A MULTI-VIEW TRINOCULAR SYSTEM FOR AUTOMATIC 3-D OBJECT MODELING AND RENDERING

F. Pedersini, A. Sarti, S. Tubaro
Dipartimento di Elettronica e Informazione (DEI), Politecnico di Milano
Piazza L. Da Vinci 32, 20133 Milano (Italy)
Tel: +39-2-2399-3647, Fax: +39-2-2399-3413
e-mail: pedersin/sarti/tubaro@elet.polimi.it

KEYWORDS: 3-D Reconstruction, Complete Restitution, Camera Motion Estimation, Edge Matching.

ABSTRACT:
For several applications of close-range photogrammetry, there is a growing interest in systems that are able to automatically perform a 3-D reconstruction of objects from stereo correspondences on CCD camera views. For such applications, a *full-3D* reconstruction of the object is often very desirable. In fact, most automatic systems for 3-D reconstruction based on stereo matching can only reconstruct the *front side* of the imaged scene. In order to obtain a *full*-3D reconstruction, it is necessary to observe the scene from several significant viewpoints. Furthermore, an exact determination of position and orientation of the cameras for all considered viewpoints (*camera-motion*) becomes crucial. In this paper we present a low-cost, high-accuracy, *full*-3D reconstruction system based on a calibrated set of three standard-resolution CCD cameras. No special positioning devices are needed as the camera-motion is retrieved for each position of the trinocular system from stereo-matching of *unconstrained* fiducial marks.

1. INTRODUCTION

In the past few years, research on close-range photogrammetry has focused more and more on problems of 3-D reconstruction of objects from multiple CCD-camera views. As a matter of fact, there exists variety of applications that would greatly benefit from the availability of systems that are able to perform an automatic 3-D reconstruction through stereo matching, particularly if they can provide a *complete restitution* of the imaged object, i.e. a full-3D description of its shape.
Typically, the available automatic systems for 3-D restitution based on stereo-matching can only provide a description of the *front side* of the imaged scene. In order to achieve a *complete* scene reconstruction it is necessary to observe the scene from a multitude of significant viewpoints, in which case the problem of the determination of relative position and orientation of the camera system in different viewpoints becomes crucial.
If we assume (without loss of generality) that the scene is static and the camera system is moving around it, then the above problem becomes that of the motion estimation of the acquisition system, to approach whom a variety of techniques has been proposed in literature. Such methods can be roughly divided into three main categories:

- *A-priori* solution of the motion estimation problem, by employing high-precision mechanical positioning devices for deciding the relative position between camera and scene before each image set is acquired.
- Detection and tracking of a set of *fiducial* points that are visible in the scene, and computation of the camera motion from the position of such points (and, possibly, some additional information on their position in the imaged scene).
- Hybrid solutions in which the motion of the camera system is subjected to particular constraints in order to simplify the motion estimation procedure. Such hybrid solutions usually trade computational complexity for mechanical complexity, with consequent increment of their total cost.

In this paper we introduce a low-cost 3-D restitution system, able to produce highly accurate reconstruction results. The system is based on a calibrated trinocular camera set with an arbitrary geometry. The camera system aims at an object placed on a support over which a set of fiducial points is marked. As there are no constraints on the motion of the camera system, in order to acquire a sequence of image triplets we may either move the camera set or the whole scene, without using any special positioning device.

As a first step of the *full*-3D reconstruction procedure, a partial 3D reconstruction from stereo is carried out for each triplet. This reconstruction is based on the accurate detection and matching of luminance edges. Furthermore, the 3D location of fiducial marks is determined with particularly high accuracy through subpixel point detection, stereo-matching and back-projection. Camera motion is thus determined as the rigid motion that best overlaps the sets of 3D fiducial points retrieved from each image triplet. In order to do so, the sum of the distances between corresponding points with respect to the motion parameters is used as a cost function for a minimization process, and special care is used for the selection of the starting point. All 3D edges are then merged together (by using camera motion information) in order to obtain a complete description of the object. Finally, surface interpolation and texture-mapping are performed.

