Multi-Camera Acquisitions for High-Accuracy 3D Reconstruction

Federico Pedersini, Augusto Sarti, Stefano Tubaro

Dip. di Elettronica e Informazione (DEI), Politecnico di Milano
Piazza L. Da Vinci 32, 20133 Milano, Italy
E-mail: pedersin/sarti/tubaro@elet.polimi.it

Abstract. In this paper we present our global approach to accurate 3D reconstruction with a calibrated multi-camera system. In particular, we illustrate a simple and effective adaptive technique for the self-calibration of CCD-based multi-camera acquisition systems. We also propose a general and robust approach to the problem of close-range partial 3D reconstruction of objects from stereo-correspondences. Finally, we introduce a method for performing an accurate patchworking of the partial reconstructions, based on 3D feature matching.

1. Introduction

In the past few years, there has been a fast proliferation of methods for the 3D reconstruction of objects from the analysis of camera images. A large number of these applications are aimed at the problem of *content creation* for the market of multi-media applications. There is a considerable number of applications, however, in which the accuracy of the 3D reconstruction plays a crucial role. For example, applications of close-range digital photogrammetry aimed at the preservation and restoration of 3D works of art require effective methods for accurate, quantitative, reproducible and repeatable 3D reconstruction. In this case, in fact, suitable 3D modeling methods should be sufficiently accurate as to match the performance of the methods that are commonly adopted for the 3D relief of works of art; and to guarantee that such measurements will be reproducible and can be repeated along time for monitoring purposes.

The most popular non-invasive approaches to 3D reconstruction of mid-sized objects are based on stereo-correspondences. Such methods are based on the detection of features (e.g. points, edges, luminance profiles) on the available images of the object. When the camera parameters (position, orientation and other intrinsic physical parameters) are known (*calibrated* case), the process of determining the correspondences is helped by some rigidity constraints such as the coplanarity of corresponding visual rays (epipolar constraint), and the 3D coordinates of the features can be determined through geometric triangulation [1,2]. When, on the contrary, the camera parameters are not known (*uncalibrated* analysis), the determination of the feature correspondences becomes more complex as it can only rely on projective constraints and invariants. Several robust matching techniques have been developed for uncalibrated acquisitions [19]; such methods are usually based on the progressive application of a variety of projective constraints on sets of uncalibrated views, in order to narrow-down a list of candidate matches (generated through a correlation-based approach) to a final set of confidence matches.

Reinhard Koch, Luc Van Gool (Eds.): SMILE '98, LNCS 1506, pp. 124-138, 1998.

In general, the 3D reconstruction methods based on feature matching can be classified into two categories:

- *monocular approach*: a number of uncalibrated views are acquired and processed all together (*global* approach) or in subgroups (*local* approach) in order to jointly estimate camera motion and object structure. In the global approach, one or more cameras are employed for acquiring a number of images of the object from a variety of viewpoints [6]. The pose of the cameras and the 3D coordinates of the features are found through a joint analysis of the image features extracted from all the available views. In the local approach a video sequence of the object is acquired in such a way to "cover" all portions of the object. Then the views are partitioned into subgroups to be processed separately using uncalibrated methods based on projective invariants and constraints.
- *multi-ocular approach*: a set of cameras is mounted on a rigid support and calibrated, so that all camera parameters are known beforehand. Several multi-ocular views of the object are acquired from a variety of viewpoints. From the analysis of each multi-view a "local" surface is generated. All local surfaces are then fused together into a single one, by using some global constraints [6,7,8].

In general, the global monocular approach estimates the 3D coordinates of some object features with best accuracy. Due to its global treatment of the data, however, it tends to produce a sparse set of 3D features that cannot be easily interpolated into a global surface unless some *a-priori* information on the object is available. Partitioning the views into smaller groups for a more "local" approach makes it easier to deal with the complexity of the surface topology but generally causes a reduction of the accuracy and is quite difficult to perform on an automatic basis. This partitioning becomes necessary when dealing with video sequences, but the subgroups of views tend to be "aligned" with each other, which is not the optimal positioning for feature matching purposes. On the other hand, acquiring a monocular sequence is certainly the simplest way to perform an acquisition campaign.

The local multi-camera approach, exhibits some interesting characteristics:

- a multi-camera acquisition system induces a "natural" partition of the views, which becomes optimal when the cameras are well-positioned on the rigid frame;
- the acquisition system can be quite easily calibrated, and the estimated parameters can be used for validating all feature matches; the calibration can be made adaptive in order to compensate for the drift of the parameters throughout the acquisition process;
- the accuracy of a well-calibrated system is at photogrammetric level;
- each calibrated triplet generates a "local" surface *patch* of modest topological complexity; all patches can be merged into a more complex global surface through "patchworking".

