Metric 3D Surface Reconstruction from Uncalibrated Image Sequences

Marc Pollefeys, Reinhard Koch, Maarten Vergauwen, and Luc Van Gool

K.U.Leuven, ESAT-PSI, Kard. Mercierlaan 94, B-3001 Heverlee, Belgium
firstname.lastname@esat.kuleuven.ac.be

Abstract. Modeling of 3D objects from image sequences is one of the challenging problems in computer vision and has been a research topic for many years. Important theoretical and algorithmic results were achieved that allow to extract even complex 3D scene models from images. One recent effort has been to reduce the amount of calibration and to avoid restrictions on the camera motion. In this contribution an approach is described which achieves this goal by combining state-of-the-art algorithms for uncalibrated projective reconstruction, self-calibration and dense correspondence matching.

1 Introduction

Obtaining 3D models from objects is an ongoing research topic in computer vision. A few years ago the main applications were robot guidance and visual inspection. Nowadays however the emphasis is shifting. There is more and more demand for 3D models in computer graphics, virtual reality and communication. This results in a change in emphasis for the requirements. The visual quality becomes one of the main points of attention.

The acquisition conditions and the technical expertise of the users in these new application domains can often not be matched with the requirements of existing systems. These require intricate calibration procedures every time the system is used. There is an important demand for flexibility in acquisition. Calibration procedures should be absent or restricted to a minimum.

Additionally, the existing systems are often build around specialized hardware (e.g. laser range finders or stereo rigs) resulting in a high cost for these systems. Many new applications however require robust low cost acquisition systems. This stimulates the use of consumer photo- or video cameras.

In this paper we present a system which retrieves a 3D surface model from a sequence of images taken with off-the-shelf consumer cameras. The user acquires the images by freely moving the camera around the object. Neither the camera motion nor the camera settings have to be known. The obtained 3D model is a scaled version of the original object (i.e. a *metric* reconstruction), and the surface albedo is obtained from the image sequence as well.

Other researchers have presented systems for extracting 3D shape and texture from image sequences acquired with a freely moving camera. The approach of

Reinhard Koch, Luc Van Gool (Eds.): SMILE '98, LNCS 1506, pp. 139–154, 1998.

Tomasi and Kanade [32] used an affine factorization method to extract 3D from image sequences. An important restriction of this system is the assumption of orthographic projection.

Another type of system starts from an approximate 3D model and camera poses and refines the model based on images (e.g. *Facade* proposed by Debevec et al. [5]). The advantage is that less images are required. On the other hand a preliminary model must be available and the geometry should not be too complex.

Our system uses full perspective cameras and does not require prior models. It combines state-of-the-art algorithms of different domains: *projective reconstruction*, *self-calibration* and *dense depth estimation*.

Projective Reconstruction: It has been shown by Faugeras [7] and Hartley [12] that a reconstruction up to an arbitrary projective transformation was possible from an uncalibrated image sequence. Since then a lot of effort has been put in reliably obtaining accurate estimates of the projective calibration of an image sequence. Robust algorithms were proposed to estimate the fundamental matrix from image pairs [33,36]. Based on this, an algorithm which sequentially retrieves the projective calibration of a complete image sequence has been developed [1]. A more recent version based on the trifocal tensor was presented in [9].

Self-Calibration: Since a projective calibration is not sufficient for many applications, researchers tried to find ways to automatically upgrade projective calibrations to metric (i.e. euclidean up to scale). Typically, it is assumed that the same camera is used throughout the sequence and that the intrinsic camera parameters are constant. This proved a difficult problem and many researchers have worked on it [8,22,35,13,25,34,15,26]. One of the main problems is that critical motion sequences exist for which self-calibration does not result in a unique solution [31]. We proposed a more pragmatic approach [27,28] which assumes that some parameters are (approximately) known but which allows others to vary. Therefore this approach can deal with zooming/focusing cameras. Others have proposed similar approaches [2,16].

Dense Depth Estimation: Since the calibration of the image sequence has been estimated we can use stereoscopic triangulation techniques between image correspondences to estimate depth. The difficult part in stereoscopic depth estimation is to find dense correspondence maps between the images. The correspondence problem is facilitated by exploiting constraints derived from the calibration and from some assumptions about the scene. We use an approach that combines local image correlation methods with a dynamic programming approach to constrain the correspondence search [21]. This technique was first proposed by Gimmel'Farb [10] and further developed by others [4,6,19].

The rest of the paper is organized as follows: In section 2 a general overview of the system is given. In the subsequent sections the different steps are explained in more detail: projective reconstruction (section 3), self-calibration (section 4), dense matching (section 5) and model generation (section 6). Section 7 concludes the paper.

2 Overview of the Method

The presented system gradually retrieves more information about the scene and the camera setup. The first step is to relate the different images. This is done pairwise by retrieving the epipolar geometry. An initial reconstruction is then made for the first two images of the sequence. For the subsequent images the camera pose is estimated in the projective frame defined by the first two cameras. For every additional image that is processed at this stage, the interest points corresponding to points in previous images are reconstructed, refined or corrected. Therefore it is not necessary that the initial points stay visible throughout the entire sequence. The result of this step is a reconstruction of typically a few hundred interest points. The reconstruction is only determined up to a projective transformation.