2. THE SYSTEM SETUP

2.1. Camera Model

What it is normally meant with *camera model* is the

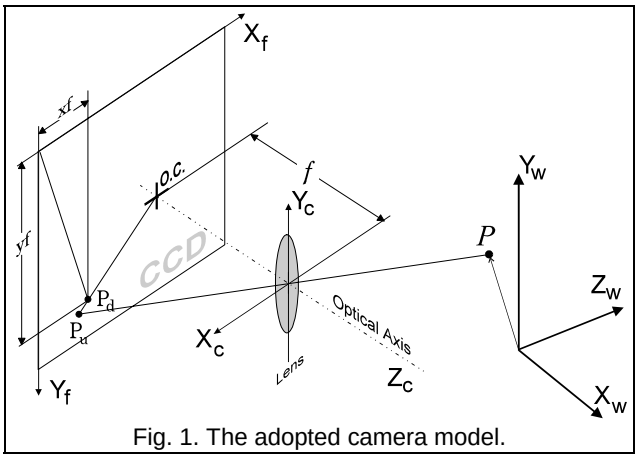

Fig. 1. The adopted camera model.

mathematical relationship between the position of a point in the three-dimensional space imaged by the camera, and the corresponding position of that point on the image plane. The scheme of the camera model that we adopted for our 3D reconstruction system is shown in Fig. 1, where three reference frames are visible:

- W*orld reference frame*, attached to the imaged scene;
- C*amera reference frame*, attached to the camera system. Notice that the Z axis is the optical axis, while the X and Y axes are parallel to the image plane.
- *Image reference frame*, where the x_f and y_f axes respectively define the horizontal and vertical directions in the digital image.

The camera model consists of a set of equations that map the 3D world-coordinates (X_w, Y_w, Z_w) of a generic point P into the 2D coordinates (x_f, y_f) of its projection onto the image plane:

- *Change of reference* frame from *world*-coordinates to *camera*-coordinates:

$$P_{cam} = \begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{bmatrix} = \mathbf{R} \cdot P_W + \mathbf{T};$$

  R and **T** being the rotation matrix and the translation vector, respectively.

- *Perspective projection* of a scene point to the image plane (the center of projection is the center of the lens and the projection plane is the camera CCD sensor):

$$p_u = \begin{bmatrix} x_u \\ y_u \end{bmatrix} = P_{cam} \cdot \frac{f}{z_{cam}}.$$

- *Lens distortion shift* of the point position p_u, predicted via perspective projection, to the actual position p_d on the image plane. When standard-resolution CCD cameras are being used, only radial distortion is normally considered. In this case, in fact, the tangential distortion can be neglected. *Radial distortion* is usually approximated by a power series:

$$r_u = r_d \cdot \left(1 + k_3 \cdot r_d^2 + k_5 \cdot r_d^4 + \ldots\right).$$

  This series can be truncated at the fifth order (only the first two coefficients are used) as the residual error results as being far below 1 pixel [4].

- *Change of coordinate frame* from *camera*-coordinates $p_d=(x_d,y_d)$. to *image*-coordinates $p_f=(x_f,y_f)$. This operation simply consists of a 2-D translation and scale change (see Fig. 1):

$$x_f = C_X + \frac{x_d}{d_X};$$

$$y_f = C_Y + \frac{y_d}{d_Y};$$

  d_x and d_y being the horizontal and vertical size of an image pixel, respectively, and (C_x,C_y) the *image*-coordinates of the optical center *OC*.

As we can see from the above description, a camera model is completely specified by a limited set of parameters. In particular, the *intrinsic* parameters (f, k_3, k_5, C_x, C_y) incorporate the physical characteristics of the camera, while the *extrinsic* parameters (**R** and **T**) define the projective geometry, and they all are estimated through an appropriate calibration procedure.

2.2. Multi-View Geometry

The multi-camera acquisition system is designed in such a way to guarantee a satisfactory camera geometry for back-projection. More precisely, the cameras are placed far enough from each other to guarantee an accurate 3D triangulation, and they are approximately pointed to a common *center* in the scene. By doing so, in fact, we maximize the 3D volume which is being simultaneously imaged by all cameras.