In this article we illustrate our calibrated reconstruction approach based on adaptive self-calibration, local stereo-matching approach and global patchworking, with the goal of obtaining a high-accuracy reconstruction of unstructured 3D objects.

2. Calibration

All calibrated 3D reconstruction methods are critically dependent on the accuracy with which the camera parameters, i.e. the geometrical, optical and electric characteristics of the camera system (camera position and orientation, focal length, pixel size, location of the optical center, nonlinear distortion coefficients, etc.) are known. In the past few years several approaches to the calibration problem have been proposed. Such methods apply to electronic cameras the same techniques that were traditionally used for the calibration of photogrammetric cameras [9,10,11]. The camera characteristics are, in fact, computed through a proper processing of the image of a test object (calibration target-frame) placed in the scene. The accuracy of the camera model can be arbitrarily improved by employing an adequate number of parameters therefore, when the goal is that of improving the calibration accuracy as much as possible, the pattern accuracy becomes the major bottleneck. For this reason, we developed an advanced photogrammetric method that jointly estimate the camera parameters and the geometry of the calibration target-set in a more accurate fashion (*self-calibration*). This method is based on a *multi-camera, multi-view* calibration approach, and performs an accurate self-calibration on the multi-camera system from the analysis of several views of a simpler calibration target-frame, such as a marked planar surface (a printed sheet of paper glued on a glass surface) or some other even simpler structure. In fact, not only is this technique able to estimate the camera parameters, but it can also determine the 3D position of the targets on the calibration frame, which can be just roughly known or not known at all. Finally, we developed method for making the calibration robust against the inevitable parameter drift that takes place during the acquisition process. Such method detects and tracks some "safe" features that are naturally present in the scene, and use their image coordinates for making the calibration process adaptive.

2.1 Calibration Strategy

The camera model allows us to compute the image coordinates of the projection of an object point *P*, given its coordinates. This model can also be seen as a function *g* that maps a point *m* in the model space into a point *d* in the observation space. This set of equations is called *direct model*. For each fiducial point P_i and for each camera C_j, the direct model provides us image coordinates in the form

$$d_{i,j} = g\left(P_i\ , C_j\right) = g\left(m_{i,j}\right); \qquad d_{i,j} = \left[x_i \quad y_i\right]; \qquad m_{i,j} = \left[P_i \quad C_j\right];$$

where C_j is the set of the 11 parameters which define the model of the *j*-th camera

$$C_j = \left[\phi, \theta, \psi,\ t_X, t_Y, t_Z,\ f,\ k_3, k_5,\ c_X, c_Y\right],$$

ϕ, θ and ψ are the 3 angles that characterize the rotation matrix $\boldsymbol{R}$, $\boldsymbol{T}=\{t_X,t_Y,t_Z\}$ is the translation vector, f is the focal length of the lens, k_3 and k_5 are the radial lens distortion coefficients, $OC=\{c_X,c_Y\}$ is the optical center on the image plane. $\boldsymbol{R}$ and $\boldsymbol{T}$ are usually referred to as *extrinsic parameters*, while the other five are called *intrinsic parameters*, as they characterize the camera. Rewriting the above equations in matrix form and extending them to all the considered fiducial points and all the cameras, the direct model becomes

$$\begin{array}{rccc} g(\cdot): & \Re^{3N+11\cdot V} & \rightarrow & \Re^{2N\cdot V} \\ & \mathbf{m} & \mapsto & \overline{\mathbf{d}} = g(\mathbf{m}) \end{array}$$

$g(\cdot)$ being the nonlinear function that maps the 3D world coordinates of N fiducial points, given the camera parameters of V cameras, into the V sets of N two-dimensional image coordinates of the perspective projection in each camera.
With this formalization of the camera model, the problem of camera calibration becomes that of computing the model vector **m** by exploiting the knowledge about the observed data **d** and the direct model $g(\cdot)$. In other words, the solution of the problem is given by the inverse of the model function $\mathbf{m}=g^{-1}(\mathbf{d})$, where **d** and g() are known. This corresponds to the classical formulation of an inverse problem. In this sense, the camera calibration problem is a typical inverse problem.

Multi-view, multi-camera calibration - It is well-known that, in order to obtain satisfactory and reliable calibration results, it is necessary for the fiducial points to "fill up" the entire scene space. This fact usually forces to construct 3D calibration patterns that are as large as the object to be reconstructed. In this case, calibration is possible only for scenes of limited spatial extension. When scenes of larger size are considered, 3D pattern cannot be employed anymore. In fact, an accurate 3D pattern is very expensive to build and quite cumbersome to handle. For this reason a *multi-view* calibration set-up has been developed, that allows us to generate the desired 3D set of fiducial points through multiple acquisition of a smaller and simpler calibration pattern, such as a planar target. In fact, the pattern can be placed in several different positions, so that the entire 3D volume of interest will be "scanned". The relative motion between different positions of the pattern is not measured, therefore the position of the fiducial points in the scene is only partially known. This, of course, complicates the calibration problem, as the relative motion of the pattern must be estimated. In fact, we have six extra unknowns per additional target position. However, as the motion of the pattern from view to view, is the same for all the cameras, each camera will give its contribution to the estimation of the pattern motion. In fact, with respect to the case of calibrating one camera with V pattern views, each new camera to calibrate adds 11 more unknowns, while providing approx. $2NV$ equations (the image coordinates of NV fiducial points).