The next step is to restrict the ambiguity of the reconstruction to a metric one. In a projective reconstruction not only the scene, but also the camera is distorted. Since the algorithm deals with unknown scenes, it has no way of identifying this distortion in the reconstruction. Although the camera is also assumed to be unknown, some constraints on the intrinsic camera parameters (e.g. rectangular or square pixels, constant aspect ratio, principal point in the middle of the image, ...) can often still be assumed. A distortion on the camera mostly results in the violation of one or more of these constraints. A metric reconstruction/calibration is obtained by transforming the projective reconstruction until all the constraints on the cameras intrinsic parameters are satisfied.

At this point the system effectively disposes of a calibrated image sequence. The relative position and orientation of the camera is known for all the viewpoints. This calibration facilitates the search for corresponding points and allows us to use a stereo algorithm that was developed for a calibrated system. This step allows to find correspondences for most of the pixels in the images.

From these correspondences the distance from the points to the camera center can be obtained through triangulation. These results are refined and completed by combining the correspondences from multiple images.

A dense metric 3D surface model is obtained by approximating the depth map with a triangular wire frame. The texture is obtained from the images and mapped onto the surface.

In figure 1 an overview of the systems is given. It consists of independent modules which pass on the necessary information to the next modules. The first module computes the projective calibration of the sequence together with a sparse reconstruction. In the next module the metric calibration is computed from the projective camera matrices through self-calibration. Then dense correspondence maps are estimated. Finally all results are integrated in a textured 3D surface reconstruction of the scene under consideration.

Throughout the rest of the paper the different steps of the method will be explained in more detail. An image sequence of the Arenberg castle in Leuven will be used for illustration. Some of the images of this sequence can be seen in Figure 2. The full sequence consists of 24 images recorded with a video camera.

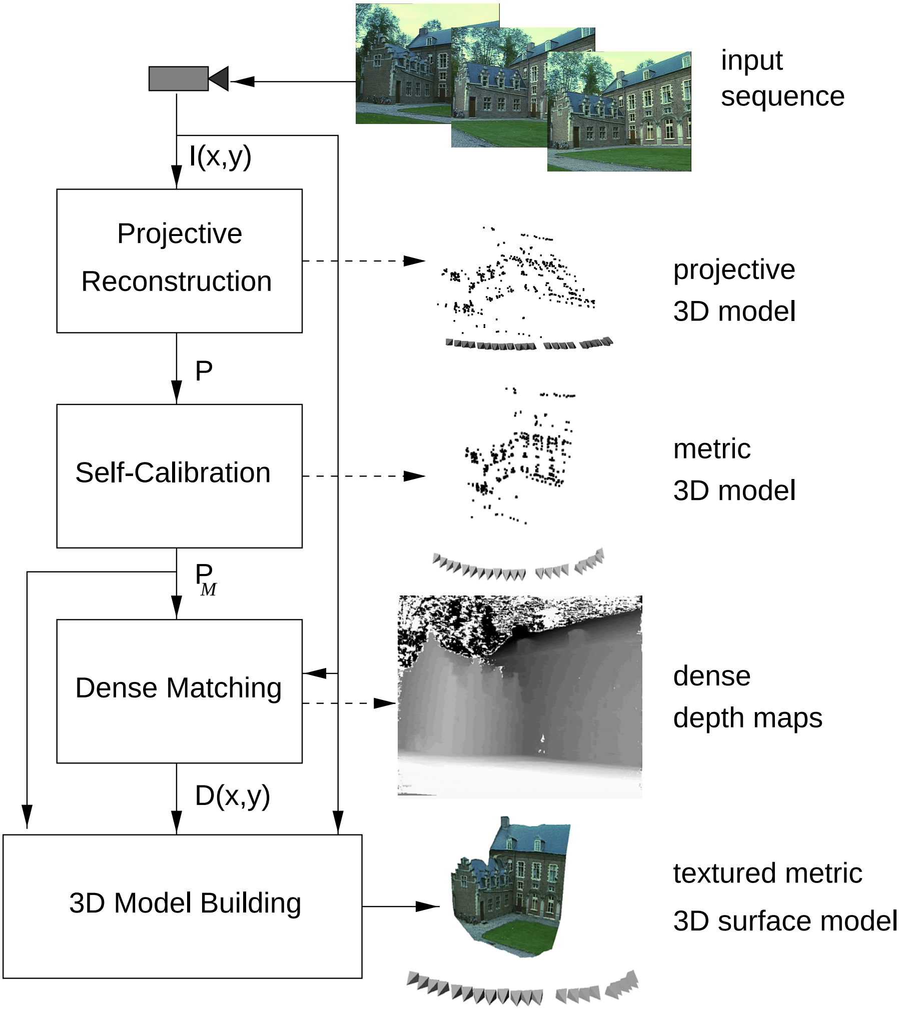

Fig. 1. Overview of the system: from the image sequence ($I(x,y)$) the projective reconstruction is computed; the projection matrices P are then passed on to the self-calibration module which delivers a metric calibration $P_{\mathcal{M}}$; the next module uses these to compute dense depth maps $D(x,y)$; all these results are assembled in the last module to yield a textured 3D surface model. On the right side the results of the different modules are shown: the preliminary reconstructions (both projective and metric) are represented by point clouds, the cameras are represented by little pyramids, the results of the dense matching are accumulated in dense depth maps (light means close and dark means far).