It is well-known that the minimum number of cameras that is required for obtaining a 3D description of the scene is two. Increasing the number of cameras, however, improves the precision and the reliability of the 3D reconstruction dramatically [6]. In particular, the introduction of a third camera has been proven to provide the most significant improvement with respect to the binocular setup, with a minimum cost increment. The results presented in this paper are thus obtained by using a *trinocular* system.

Once the cameras are properly placed, *camera calibration* is performed in order to determine the *intrinsic* and *extrinsic* parameters of the acquisition system. In order to do so, a set of *calibration target points*, whose 3D *world*-coordinates are known with good precision, is used. It is worth noticing that the camera model introduced above is used throughout the whole 3D reconstruction chain (matching and back-projection algorithms). This means that there is no need of image rectification or any constraint on the camera geometry.

2.3. The Object Acquisition Procedure

An image triplet acquired with the above camera system allows us to retrieve enough information for a 3D reconstruction of the *front side* of the scene. In order to obtain a more complete (*full*-3D) description of an object, however, it is necessary to acquire image triples from many different viewpoints, so that the whole visible surface of the object will be imaged and reconstructed. The acquisition procedure will thus consist of a series of

image triplets (*trinocular views*), each of which is taken from a different viewpoint. The viewpoint can be changed by moving either the camera system or the object.
In order to perform the estimation of the *camera motion* between different viewpoints in *world*-coordinates, the presence of some reference points (*fiducial marks*) becomes necessary. These targets are placed in the scene in such a way that the number of fiducial points that are visible in all images of two consecutive trinocular views exceeds a specified minimum. This allows us to compute the 3D position of all visible targets with respect to the *world* reference frame, and to merge the 3D information extracted from the individual trinocular views.

3. ESTIMATION OF CAMERA MOTION

Several techniques for the camera motion from point-feature correspondence are available in the literature. Most of such methods perform motion estimation from two-dimensional data by applying a *rigidity* constraint to a set of matched points on monocular views [1,2,3]. Vectors from optical centers and corresponding points on the image planes are, in fact, bound to be coplanar (*essential constraint*), which results in a scalar equation for each pair of corresponding points in different views.
As already anticipated in the Introduction, since the acquisition system consists of a calibrated set of three cameras, camera motion estimation can be performed directly in the three-dimensional space. In fact, for each trinocular view we can accurately determine the 3D co-ordinates of the fiducial points, relative to the camera frame.
As a first step, fiducial marks are located with subpixel accuracy on the image plane. Point correspondence between them is then computed by using a stereo-matching algorithm that exploits epipolar constraints for reducing the search space of correspondences and guaranteeing the absence of ambiguities. Finally, the fiducial points can be re-projected in the 3D space by using the camera calibration parameters.
Once the 3D co-ordinates of the fiducial points, relative to the camera system, are retrieved for each image triplet, we can recover the camera motion as that rigid motion that best overlaps the two sets of 3D fiducial points that are being considered. This can be done through a minimization process that uses the sum of the distances between corresponding points as a cost function.
The minimization algorithm that determines the motion parameters is nonlinear as, besides estimating the translation vector, it computes the Euler angles that best describe the rotation of the camera system. As a consequence, in order to prevent the algorithm from finding undesired local minima, it is of crucial importance operating a careful selection of the starting point. A sufficiently accurate estimate of the camera motion can be obtained through a linear least square algorithm, provided we adopt an affine representation of the rigid motion itself (translation vector and rotation matrix). One should keep in mind, however, that a rotation matrix represents a super-parametrization of a rigid rotation (3x3 matrices are used for describing elements of the three-dimensional rotation manifold $SO(3)$), therefore the linear minimization process generally returns matrices that do not satisfy the orthogonality constraint. By projecting the estimated matrix onto $SO(3)$, however, we obtain a rotation matrix that is accurate enough to be safely used as a starting point for the non-linear minimization process.

4. SCENE RECONSTRUCTION

The scene reconstruction procedure is divided into the following steps:
a) Camera setup and calibration;
b) Estimation of 3D edges for each triplet;
c) 3D localization of the fiducial points for each triplet;
d) Camera motion estimation and conversion of all 3D edges into *world*-coordinates;
e) 3D surface interpolation.
After camera setup and calibration, several trinocular views of the scene are acquired from different viewing directions (see, for example, Figs. 2 and 5). In the examples presented in this paper, the change of viewpoint is obtained by moving object and support.