Self-calibration - Thanks to the large number of constraints and to the fact that, through multiple pattern positioning, the fiducial points end up covering the whole scene space, the self-calibrating approach can lead to the best results that can be obtained with such low-cost calibration setups, in terms of global accuracy throughout the scene space. In fact, experimental results have shown that, under proper conditions, the achieved calibration accuracy, with this approach, reaches the limit imposed by the accuracy with which the position of the fiducial points is known, with respect to the pattern frame. In order to further improve the accuracy of the calibration, it is either necessary to use calibration patterns of higher precision, for which the fiducial point coordinates have been determined with high accuracy (e.g. with photogrammetric techniques) or to adopt a self-calibrating approach, which, besides estimating the camera parameters, refines the estimates of the a-priori given coordinates of the fiducial points. The complexity of the former solution is the same

as in the previous case, but it requires expensive calibration patterns. As the aim of this work to obtain high performance at low cost, we focused on the latter solution.
The self-calibration problem is much more undetermined than the previously considered one, because the calibration points coordinates *WP* are considered only approximately known. In other words, also the data points become, to some extent, unknowns to be estimated. The a-priori knowledge about the data generally consists of a rough estimate of the world-coordinates. The proposed technique is able not only to further improve the accuracy of the estimated camera parameters, but to refine the a-priori given estimate of the world coordinates as well.

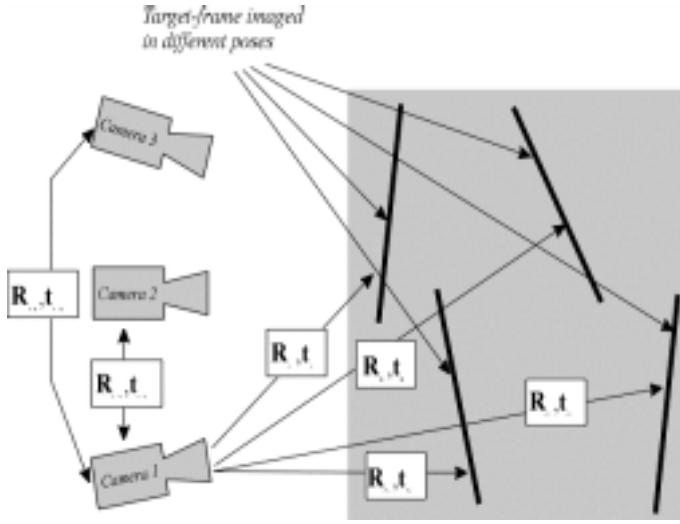

Fig. 1. General scheme for the multi-view multi-camera approach to self-calibration

The calibration target-set that we adopted is planar as the pixel size is assumed known [9]. A planar target-set is much simpler to build compared to a 3D target-frame as it can be easily constructed, for example, by gluing a laser-printed sheet of paper on a rigid planar surface. This procedure also gives us some *a-priori* information on the coordinates of the targets (and their uncertainty), relative to a frame attached to the surface. A 3D calibration target-frame, on the other hand, would require an accurate 3D measurement of the coordinates of the targets (generally through some photogrammetric technique [11]). An example of application of our self-calibration approach is reported in Fig. 2.

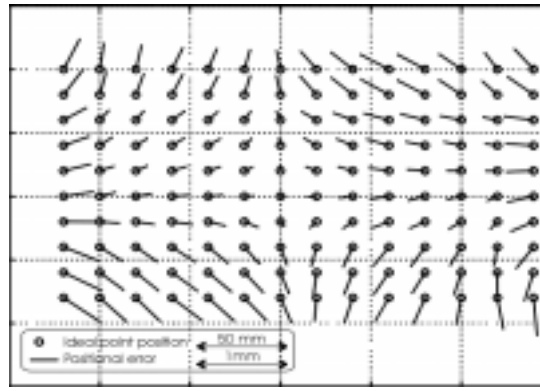

Fig. 2. *A-priori* coordinates of the fiducial points of the target-set (laser-printed circles on a sheet of A4 paper, glued to a flat surface) and corresponding *a-posteriori* corrections estimated through self-calibration. The orientation of the (magnified) correction vectors denotes the deformation of the sheet of paper due to the action of the dragging mechanism of the laser printer.