Fig. 2. Some images of the Arenberg castle sequence. This sequence is used throughout the paper to illustrate the different steps of the reconstruction system.

2.1 Notations

In this section some notations used in this paper are introduced. A detailed explanation of the basic concepts can be found in [23]. Projective geometry and homogeneous coordinates are used throughout this paper. Metric entities are indicated with a subscript $\mathcal{M}$.

The following equation is used to describe the perspective projection of the scene onto the images

$$m \propto \mathbf{P}M \tag{1}$$

where $\mathbf{P}$ is a 3×4 projection matrix describing the perspective projection process, $M = [X Y Z\, 1]^\top$ and $m = [x\, y\, 1]^\top$ are vectors containing the homogeneous coordinates of the world points respectively image points. Note that $\propto$ will be used throughout this paper to indicate equality up to a non-zero scale factor. Indexes i and j will be used for points (e.g. M_i), indexes k and l for views (e.g. $\mathbf{P}_k$).

In the metric case the camera projection matrix factorizes as follows:

$$\mathbf{P}_\mathcal{M} = \mathbf{K}[\mathbf{R}\,|\text{-}\mathbf{R}t] \tag{2}$$

Here $(\mathbf{R}, t)$ denotes a rigid transformation (i.e. $\mathbf{R}$ is a rotation matrix and t is a translation vector) which indicate the position and orientation of the camera, while the upper triangular calibration matrix $\mathbf{K}$ encodes the intrinsic parameters of the camera:

$$\mathbf{K} = \begin{bmatrix} f_x & s & u_x \\ & f_y & u_y \\ & & 1 \end{bmatrix} \tag{3}$$

where f_x and f_y represent the focal length divided by the pixel width resp. height, (u_x, u_y) represents the principal point and s is a factor which is zero for rectangular pixels.

The following notations are used for the epipolar geometry: $\mathbf{F}_{kl}$ is the fundamental matrix for views k and l, e_{kl} is the epipole corresponding to this fundamental matrix in view l.

3 Projective Reconstruction

At first the images are completely unrelated. The only assumption is that the images form a sequence in which consecutive images do not differ too much. Therefore the local neighborhood of image points originating from the same scene point should look similar if images are close in the sequence. This allows for automatic matching algorithms to retrieve correspondences.

3.1 Relating the Images

It is not feasible to compare every pixel of one image with every pixel of the next image. It is therefore necessary to reduce the combinatorial complexity. In addition not all points are equally well suited for automatic matching. The local neighborhoods of some points contain a lot of intensity variation and are therefore easy to differentiate from others. An interest point detector (i.e. the Harris corner detector [11]) is used to select a certain number of such suited points. These points should be well located and indicate salient features that stay visible in consecutive images. Correspondences between these image points need to be established through a matching procedure.

Matches are determined through normalized cross-correlation of the intensity values of the local neighborhood. Since images are supposed not to differ too much, corresponding points can be expected to be found back in the same region of the image. Therefore at first only interest points which have similar positions are considered for matching. When two points are mutual best matches they are considered as potential correspondences.

Since the epipolar geometry describes the complete geometry relating two views, this is what should be retrieved. Computing it from the set of potential matches through least squares does in general not give satisfying results due to its sensitivity to outliers. Therefore a robust approach should be used. Several techniques have been proposed [33,36] based on robust statistics [29]. Our system incorporates the RANSAC (RANdom SAmpling Consesus) approach used by Torr [33]. Table 1 sketches this technique.

```
• repeat
    - take minimal sample (7 matches)
    - compute $\mathbf{F}$
    - estimate $\%inliers$
  until $P_{\texttt{OK}}(\%inliers, \#trials) > 95\%$
• refine $\mathbf{F}$ (using all inliers)
```

Table 1. Robust estimation of the epipolar geometry from a set of matches containing outliers using RANSAC ($P_{\texttt{OK}}$ indicates the probability that the epipolar geometry has been correctly estimated).

Once the epipolar geometry has been retrieved, one can start looking for more matches to refine this geometry. In this case the search region is restricted to a few pixels around the epipolar lines.

3.2 Initial Reconstruction

The two first images of the sequence are used to determine a reference frame. The world frame is aligned with the first camera. The second camera is chosen so that the epipolar geometry corresponds to the retrieved $\mathbf{F}_{12}$ (see [23]).

$$\begin{array}{l} \mathbf{P}_1 = [\quad \mathbf{I}_{3\times3} \quad\quad 0 \quad] \\ \mathbf{P}_2 = [\ [e_{12}]_\times \mathbf{F}_{12} + e_{12} a^\top \ \ a_4 e_{12}\ ] \end{array} \tag{4}$$

where $[e_{12}]_\times$ indicates the vector product with e_{12}. Equation 4 is not completely determined by the epipolar geometry (i.e. $\mathbf{F}_{12}$ and e_{12}), but has 4 more degrees of freedom (i.e. $a_i, i = 1 \ldots 4$). $a = [a_1 a_2 a_3]^\top$ determines the position of the plane at infinity and a_4 determines the global scale of the reconstruction. To avoid some problems during the reconstruction it is recommended to determine a in such a way that the plane at infinity does not cross the scene. Our implementation follows the quasi-Euclidean approach proposed in [1], but an alternative would be to use Hartley's cheirality [13] or oriented projective geometry [18]. Since there is no way to determine the global scale from the images, a_4 can arbitrarily be chosen to $a_4 = 1$.