Reconstruction of 3-D edges: For each trinocular view, a 3D reconstruction of luminance edges is performed. This is done through detection, matching and back-projection of all visible edges of the scene. Luminance edges are detected by using an optimized version of Canny's edge detector [7]. The detected edges are then passed to an *edge selector*, in order to keep only those that carry a significant information (e.g. edges that too short are discarded) and labeled. For each labeled edge, the stereo-corresponding *(homologous)* ones in the other two images are searched on the epipolar space. Notice that, as the radial distortion is taken into account, the epipolar lines are actually represented by curves.
Using more than two views dramatically speeds up the search of homologous edges. Moreover, matching ambiguities, typical of binocular systems, are overcome with a proper placement of the third camera.
Due to a different fragmentation of the same luminance edge in different images, it may happen that a single edge in one image needs to be matched to several edges in the others. For this reason, not only is the proposed edge matching algorithm capable of finding "one-to-one" correspondences, but it can also handle correspondences between *subsets of edges* that are portions of the same fragmented one.
Once the trinocular edge matching is completed, each edge triplet is back-projected onto the 3D scene space. In order to do so, each edge is first approximated by a chain of line segments (the desired level of accuracy can be decided by adjusting the average segment length). For each representative edge triplet, the back-projected point in the 3D space is determined by selecting the closest point to the lines that pass through the optical center and the edge point of each camera. The procedure returns a list of 3D edges described by their representative 3D points. All 3D edges, of course, are relative to the reference frame of its corresponding trinocular image.

3-D localization of reference marks: for each trinocular view, all visible fiducial marks are located with subpixel accuracy. A point-matching is then performed over such points and, wherever a matching is found, the back-projected point is determined. By doing so, we obtain, for each trinocular view, the 3D *camera*-coordinates of a subset of the fiducial marks. Some *a-priori* knowledge on the relative position of the targets in the scene will help identifying and labeling them. After labeling them, the reference points can be matched throughout different

triplets, thus allowing us to compute the camera motion between them. As described above, the rigid motion that best overlaps targets of different triplets is taken as the relative motion of the camera system from one triplet to another. This operation is carried out for all the consecutive pairs of views. By using camera motion, the coordinates of all 3D edges can now be converted into a common reference frame. At this point, if camera motion has been accurately estimated, a simple merging of 3D edges obtained from each triplet provides a complete description of the observed object.

3-D surface interpolation: In some cases it is highly desirable to obtain a 3D model whose shape is described by a closed surface rather than by edges. Moreover, for applications like image synthesis or virtual reality, there is a need for 3D models where besides the shape, also the original pictorial information on the surface (texture) is recovered. For this reason, the last processing step, is the construction of a surface that, by passing through all edges, approximates the object surface. The 3D surface is obtained using an optimized surface interpolation technique which, in fact, is a discretization of the *thin-plate spline algorithm* (*Discrete Smooth Interpolation* [5]). This technique allows the presence of local discontinuities in the interpolated surface, while performing a spline-like interpolation on smooth surface regions. This technique is particularly suitable for interpolation of 3D shape information, being the shape information typically characterized by edges and depth discontinuities (i.e. at object borders).
The operation of recovering the original pictorial information (*texture*) is done through a back-projection of the luminance information associated to the original images (from the original viewpoint to the scene space). Roughly speaking, the images are projected on the interpolate surface in a similar way as a "slide projector" would do it. In order to obtain good quality results from this *texture mapping* operation, particular care must be used in compensating the different conditions of illumination for the different viewpoints of the original images. The quality of the texture mapping is also affected by the quality of camera calibration and camera motion estimation, which normally causes undesirable errors in overlapping the texture from different viewpoints.