2.2 Adaptive Calibration

In order to extract 3D information from the scene views the *camera parameters* must be known with good accuracy throughout the whole acquisition campaign. As *camera calibration* is performed before the beginning of an acquisition session, problems of parameter drift may occur. In fact, when long video sequences are acquired, the stability of the camera parameters measured at the beginning becomes a crucial problem as mechanical shocks, vibrations or thermal effects on cameras and supports, can cause small variations of the initial camera set-up. This drift of the camera parameters leads to significant 3D reconstruction errors, as the 3D back-projection is rather ill-conditioned with respect to the camera parameters. In order to overcome this problem, we detect and track any changes in the acquisition system and, whenever possible, we apply an on-the-fly correction of the camera parameters. By doing so, the calibration holds accurate throughout the acquisition campaign.
Our approach does not require us to place targets in the scene or to use any *a-priori* knowledge, but exploits luminance features that are already present in the scene (e.g. corners and spots) which can be located in the image with high precision. After the localization process, which is performed with sub-pixel accuracy, a matching operation is performed among the n sets (n being the number of cameras) of feature points, which returns a set of n-tuples of homologous points. The matched n-tuples will be then back-projected into the 3D scene space. If the camera parameters change, then the back-projection will be affected by larger errors, with respect to the predicted pre-calibration accuracy. A proper analysis of the magnitude and the temporal changes of the back-projection error allows us to reveal and characterize any incidental modifications of the camera parameters. Furthermore, if the set of matched

n-tuples is informative enough, the proposed technique allows to accurately measure the occurred modification and, therefore, to re-calibrate the system.
Our approach can be seen as composed of two main steps

- check on the validity of the current camera parameters through the estimation of the back-projection's accuracy
- analysis of the temporal changes of the back-projection's accuracy, in order to reveal increments in the reconstruction error that could likely denote a change in the system parameter

The first step of the algorithm consists in the detection of the significant image features that will be used as control points. Our method is based on the techniques presented in [13,14,15]. In order to obtain super-resolution in the image localization accuracy, an algorithm for the local modeling of the image Laplacian function has been developed and employed in the localization procedure. The obtained results show that the introduced improvements has allowed to reach a localization accuracy better than 0.2 pixel [18].
Over the obtained sets of image points, an *n*-partite matching algorithm is applied, in order to find the stereo-corresponding *n*-tuples. The matching criterion is based not only on the epipolar geometry defined by the current calibration, as the calibration is not considered as reliable in this application, but also on the similarity of the local luminance profiles. All the matched *n*-tuples are then back-projected in the 3D scene space, and an "*accuracy index*" is computed for each match, based on the back-projection error. The statistical distribution of this index over the matched points and its temporal behavior are then analyzed, in order to reveal any increment of the accuracy index that could very likely denote a change in the system parameters. Moreover, at the beginning of the sequence, the back-projected points that are most accurate and are fixed in the scene are selected as *control points*. These are the points that could then be used as 3D fiducial points for the re-calibration of the system. In fact, if the number of matched points is sufficient, it is possible to perform a reliable re-calibration of the system. When a change in the camera system has been detected, the current set of matched *n*-tuples of image features is exploited, in order to recover the new camera parameters.
Assuming that the camera system is not subjected to a rigid motion with respect to the scene throughout the acquisition session, at the beginning of the sequence the most accurate and stable (fixed) back-projected points are detected and used as *control points*. These are the candidate points to be used as 3D targets for parameter correction, provided that their number is sufficient. When a parameter change is detected, the current set of matched *n*-tuples of image features is used for recovering the new camera parameters.
Depending on the previous knowledge of the 3D position of the matched points, the algorithm adopts either a calibration or a self-calibration approach. More precisely, if the 3D position of some points had been measured at the beginning of the sequence when the system was still calibrated, then re-calibration is performed through a standard procedure that uses the available 3D points as markers. If, on the contrary, no reliable information is available about the actual 3D position of the matched points, the calibration can only be corrected through a self-calibration procedure. Self-

calibration allows to simultaneously determine the camera parameters and the 3D position of the fiducial points.
This method, however, requires a larger number of matched points for accurate results, as the self-calibration problem is much more ill-conditioned than the calibration problem. We are currently working on a modified version of the method that is able to determine a *rigid* (rather than *fixed*) set of points and perform calibration with respect to a relative (rather than absolute) frame.
The proposed technique has been tested on real sequences acquired with different trinocular camera systems, with both simulated and real variations of the camera parameters. In all experimental situations, the algorithm has been able to detect the modification of the camera parameters. Moreover, after artificial modifications of the camera system, of the same characteristics and entity of accidental ones (artificial shocks, change of focal length, etc.), the algorithm has been able to measure the drift of the parameters, thus allowing the re-calibration of the system. The results have shown that the accuracy of the re-calibration, in all cases, has reached the same accuracy of the original calibration.