Once the cameras have been fully determined the matches can be reconstructed through triangulation. The optimal method for this is given in [14]. This gives us a preliminary reconstruction.

3.3 Adding a View

For every additional view the pose towards the pre-existing reconstruction is determined, then the reconstruction is updated. This is illustrated in Figure 3.

The first steps consists of finding the epipolar geometry as described in Section 3.1. Then the matches which correspond to already reconstructed points are used to compute the projection matrix $\mathbf{P}_k$. This is done using a robust procedure similar to the one laid out in Table 1. In this case a minimal sample of 6 matches is needed to compute $\mathbf{P}_k$. Once $\mathbf{P}_k$ has been determined the projection of already reconstructed points can be predicted. This allows to find some additional matches to refine the estimation of $\mathbf{P}_k$. This means that the search space is gradually reduced from the full image to the epipolar line to the predicted projection of the point. This is illustrated in Figure 4.

Once the camera projection matrix has been determined the reconstruction is updated. This consists of refining, correcting or deleting already reconstructed points and initializing new points for new matches.

After this procedure has been repeated for all the images, one disposes of camera poses for all the views and the reconstruction of the interest points. In the further modules mainly the camera calibration is used. The reconstruction itself is used to obtain an estimate of the disparity range for the dense stereo matching.

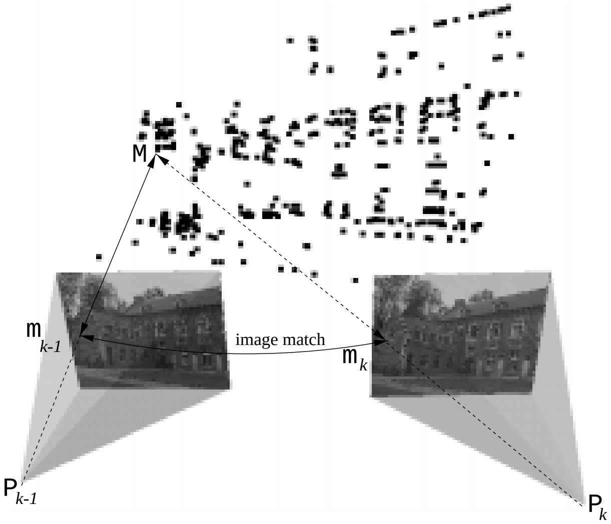

Fig. 3. Image matches (m_{k-1}, m_k) are found as described before. Since the image points, m_{k-1}, relate to object points, M_k, the pose for view k can be computed from the inferred matches (M, m_k).

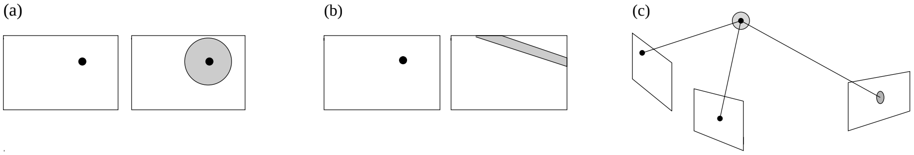

Fig. 4. (a) a priori search range, (b) search range along the epipolar line and (c) search range around the predicted position of the point.

4 Self-Calibration

The reconstruction obtained as described in the previous paragraph is only determined up to an arbitrary projective transformation. This might be sufficient for some robotics or inspection applications, but certainly not for visualization. In this section a technique to restrict this ambiguity to metric is described.

For a metric calibration the factorization of the camera projection matrices as in Equation 2 yields the physical parameters of the camera. A necessary condition for a metric reconstruction is therefore that constraints which exist on the intrinsic camera parameters are verified through this factorization.

To apply the following method to standard zooming/focusing cameras, some assumptions should be made. Often it can be assumed that pixels are rectangular or even square. If necessary (e.g. when only a short image sequence is at hand, when the projective calibration is not accurate enough or when the motion sequence is close to critical [31] without additional constraints), it can also be used that the principal point is close to the center of the image.

For the actual computations the absolute conic ω is used. This is an imaginary conic located in the plane at infinity Π_∞. Both entities are the only geometric entities which are in invariant under all Euclidean transformations. The plane at infinity and the absolute conic respectively encode the affine and metric properties of space. This means that when the position of Π_∞ is known in a projective framework, affine invariants can be measured. Since the absolute conic is invariant under Euclidean transformations its image only depends on the intrinsic camera parameters (focal length, ...) and not on the extrinsic camera parameters (camera pose). The following equation applies for the dual image of the absolute conic:

$$\omega_k^* \propto \mathbf{K}_k \mathbf{K}_k^\top \tag{5}$$

Therefore constraints on the intrinsic camera parameters are readily translated to constraints on the dual image of the absolute conic. This image is obtained from the absolute conic through the following projection equation:

$$\omega_k^* \propto \mathbf{P}_k \Omega^* \mathbf{P}_k^\top \tag{6}$$

where Ω^* is the dual absolute quadric which encodes both the absolute conic and its supporting plane, the plan at infinity. The constraints on ω_k^* can therefore be back-projected through this equation. The result is a set of constraints on the position of the absolute conic (and the plane at infinity).