5. EXPERIMENTAL RESULTS

Some examples of application of the system are presented in this paper. Two sample objects, a fish-shaped hand-crafted object and a toy train engine, have been used to test the proposed *full*-3D reconstruction procedure. Each object has been placed on a low-cost moving support in front of the trinocular camera system. The cameras are placed at the vertices of a triangle in order to avoid matching ambiguities and to guarantee favorable conditions in the relative epipolar geometry. Figures 2 and 5 show, for each object, one of the original images taken by the camera system.
Figure 3 shows, for the object "train", a view of the 3D edges localized in one trinocular shot. Thanks to the accuracy of the 3D edge reconstruction and camera calibration algorithms, with this technique it has been possible to achieve a relative accuracy of 200/300 ppm in the 3D location of edges.
Figures 4 and 6 show the obtained reconstruction after merging the 3D edges obtained from all trinocular views. As we can see, edges from different triplets merge in a very precise fashion, which confirms the quality of the *camera motion* estimation. The maximum diameter of the bundles of homologous edges results as being smaller than 0.5 mm, which corresponds to a relative precision of 300/400 ppm.
Figure 7 shows, for the object "fish", a synthetic perspective view of the reconstructed surface of the object, where the pictorial information has been mapped from the original images through texture-mapping. The fidelity of the rendering and the sharpness of the projected texture prove the good quality of the proposed texture-mapping procedure.

6. CONCLUSIONS

The experimental results have shown that, in spite of the low cost of the system, the achieved level of accuracy is quite high. In fact, consider just one trinocular view, the 3D co-ordinates of visible sharp edges can be computed with a precision of about 200/300 ppm. When considering a series of trinocular views for a complete reconstruction of the scene, the accuracy remains nearly unchanged (300/400 ppm), which emphasizes the quality of the camera motion estimate.

Further improvements of the proposed reconstruction method are currently under development, especially those related to the 3D edge-merging process and the problem of "full-3D" interpolation of surfaces of complex volumes. In particular, we are focusing on the integration of volumetric reconstruction methods and the above technique.

REFERENCES:

[1] C. Braccini, G. Gambardella, A. Grattarola, S. Zappatore: "Motion estimation of rigid bodies: effects of the rigidity constraints." *EURASIP, 1986. Signal Processing III: theories and applications*. pp. 645-648.

[2] T.S. Huang, O.D. Faugeras: "Some properties of the *E* matrix in two-view motion estimation." *IEEE Trans on Pattern Analysis and Machine Intelligence*. Vol. 11, No. 12, Dec. 1989, pp. 1310-1312.

[3] S. Soatto, R. Frezza, P. Perona: "Motion estimation on the Essential Manifold." In: *Computer Vision - ECCV '94. Third European Conference on Computer Vision.* Proceedings. Vol.II., Stockholm, Sweden, 2-6, May 1994. pp. 61-72.

[4] Tsai, R.Y., 1987. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using off-the-shelf TV Cameras and Lenses. IEEE Journal on Robotics and Automation, Vol. RA-3, No. 4, pp. 323-344.

[5] Mallet, J.L., 1989. Discrete Smooth Interpolation. ACM Transactions on Graphics, vol.8, N.2, pp.121-144.

[6] Ayache, N., 1991. Artificial vision for Mobile Robots. MIT Press.

[7] Canny, J., 1986. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.8, No.6, pp.679-698.

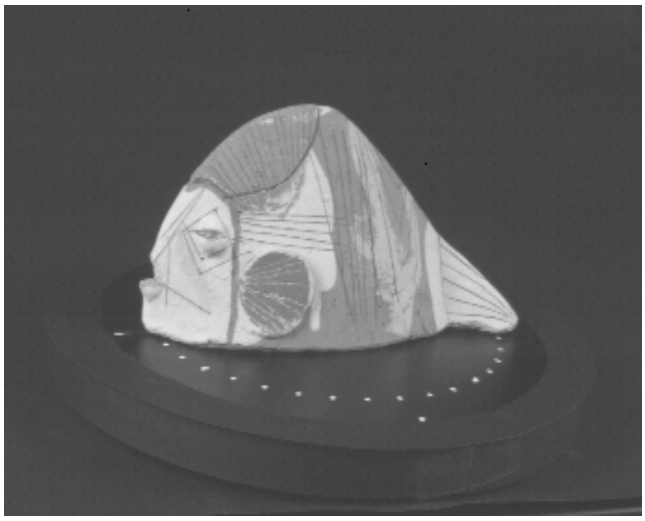

Figure 2. One of the images of the acquisition sequence of the scene “FISH” (left camera).