3. Partial Reconstruction

Our approach to local reconstruction is based on feature correspondence. Image features that are most often used for 3D reconstruction are points, luminance edges and luminance profiles. Such features tend to provide information of different nature. Point and edge matching is generally a very precise and reliable process, but it usually results in a sparse set of 3D data. Conversely, matching the luminance profiles of small areas tends to generate a much denser set of 3D data but it is rather sensitive to the unavoidable viewer-dependent perspective/radiometric distortions, therefore this approach tends to be less stable and reliable. For this reason we developed a general and robust solution to the problem of 3D reconstruction from stereo correspondence of luminance patches. The method is largely independent on the camera geometry, and employs a calibrated set of three or more standard TV-resolution CCD cameras, which provides enough redundancy for removing possible matching ambiguities. The robustness of the approach can also be attributed to the *physicality* of the matching process, which is actually performed in the 3D space rather than on the image plane. In order to do so, besides the 3D location of the surface patches, it estimates their local orientation in 3D space as well, so that the geometric distortion of the luminance patch can be included in the model. Finally, the method takes into account the viewer-dependent radiometric distortion.

3.1 Edge-Based Approach

As a preliminary operation, we perform partial reconstruction from edge matching, in order to obtain reliable and accurate 3D data to begin with. The same type of features will later be used for egomotion estimation as well (which is based on 3D contour matching in object space). In order to be able to use edges for accurate egomotion estimation, however, we need to detect them with great accuracy. We do this by first

using a traditional edge detector, we then retrieve the subpixel location of the edge points through an interpolation process which takes the luminance gradient into account. Finally, a rule-based contour tracking method is employed for determining the correct connection between edge points.
The search for homologous edges on different views is performed along *epipolar lines*. Notice that using more than two cameras allows us to avoid problems of matching ambiguity. With three cameras, in fact, not only can we always select the best pair of views for a specific stereo-correspondence (sharp intersection between edge and epipolar lines), but we can validate the match through a check on the third view. In fact, the edge point must lie on the intersection of the two epipolar lines associated to the homologous edge points on the other views. Once the matches are found, each set of corresponding contours is back-projected onto the 3D scene space by looking for the point at minimum distance from the three homologous visual rays.

3.2 Area-Based Approach

The luminance patches used by most area-matching techniques are normally assumed to have the same shape in all views. It is quite clear, however, that this hypothesis is acceptable only when the angles between the viewing directions of the three cameras are not too wide, which is not our case. As a consequence, we need to take into account the perspective distortion of the shape of the patch, when back-projected onto the object surface and then re-projected onto the other image plane. In order to do so, we assume the 3D surface to be locally flat, which means that it can be approximated by a plane within the back-projected surface patch.
In the other view we search, along the distorted (due to radial distortion) epipolar line, for the patch that best matches the first one. The projective distortion of the patch is accounted for by estimating, both position and orientation of the patch. In practice, the minimum of a *similarity function* between a patch of the actual image and a re-projected patch after perspective warping is searched for as a function of position and local orientation of the tangent plane of the object surface.
Area matching requires a comparison between the actual luminance profile of a patch with the one that we obtain by *transferring* luminance profiles of other views through a specific 3D surface model. Let S be a surface patch in object space, obtained by back-projecting a reference image patch of any of the views onto the plane $\mathbf{s}^{\mathrm{T}}\mathbf{x}=0$, and let $\mathsf{S}^{(i)}$ be its *i*-th view. The transfer of projective coordinates from the *j*-th view to the *i*-th view through the plane $\mathbf{s}^{\mathrm{T}}\mathbf{x}=0$, can be expressed as a homography (an invertible linear projective transformation) of the form

$$\mathbf{u}^{(i)} = \mathbf{M}_{ij}(\mathbf{s})\mathbf{u}^{(j)} = 0 \ ,$$

where $\mathbf{M}_{ij}(\mathbf{s})$ is a 3 by 3 matrix which depends on the parameters of the plane over which the patch lies. This homography allows us to express the luminance transfer from the *j*-th view to the *i*-th view as

$$I_j^{(i)}(\mathbf{u}^{(i)}) = g_j^{(i)} I^{(i)}(\mathbf{M}_{ji}(\mathbf{s})\mathbf{u}^{(i)}) + \Delta_j^{(i)}$$

where $g^{(i)}_j$ is a correction factor (*gain*) that accounts for electrical differences in the camera sensors, while $\Delta^{(i)}_j$ is an additive radiometric correction (*offset*) which accounts for non-Lambertian effects of the surface reflectivity (reflection's migration with the viewpoint). Notice that the Lambertian component of the surface reflectivity does not appear in the above expression as it is the same for all views.