Our systems first uses a linear method to obtain an approximate calibration. This calibration is then refined through a non-linear optimization step in a second phase.

4.1 Initial Calibration

To obtain a linear algorithm some assumptions have to be made. If the pixels are square and the principal point is in the middle of the image, the image can be transformed to obtain the following intrinsic camera parameters:

$$\mathbf{K}_k = \begin{bmatrix} f_k & 0 & 0 \\ & f_k & 0 \\ & & 1 \end{bmatrix} \tag{7}$$

This simplifies Equation (6) as follows:

$$\lambda \begin{bmatrix} f_k^2 & 0 & 0 \\ 0 & f_k^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{P}_k \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \\ c_2 & c_5 & c_6 & c_7 \\ c_3 & c_6 & c_8 & c_9 \\ c_4 & c_7 & c_9 & c_{10} \end{bmatrix} \mathbf{P}_k^\top \tag{8}$$

with λ an explicit scale factor. From the left-hand side of Eq. (8) it can be seen that the following equations have to be satisfied:

$$\omega_k^{*(11)} = \omega_k^{*(22)}, \tag{9}$$

$$\omega_k^{*(12)} = \omega_k^{*(13)} = \omega_k^{*(23)} = 0 \tag{10}$$

$$\omega_k^{*(21)} = \omega_k^{*(31)} = \omega_k^{*(32)} = 0 \ . \tag{11}$$

with $\omega_k^{*(ij)}$ representing the element on row i and column j of ω_k^*. Note that due to symmetry (10) and (11) result in identical equations. These constraints can thus be imposed on the right-hand side, yielding 4 independent linear equations in $c_i, i = 1 \ldots 10$ for every image:

$$\begin{aligned} P_k^{(1)} \Omega^* P_k^{(1)\top} &= P_k^{(2)} \Omega^* P_k^{(2)\top} \\ 2P_k^{(1)} \Omega^* P_k^{(2)\top} &= 0 \\ 2P_k^{(1)} \Omega^* P_k^{(3)\top} &= 0 \\ 2P_k^{(2)} \Omega^* P_k^{(3)\top} &= 0 \end{aligned}$$

with $P_k^{(j)}$ representing row j of $\mathbf{P}_k$ and Ω_* parameterized as in (8). The rank 3 constraint can be imposed by taking the closest rank 3 approximation (using SVD for example). This approach holds for sequences of 3 or more images. The special case of 2 images can also be dealt with, but with a slightly different approach. For more details see [28].

4.2 Refined Calibration

To refine the calibration Eq. (6) is used directly in a non-linear least squares criterion. In this case the user is free to specify the constraints which should be imposed. Every intrinsic parameter can be known, fixed or free. The dual image absolute conics ω_k^* should be parameterized in such a way that these constraints are enforced. For the absolute quadric $\mathbf{\Omega}^*$ a parameterization should be used which takes into account the symmetry and the rank 3 constraint. Since $\mathbf{\Omega}^*$ is only determined up to scale this leaves us with a minimum parameterization of 8 parameters. This can be done by putting $\Omega_{33}^* = 1$ and by calculating Ω_{44}^* from the rank 3 constraint. The following parameterization satisfies these requirements:

$$\Omega^* = \begin{bmatrix} \mathbf{KK}^\top & -\mathbf{KK}^\top a \\ -a^\top \mathbf{KK}^\top & a^\top \mathbf{KK}^\top a \end{bmatrix} \ . \tag{12}$$

Here a defines the position of the plane at infinity $\Pi_\infty = [\, a^\top \ 1]^\top$. In this case the transformation from projective to metric is particularly simple:

$$\mathbf{T}_{\mathcal{P}\to\mathcal{M}} = \begin{bmatrix} \mathbf{K}^{-1} & 0 \\ a^\top & 1 \end{bmatrix} \tag{13}$$

An approximate solution to these equations can be obtained through non-linear least squares. The following criterion should be minimized ($\|.\|_F$ is the Frobenius norm):

$$\min \sum_{i=1}^{n} \left\| \frac{\mathbf{K}_i \mathbf{K}_i^\top}{\|\mathbf{K}_i \mathbf{K}_i^\top\|_F} - \frac{\mathbf{P}_i \Omega^* \mathbf{P}_i^\top}{\|\mathbf{P}_i \Omega^* \mathbf{P}_i^\top\|_F} \right\|_F^2 \ . \tag{14}$$

5 Dense Depth Estimation

Only a few scene points are reconstructed from feature tracking. Obtaining a dense reconstruction could be achieved by interpolation, but in practice this does not yield satisfactory results. Small surface details would never be reconstructed in this way. Additionally, some important features are often missed during the corner matching and would therefore not appear in the reconstruction.

These problems can be avoided by using algorithms which estimate correspondences for almost every point in the images. At this point algorithms can be used which were developed for calibrated stereo rigs.