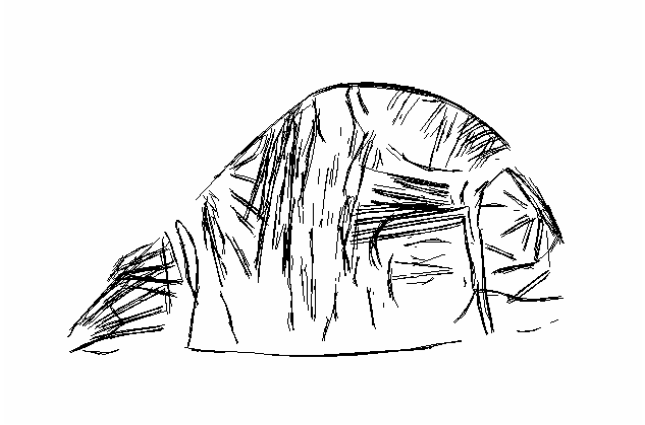

Figure 3. Scene “FISH”: Merging of the 3-D edges obtained from all the views.

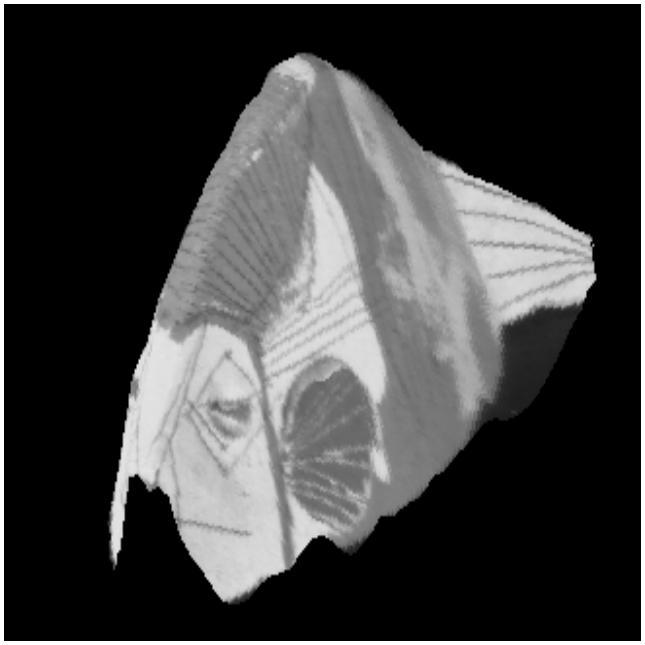

Figure 4. Scene “FISH”: 3-D surface interpolation and restitution of the original luminance texture.

Figure 5. One of the images of the acquisition sequence of the scene “TRAIN” (left camera).

Figure 6. Scene “TRAIN”: The 3-D edges obtained from one trinocular shot.

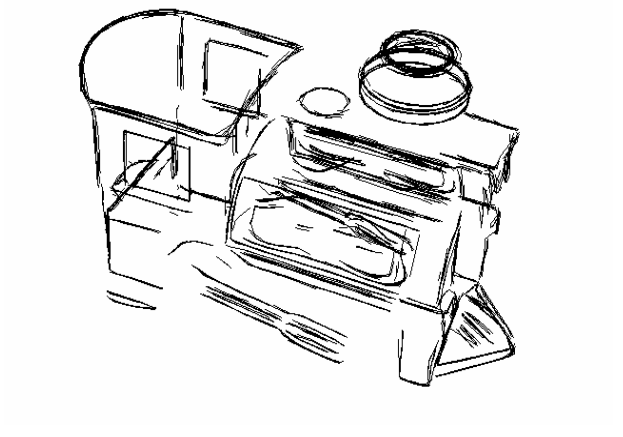

Figure 7. Scene “TRAIN”: Merging of the 3-D edges obtained from all the views.