu/animation/ [August 2003].
[33] Cyberware. Available from http://www.cyberware.com [August 2003].
[34] Jack. Available from http://www.eds.com/products/plm/efactory/jack/classic_jack.shtml [August 2003].
[35] Lightwave. Available from http://www.lightwave3d.com [August 2003].
[36] Maya. Available from http://www.aliaswavefront.com/en/WhatWeDo/maya/see/solutions/soln_intro.shtml [August 2003].
[37] Photomodeler. Available from http://www.photomodeler.com [August 2003].

[38] Points2Polys. Available from http://www.paraform.com [August 2003].
[39] Polyworks. Available from http://www.innovmetric.com [August 2003].
[40] Ramsis. Available from http://www.ramsis.de [August 2003].
[41] SculpLand. Available from http://www.sanynet.ne.jp/~nakajima/SculpLand.html [August 2003].
[42] Studio Magic. Available from http://www.geomagic.com [August 2003].
[43] Taylor. Available from http://www.taylor.com [August 2003].
[44] Vitus. Available from http://www.vitus.de/english/home_en.html [August 2003].