Realistic Surface Reconstruction of 3D Scenes from Uncalibrated Image Sequences

Reinhard Koch[1], Marc Pollefeys, and Luc Van Gool
K.U.Leuven, Dept. Elektrotechniek, ESAT-PSI
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium
{firstname.lastname}@esat.kuleuven.ac.be

Abstract

This contribution addresses the problem of obtaining 3D models from image sequences. A 3D surface description of the scene is extracted completely from a set of uncalibrated camera images of the scene. No prior knowledge about the scene or about the camera is needed to build the 3D models. The only assumptions are the rigidity of the scene objects and opaque object surfaces.

The modeling system described here uses a 3-step approach. First, the camera pose and intrinsic parameters are calibrated by tracking salient feature points throughout the sequence. Next, consecutive images of the sequence are treated as stereoscopic image pairs and dense correspondence maps are computed by area matching. Finally, dense and accurate depth maps are computed by linking together all correspondences over the viewpoints. The depth maps are converted to triangular surfaces meshes that are texture mapped for photo-realistic appearance. The feasibility of the approach has been tested on both real and synthetic data and is illustrated here on several outdoor image sequences.

1. Introduction

The use of three-dimensional surface models for the purpose of visualization is gaining importance. Highly realistic 3D models are readily used to visualize and simulate events, like in flight simulators, in the games and film industry or for product presentations. The range of applications spans from architecture visualization over virtual television studios, virtual presence for video communications to general "virtual reality" applications.

A limitation to the widespread use of these techniques is currently the high costs of such 3D models since they have to be produced manually. Especially if existing objects are to be reconstructed, the measurement process for obtaining the correct geometric and photometric data is tedious and time consuming. Traditional solutions include the use of stereo rigs, laser range scanners and other 3D digitizing devices. These devices are often very expensive, require careful handling and complex calibration procedures and are designed for a restricted depth range only.

In this work an image based approach is proposed which avoids most of the problems mentioned above. The scene which has to be modeled is recorded from different viewpoints by a video camera. The relative position and orientation of the camera and its calibration parameters will automatically be retrieved from the image data by the proposed algorithms. Hence, there is no need for measurements in the scene or calibration procedures whatsoever. There is also no restriction on range, it is just as easy to model a small object, as to model a complete landscape. The proposed method thus offers a previously unknown flexibility in 3D model acquisition. In addition, any photographic recording device - e.g. camcorder, digital camera, or even standard photographic film camera - is sufficient for scene acquisition. Hence, increased flexibility is accompanied by a decrease in cost.

In this contribution we will discuss the complete and automatic modeling system that is capable of computing accurate and dense 3D surface models from uncalibrated image sequences. We review the state of the art for scene reconstruction from images in section 2. Section 3 gives an overview of the proposed system and discusses the steps needed for depth estimation from image sequences. Section 4 deals with the 3D model generation and the creation of textured surfaces. In sect. 5 several experiments on real outdoor sequences are performed. Objects of different scale are modeled and different imaging sensors are used to demonstrate the quality and flexibility of the proposed reconstruction system.

[1]The author is now with the Institute of Computer Science, Division Multimedia Information Processing, at the University of Kiel, Germany.

2 State of the art

There have been numerous approaches to reconstruct and to visualize existing 3D environments from image sequences. Two main directions of research have evolved: geometry-based and image-based scene representations. Both methods aim at realistic capture and fast visualization of 3D scenes from image sequences.

Image-based rendering approaches like plenoptic modeling [24], lightfield rendering [22] and the lumigraph [11] have lately received a lot of attention, since they can capture the appearance of a 3D scene from images only, without the explicit use of 3D geometry. Thus one may be able to capture objects with very complex geometry and with non-lambertian surface reflectivity that can not be modeled otherwise. Basically one caches all possible views of the scene and retrieves them during view rendering. The price to pay for this approach is a very high amount of data and a tedious image acquisition. In fact, one has to obtain the radiance of the scene from light rays in all possible positions and orientations, which is a 5-dimensional function.

Panoramic image mosaics are another way to represent the environment from a restricted set of viewpoints. Panoramics are obtained by rotating the camera around a fixed viewpoint and allow highly realistic rendering from this viewpoints [15, 30].

Geometric 3D modeling approaches generate polygonal (triangular) surface meshes of a scene. A limited set of calibrated camera views of the scene is sufficient for reconstruction. Texture mapping adds the necessary fidelity for photo-realistic rendering to the object surface. Methods were reported that generate complete environments from sets of panoramic images when the camera pose is known by instrumentation [31], and for semi-automatic modeling of architectural scenes when the class of objects is restricted to simple shapes [4, 29].

Common to all approaches is the image acquisition part. Images of the observed scene have to be taken and evaluated to obtain a realistic representation. Here we can distinguish between calibrated and uncalibrated image acquisition. In the case of fully calibrated image sequences the pose and orientation for each acquisition viewpoint is known a priory, through the use of external camera calibration devices or when using mechanical pose control with a camera on a robot arm. The need to obtain a precise calibration from external measurements places severe restrictions on the image acquisition process and limits the applicability for real scenes.