If a reference patch produces reliable 3D information, then it can be used for 3D surface reconstruction. Once all reference regions have been considered, surface interpolation is carried out and the area matching process can start over with a smaller patch size. In this case the previously estimated surface can be used for initializing the search in the next step and speeding up the process.
As a general rule, we need to make sure that the maximum size of the patch is small enough to guarantee a limited matching error. On the other hand, we know that the area matching process is based on the minimization of a highly nonlinear similarity function, therefore we can expect the process to be quite sensitive to local minima. In order to avoid such a problem, we can use an initial *guess* of the surface shape, which helps the minimization process converge to a global minimum and dramatically speeds up the matching process by reducing the size of the search space. In principle,

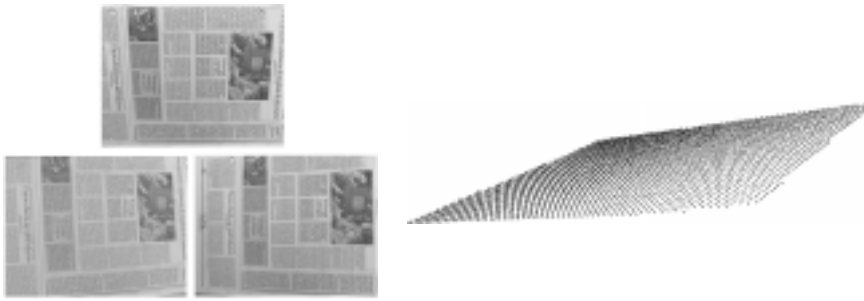

Fig. 3. Original views of the newspaper's page (glued to a planar surface) and 3D points reconstructed through area matching.

any method can be used for obtaining the initial surface. In our case we used the surface obtained through edge-matching [1,2], whose reliability is guaranteed by the accuracy of the camera model and the calibration procedure. As the result of the edge-matching is usually a sparse, though accurate, set of 3D points, such data is interpolated in order to generate the initial surface. We interpolated the 3D data by means of a modified and optimized version of the edge-preserving Discrete Smooth Interpolator (DSI) [16].
Some experiments of 3D scene reconstruction have been carried out on several test scenes. The first test presented in this paper concerns the measurement of the accuracy of the area matching using a flat object placed at about 1.2 m of distance from the camera system. The surface reconstruction resulted to be flat with 0.1 mm of standard deviation (see Fig. 3).

A second experiment concerned the 3D reconstruction of a tele-conferencing scene. The acquisition was made with a trinocular camera system at CCETT, France, within the ACTS "PANORAMA" Project (see Fig.4). No initial reconstruction was used for area matching. Instead, progressive-scan initialization was performed. The results of the area matching procedure are visible in Fig. 4. As we can see, the quality of the

reconstruction greatly benefits from the fact that the geometric distortion is included in the model.

Fig. 4. One of the three original views of a conference scene (above) and two virtual views of the 3D points reconstructed through 3D area matching (below).

4. Motion Estimation through Line Matching

In order to be able to merge 3D data coming from different partial reconstructions, we need to accurately estimate the rigid motion that the acquisition system undergoes between two multi-view acquisitions. In order to do so, one could employ high-precision mechanical devices for positioning the camera system (or the object) before acquiring a multi-view. This *a-priori* solution of the ego-motion problem, however, is usually quite expensive and not very flexible. In alternative, one can perform detection and tracking of some image features throughout the acquisition process, and use the location of such features for estimating the camera motion. This last approach becomes particularly interesting when the features to be extracted are part of the scene to be reconstructed rather than being artificially added to it. Adding special *markers* to the imaged scene is, in fact, common practice in photogrammetry but, besides making the egomotion retrieval more invasive, it requires a certain expertise and slows down the acquisition process [8]. Conversely, natural point-like features that are already present in the scene are difficult to safely extract and accurately locate. Scene features that can be quite safely detected are, instead, luminance edges [6]. These features are more likely to be naturally present in the scene and rather easy to detect, which makes them good candidate features for egomotion estimation.

Our method is based on the analysis of 3D contours in the imaged scene. Having adopted a calibrated multi-ocular camera system [9,11], the estimation is entirely performed in the 3D space. In fact, all edges of each one of the multi-views are previously localized, matched and back-projected onto the object space [12]. Roughly speaking, the method searches for the rigid motion that best merges the sets of 3D edges that are extracted from each one of the multiple views.

After partitioning the 3D contours in lines and curves, we proceed as follows:

1. rough egomotion estimation from straight contours:
    - matching of straight contours
    - motion estimation through minimization of the distance between homologous contours
2. egomotion refinement using curved contours:
    - matching of curved contours
    - motion estimation through a minimization of the distance between homologous curved contours.

Notice that, as a first approximation of the egomotion is already available, the matching of curved contours is a rather simple operation compared with the matching of straight lines.