5.1 Rectification

Since we have computed the calibration between successive image pairs we can exploit the epipolar constraint that restricts the correspondence search to a 1-D search range. It is possible to re-map the image pair to standard geometry with the epipolar lines coinciding with the image scan lines [19]. The correspondence search is then reduced to a matching of the image points along each image scan-line. This results in a dramatic increase of the computational efficiency of the algorithms by enabling several optimizations in the computations. The rectification procedure is illustrated in Figure 5. For some motions (i.e. when the epipole is located in the image) standard rectification based on planar homographies is not possible and a more advanced procedure should be used [30].

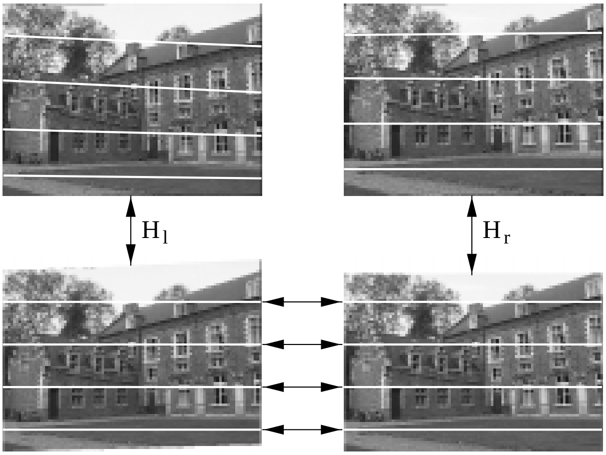

Fig. 5. Through the rectification process the image scan lines are brought into epipolar correspondence. This allows important gains in computational efficiency and simplification of the dense stereo matching algorithm.

5.2 Dense Stereo Matching

In addition to the epipolar geometry other constraints like preserving the order of neighboring pixels, bidirectional uniqueness of the match, and detection of occlusions can be exploited. These constraints are used to guide the correspondence towards the most probable scan-line match using a dynamic programming scheme [6].

For dense correspondence matching a disparity estimator based on the dynamic programming scheme of Cox *et al.* [4], is employed that incorporates the above mentioned constraints. It operates on rectified image pairs (I_k, I_l) where the epipolar lines coincide with image scan lines. The matcher searches at each pixel in image I_k for maximum normalized cross correlation in I_l by shifting a small measurement window (kernel size 5x5 to 7x7 pixel) along the corresponding scan line. The selected search step size ΔD (usually 1 pixel) determines the search resolution. Matching ambiguities are resolved by exploiting the ordering constraint in the dynamic programming approach [19]. The algorithm was further adapted to employ extended neighborhood relationships and a pyramidal estimation scheme to reliably deal with very large disparity ranges of over 50% of image size [6].

5.3 Multiview Matching

The pairwise disparity estimation allows to compute image to image correspondence between adjacent rectified image pairs, and independent depth estimates for each camera viewpoint. An optimal joint estimate is achieved by fusing all independent estimates into a common 3D model. The fusion can be performed in an economical way through controlled correspondence linking. The approach utilizes a flexible multi viewpoint scheme which combines the advantages of small baseline and wide baseline stereo [21].

Assume an image sequence with $k = 1 \rightarrow n$ images. Starting from a reference view point k the correspondences between adjacent images $(k+1, k+2, ..., n)$ and $(k-1, k-2, ..., 1)$ are linked in a chain. The depth for each reference image point m_k is computed from the correspondence linking that delivers two lists of image correspondences relative to the reference, one linking down from $k \rightarrow 1$ and one linking up from $k \rightarrow n$. For each valid corresponding point pair $(\mathbf{x_k}, \mathbf{x_l})$ we can triangulate a depth estimate $d(x_k, x_l)$ along S_{m_k} with e_l representing the depth uncertainty. The left part of Figure 6 visualizes the decreasing uncertainty interval during linking.

While the disparity measurement resolution ΔD in the image is kept constant (at 1 pixel), the reprojected depth error e_l decreases with the baseline. Outliers are detected by controlling the statistics of the depth estimate computed from the correspondences. All depth values that fall within the uncertainty interval around the mean depth estimate are treated as inliers. They are fused by a 1-D kalman filter to obtain an optimal mean depth estimate. Outliers are undetected correspondence failures and may be arbitrarily large. As threshold to detect the outliers we utilize the depth uncertainty interval e_l.

The result of this procedure is a very dense depth map. Most occlusion problems are avoided by linking correspondences from up and down the sequence. An example of such a very dense depth map is given in Figure 6.

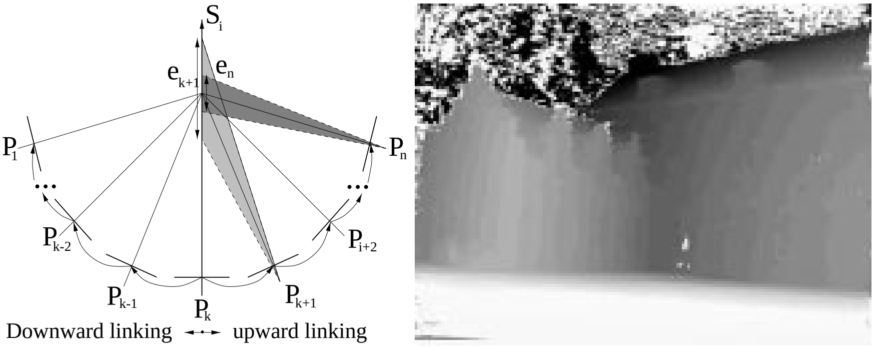

Fig. 6. Depth fusion and uncertainty reduction from correspondence linking (left), Resulting dense depth map (light means near and dark means far) (right).

6 Building the Model

The dense depth maps as computed by the correspondence linking must be approximated by a 3D surface representation suitable for visualization. So far each object point was treated independently. To achieve spatial coherence for a connected surface, the depth map is spatially interpolated using a parametric surface model. The boundaries of the objects to be modeled are computed through depth segmentation. In a first step, an object is defined as a connected region in space. Simple morphological filtering removes spurious and very small regions. We then employ a bounded thin plate model with a second order spline to smooth the surface and to interpolate small surface gaps in regions that could not be measured. If the object consist of dominant planar regions, the local surface normal may be exploited to segment the object into planar parts [20].

The spatially smoothed surface is then approximated by a triangular wireframe mesh to reduce geometric complexity and to tailor the model to the requirements of Computer Graphics visualization systems. The mesh triangulation currently utilizes the reference view only to build the model. The surface fusion from different view points to completely close the models remains to be implemented. Sometimes it is not possible to obtain a single metric framework for large objects like buildings since one may not be able to record images continuously around it. In that case the different frameworks have to be registered to each other. This will be done using available surface registration schemes [3].

Texture mapping onto the wire-frame model greatly enhances the realism of the models. As texture map one could take the reference image texture alone and map it to the surface model. However, this creates a bias towards the selected image and imaging artifacts like sensor noise, unwanted specular reflections or the shading of the particular image is directly transformed onto the object. A better choice is to fuse the texture from the image sequence in much the same way as depth fusion.

The viewpoint linking builds a controlled chain of correspondences that can be used for texture enhancement as well. The estimation of a robust mean texture will capture the static object only and the artifacts (e.g. specular reflections or pedestrians passing in front of a building) are suppressed [17]. The texture fusion could also be done on a finer grid, yielding a super resolution texture [24].

An example of the resulting model can be seen in Figure 7.

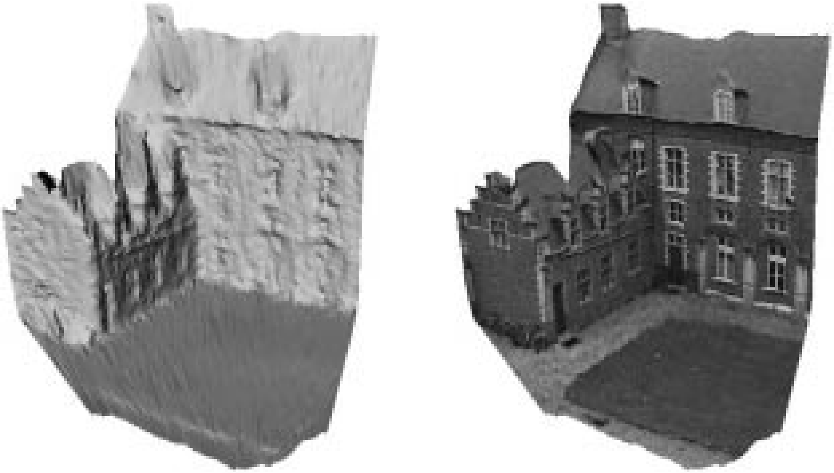

Fig. 7. 3D surface model obtained automatically from an uncalibrated image sequence, shaded (left), textured (right).

7 Conclusion

An automatic 3D scene modeling technique was discussed that is capable of building models from uncalibrated image sequences. The technique is able to extract metric 3D models without any prior knowledge about the scene or the camera. The calibration is obtained by assuming a rigid scene and some constraints on the intrinsic camera parameters (e.g. square pixels).

Work remains to be done to get more complete models by fusing the partial 3D reconstructions. This will also increase the accuracy of the models and eliminate artifacts at the occluding boundaries. For this we can rely on work already done for calibrated systems.

Acknowledgments

We would like to thank Andrew Zisserman and his team from Oxford for supplying us with robust projective reconstruction software. A specialization grant from the Flemish Institute for Scientific Research in Industry (IWT), the financial support from the EU ACTS project AC074 'VANGUARD' and the Belgian IUAP project 'IMechS' are also gratefully acknowledged.