4.1 Egomotion from Straight Lines

Line matching in 3D space is performed through a hypothesize-and-test type of procedure [17]. The first step of this method consists of formulating hypotheses on the possible couplings by selecting all those that do not violate some rules of congruence based on a set of geometrical constraints. By doing so, we drastically reduce the search space over which to test for matching correctness. At this point we can proceed with an exhaustive search through the above reduced set of hypotheses and select the match that maximizes a matching quality index.

Once the matching process is complete, the egomotion estimation can be performed rather easily by searching for the rigid motion that minimizes an appropriate *merging cost* function between two sets of 3D lines that pertain two different partial reconstructions. Notice 3D contours are generally reconstructed as chains of segments whose length and fragmentation may vary quite drastically from multi-view to multi-view. We thus proceed by first determining the 3D line portions that best fit (through linear regression) the chains of fragments of edges that have been recognized as straight. Then instead of measuring the distances between extremal points of two segments, we measure the distance between the extremal points of one segment and the line that the other segment lies upon (see Fig. 5a).

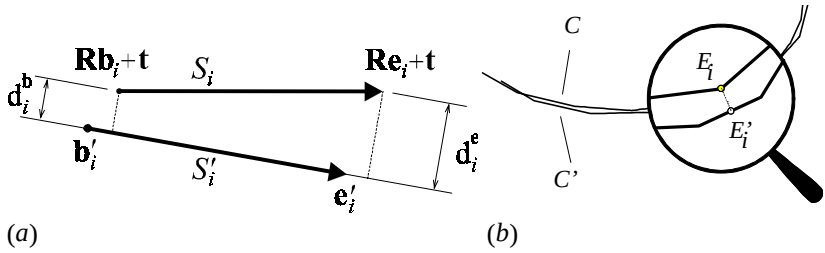

Fig. 5. Evaluation of the merging cost of two straight 3D contours (a). 3D curve matching: evaluation of the distance between two polylines (b).

Such distances are used for defining the merging cost as follows

$$C_s = \sum_{i=1}^{N} \left[\left(d_i^b \right)^2 + \left(d_i^e \right)^2 \right] .$$

In fact, the orientation of edges is usually less sensitive to fragmentation problems than their location in the 3D space [1,17].

4.2 Egomotion Refinement from Curved Contours

As already said above, curved contours are used for improving the accuracy of the egomotion's estimate. A matching process is still required but it is much simpler as we already have a first approximation of the camera motion, determined from straight edges. In fact, applying the pre-determined rigid motion to the set of curved edges, we can decide whether two curved edges are matched, depending on their global distance, which can be measured, with reference to Fig. 5a, as

$$d_g = \frac{1}{2}\left(d(C,C') + d(C',C) \right) \quad , \qquad d(C,C') = \frac{1}{N}\sum_i d(E_i,C') = \frac{1}{N}\sum_i \left\| \overline{E_i E_i'} \right\|$$

The global cost function for motion refinement is of the form $C=C_s+kC_c$, where C_s and C_c are the merging costs associated to straight and curved contours, respectively, and k is weight for balancing the two contributes.

4.3 Examples of Application

The method has been extensively tested against convergence problems and has been applied to a series of trinocular acquisitions of real images in order to evaluate qualitatively and quantitatively the accuracy of the results and the speed of convergence. Furthermore, the performance of the proposed method has been compared with that of a previously studied method [2,8] based on point correspondences between artificially added markers. Quantitative results have been obtained by measuring the maximum thickness of the bundles of edges when superimposing different sets of them with the estimated motion parameters. The performance of the proposed method has been proven to be equal to or better than that of the point-based approach, resulting in a maximum bundle size of about 100 ppm in all tests (after merging all 3D edges coming from 20 multi-views).
In Fig. 6, the results on 3D data merging are reported for an object of complex shape, in both cases of egomotion estimated through point and line correspondences. In the first case the cost function is a rigidity constraint based on the distance between reconstructed 3D points of different 3D data sets. Such points are markers that have been artificially added to the scene (white dots placed on the object's support). In the second case the egomotion is computed with the method proposed in this paper. Even though no artificially added markers have been used for the estimation, the accuracy of the estimate is comparable with that obtained through point-matching.