References

1. P. Beardsley, P. Torr and A. Zisserman, 3D Model Acquisition from Extended Image Sequences. *Proc. European Conference on Computer Vision*, Cambridge, UK, vol.2, pp.683-695, 1996.
2. S. Bougnoux, From Projective to Euclidean Space Under any Practical Situation, a Criticism of Self-Calibration, In *Proc. International Conference on Computer Vision*, pp.790-796, Bombay, 1998.
3. Y. Chen and G. Medioni, Object Modeling by Registration of Multiple Range Images, In *Proc. IEEE International Conference on Robotics and Automation*, pp.2724-2729, Sacramento (CA), 1991.
4. I. Cox, S. Hingorani and S. Rao, A Maximum Likelihood Stereo Algorithm, *Computer Vision and Image Understanding*, Vol. 63, No. 3, May 1996.
5. P. Debevec, C. Taylor and J. Malik, Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach, In *Siggraph*, 1996.
6. L. Falkenhagen, Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints. In *Proc. International Workshop on SNHC and 3D Imaging*, Rhodes, Greece, 1997.
7. O. Faugeras, What can be seen in three dimensions with an uncalibrated stereo rig. In *Proc. European Conference on Computer Vision*, pp.563-578, 1992.
8. O. Faugeras, Q.-T. Luong and S. Maybank, Camera self-calibration: Theory and experiments, In *Proc. European Conference on Computer Vision*, pp.321-334, 1992.
9. A. Fitzgibbon and A. Zisserman, Automatic camera recovery for closed or open image sequences, In *Proc. European Conference on Computer Vision*, pp.311-326, Freiburg, 1998.
10. G. Gimel'farb, Symmetrical approach to the problem of automatic stereoscopic measurements in photogrammetry, Cybernetics, 1979, 15(20, 235-247; Consultants Bureau, N.Y.
11. C. Harris and M. Stephens, A combined corner and edge detector, in *Fourth Alvey Vision Conference*, pp.147-151, 1988.
12. R. Hartley, Estimation of relative camera positions for uncalibrated cameras. In *Proc. European Conference on Computer Vision*, pp.579-587, 1992.
13. R. Hartley, Euclidean reconstruction from uncalibrated views. In *Applications of invariance in Computer Vision*, LNCS 825, Springer-Verlag, pp.237-256, 1994.
14. R. Hartley and P. Sturm, Triangulation, *Computer Vision and Image Understanding*, 68(2):146-157, 1997.
15. A. Heyden and K. Åström, Euclidean Reconstruction from Constant Intrinsic Parameters In *Proc. International Conference on Pattern Recognition*, Vienna, Austria, pp.339-343, 1996.
16. A. Heyden, K. Åström, Euclidean Reconstruction from Image Sequences with Varying and Unknown Focal Length and Principal Point, In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.

17. M. Irani and S. Peleg, Super resolution from image sequences, In *Proc. International Conference on Pattern Recognition*, Atlantic City, NJ, 1990.
18. S. Laveau, Géométrie d'un systeème de N caméras. Théorie, estimation et applications. Ph.D. thesis, Ecole Polytechnique, France, 1996.
19. R. Koch, Automatische Oberflachenmodellierung starrer dreidimensionaler Objekte aus stereoskopischen Rundum-Ansichten. *PhD thesis*, University of Hannover, Germany, 1996.
20. R. Koch, Surface Segmentation and Modeling from Stereoscopic Image Sequences, In *Proc. International Conference on Pattern Recognition*, Vienna, 1996.
21. R. Koch, M. Pollefeys and L. Van Gool, Multi Viewpoint Stereo from Uncalibrated Video Sequences, In *Proc. European Conference on Computer Vision*, Freiburg, Germany, 1998.
22. Q.-T. Luong and O. Faugeras, O. Self Calibration of a moving camera from point correspondences and fundamental matrices. In *Internation Journal of Computer Vision*, vol.22-3, 1997.
23. Moons, T. A Tutorial on Multi-View Relationships, In *Proc. SMILE workshop*, Freiburg, 1998.
24. E. Ofek, E. Shilat, A. Rappopport and M. Werman, Highlight and Reflection Independent Multiresolution Textures from Image Sequences. *IEEE Computer Graphics and Applications*, vol.17 (2), March-April 1997.
25. M. Pollefeys and L. Van Gool, A stratified approach to self-calibration. In *Proc. International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp.407-412, 1997.
26. M. Pollefeys and L. Van Gool, Self-calibration from the absolute conic on the plane at infinity, In *Proc. International Conference on Computer Analysis of Images and Patterns*, Kiel, Germany, pp. 175-182, 1997.
27. M. Pollefeys, R. Koch and L. Van Gool, Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, In *Proc. International Conference on Computer Vision*, pp.90-95, Bombay, 1998.
28. M. Pollefeys, R. Koch and L. Van Gool, Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Intrinsic Camera Parameters, to appear in IJCV.
29. P. Rousseeuw, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
30. S. Roy, J. Meunier and I. Cox, Cylindrical Rectification to Minimize Epipolar Distortion, In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.393-399, 1997.
31. P. Sturm, Critical Motion Sequences for Monocular Self-Calibration and Uncalibrated Euclidean Reconstruction. In *Proc. International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp.1100-1105, 1997.
32. C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: A factorization approach, *International Journal of Computer Vision*, 9(2):137-154, 1992.
33. P. Torr, Motion Segmentation and Outlier Detection, PhD Thesis, Dept. of Engineering Science, University of Oxford, 1995.
34. B. Triggs, The Absolute Quadric, In *Proc. International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp.609-614, 1997.
35. C. Zeller and O. Faugeras, Camera self-calibration from video sequences: the Kruppa equations revisited. INRIA, Sophia-Antipolis, France, Research Report 2793, 1996.
36. Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong, A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, *Artificial Intelligence Journal*, Vol.78, pp.87-119, October 1995.