5. Conclusions

In this paper we presented our global approach to accurate 3D reconstruction with a calibrated multi-camera system. In particular, we presented a simple and effective technique for calibrating CCD-based multi-camera acquisition systems. The proposed method was proven to be capable of highly-accurate results even when using very

Fig. 6. One of the original views of the object, fusion of all 3D edge sets through 3D point correspondences (added marks), fusion of all 3D edge sets through 3D contour matching (natural features).

simple calibration target-sets (with little or no *a-priori* information on it) and low-cost imaging devices, such as standard TV-resolution cameras connected to commercial frame-grabbers. We also showed our approach to adaptive calibration, which proved effective for keeping track of camera parameter drift through natural feature tracking.
We also proposed and illustrated a general and robust approach to the problem of close-range partial 3D reconstruction of objects from stereo-correspondences. The method is independent on the geometry of the acquisition system which could be a set of *n* cameras with strongly converging optical axes. The robustness of the approach can be mainly attributed to the physicality of the matching process, which is virtually performed in the 3D space. In fact, both 3D location and local orientation of the surface patches are estimated, so that the geometric distortion can be accounted for. The method takes into account the viewer-dependent radiometric distortion as well.
Finally, we presented a method for performing an accurate patchworking of the partial reconstructions, through 3D feature matching. The method, based on the best fusion of 3D curves, provides very accurate results even when using standard TV-resolution CCD cameras.

References

[1] N. Ayache, “Artificial vision for Mobile Robots”, MIT Press, 1991.
[2] F. Pedersini, A. Sarti, S. Tubaro: “A Multi-view Trinocular System for Automatic 3D Object Modelling and Rendering”. XVIII Int. Congress for Photogrammetry and Remote Sensing, 1996, Vienna, Austria.

[3] Y. Otha, T. Kanade, “Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming”, IEEE Trans. On PAMI, Vol. 7, N. 2, pp. 139-154, 1985.
[4] P. Pigazzini, F. Pedersini, A. Sarti, S. Tubaro: “3D Area Matching with Arbitrary Multiview Geometry”. EURASIP Signal Processing: Image Communications. Special issue on *3D video technology*, early issues of 1998.
[5] F. Pedersini, A. Sarti, S. Tubaro: “Robust Area Matching”. IEEE Intern. Conf. on Image Processing, 1997, October 26-29, 1997, Santa Barbara, CA, USA.
[6] F. Pedersini, A. Sarti, S. Tubaro: “Egomotion Estimation of a Multicamera System through Line Correspondence”. IEEE ICIP, 1997, October 26-29, 1997, Santa Barbara, CA, USA.
[7] F. Pedersini, A. Sarti, S. Tubaro: “Automatic Surface Reconstruction of 3D Works of Art”. International Conference on Electronic Imaging and the Visual Arts (EVA'97). March 19-25, 1997, Florence, Italy.
[8] F. Pedersini, A. Sarti, S. Tubaro: “3D Motion Estimation of a Trinocular System for a Full-3D Object Reconstruction”. IEEE Intern. Conf. on Image Processing, September, 1996, Lausanne, Switzerland.
[9] R. Y. Tsai, “A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using off-the-shelf TV Cameras and Lenses” - IEEE J. on Robotics and Automation, Vol. RA-3, No. 4, Aug. 1987, pp. 323-344.
[10] J. Weng, P. Cohen, M. Herniou, “Camera Calibration with Distortion Model and Accuracy Evaluation”, *IEEE Trans. on PAMI*, Oct 1992, Vol. 14, No 10, 965-980.
[11] F. Pedersini, D. Pele, A. Sarti, S. Tubaro: “Calibration and Self-Calibration of Multi-Ocular Camera Systems”. Intl. Workshop on Synthetic-Natural Hybrid Coding and Three-Dimensional (3D) Imaging (IWSNHC3DI'97). September 5-9 1997, Rhodes, Greece.
[12] F. Pedersini, S. Tubaro: , “Accurate 3D reconstruction from trinocular views through integration of improved edge-matching and area-matching techniques.” *VIII European Signal Processing Conference*, September 10-13, 1996, Trieste, Italy
[13] L. Kitchen, A. Rosenfeld, “Gray-level corner detection”, Pattern Recognition Letters, No. 1, 1982, pp. 95-102.
[14] G. Giraudon, R. Deriche “On corner and vertex detection”, Proceedings Intl. Conf. on Computer Vision and Pattern Recognition, Maui, Hawaii, June 1991, pp. 650-655.
[15] K. Rohr, “Recognizing Corners by Fitting Parametric Models”, Intl. J. of Computer Vision, Vol. 9, No. 3, 1992, pp. 213-230.
[16] J. Mallet: “Discrete Smooth Interpolation”, ACM Tr. on Graphics, Vol. 8, No. 2, 1989, pp. 121-144.
[17] Z. Zhang, O.D. Faugeras: “3D dynamic scene analysis: a stereo based approach”, Springer, 1992.
[18] F. Pedersini, A. Sarti, S. Tubaro: “Tracking Camera Calibration in Multi-Camera Sequences through Automatic Feature Detection and Matching”. IX European Signal Processing Conference, September 8 - 11, 1998, Rhodes, Greece.
[19] L. Van Gool, A. Zisserman, “Automatic 3D model building from video sequences”. European Tr. on Telecommunications, Vol. 8, No. 4, pp. 369-78, July-Aug. 1997.