In the uncalibrated case no prior knowledge of camera poses and intrinsic camera parameters is assumed and all parameters, camera pose and intrinsic calibration as well as the 3D scene structure have to be estimated from the 2D image sequence alone. The advantage of this approach is that simple photographs of the scene can be used without any prior knowledge and without additional calibration equipment. Even old footage taken with any camera system can be used for reconstruction. The method proposed in this contribution is placed in this framework. Since we do not rely on any prior calibration or scene information, we can handle a wider range of scenes than the above mentioned methods.

Faugeras and Hartley first demonstrated how to obtain uncalibrated projective reconstructions from image sequences alone [7, 12]. Since then, researchers tried to find ways to upgrade these reconstructions to metric (i.e. Euclidean but unknown scale, see [8, 33, 27]). Newest results report full self-calibration methods even for varying intrinsic parameters like focal length, which allows the unrestricted use of the camera, for example zooming [13, 26, 28]. To employ these self-calibration methods for sequence analysis they must be embedded in a complete scene reconstruction system. Beardsley et al. [1] proposed a scheme to obtain projective calibration and 3D structure by robustly tracking salient feature points throughout an image sequence. This sparse object representation outlines the object shape, but gives not sufficient surface detail for visual reconstruction. Highly realistic 3D surface models need the dense depth estimation and can not rely on few feature points alone. The work of Fitzgibbon and Zisserman recently extended the approach to model objects from all sides [9].

In [28] we extended the method of Beardsley in two directions. On the one hand the projective reconstruction was updated to metric even for varying internal camera parameters, on the other hand a dense stereo matching technique [5] was applied between two selected images of the sequence to obtain a dense depth map for a single viewpoint. From this depth map a triangular surface wire-frame was constructed and texture mapping from one image was applied to obtain realistic surface models. In [18] the approach was further extended to multi viewpoint sequence analysis. Newest results show that the proposed method allows a combined framework for image- and geometry-based scene reconstructions [19, 20]. In this contribution we concentrate on the reconstruction of geometric 3D scene models.

3. Geometric Modeling from Image Sequences

Robust camera calibration and accurate depth estimation are the key problems to be solved. In our system we use a 3-step approach that is visualized in fig. 1 with the example of modeling a building facade:

- Camera self-calibration is obtained by robust tracking of salient feature points over the image sequence,
- dense depth maps are computed between adjacent image pairs,

- depth maps are linked together by multiple view point linking to fuse depth measurements from all images into a consistent model. The model is stored as a textured 3-D surface mesh.

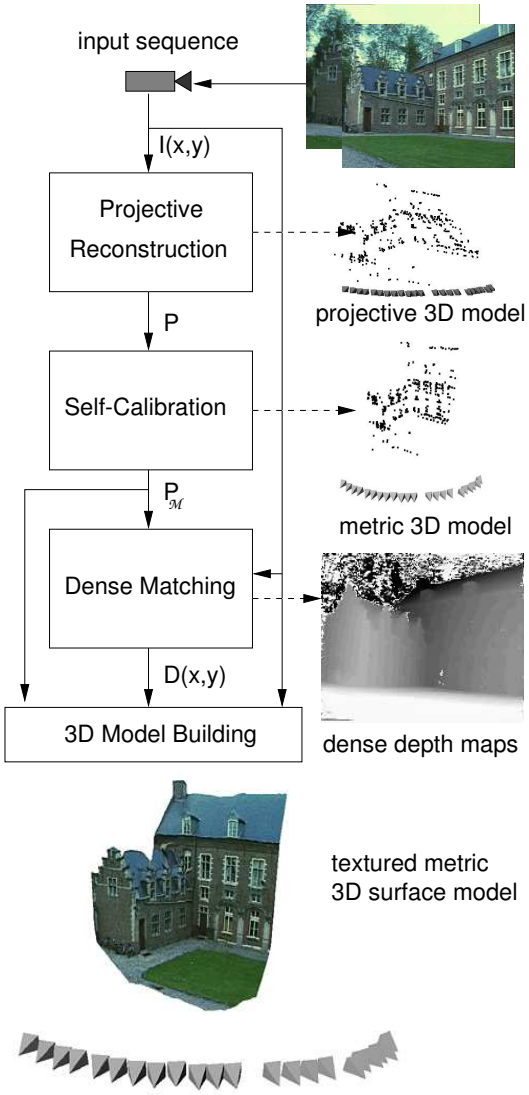

Figure 1. System overview: from the image sequence ($I(x,y)$) the projective reconstruction is computed; the projection matrices P are then passed on to the self-calibration module which delivers a metric calibration $P_{\mathcal{M}}$; the next module uses these to compute dense depth maps $D(x,y)$; all these results are assembled in the last module to yield a textured 3-D surface model. The little pyramids in front of the building symbolize the different camera positions.

3.1. Camera Calibration through Feature Tracking

Camera calibration[2] is obtained by tracking salient image features throughout the sequence. The difficulty of this step is to robustly find at least a few but very reliable correspondences that are needed for camera calibration. Salient feature points like strong intensity corners are matched using robust (RANSAC) techniques for that purpose. In a two-step procedure a projective calibration and feature point reconstruction is recovered from the image sequence which is then updated to metric calibration with a self-calibration approach.

Retrieving the projective framework. At first feature correspondences are found by extracting intensity corners in different images and matching them using a robust corner matcher [32]. In conjunction with the matching of the corners a restricted calibration of the setup is calculated (i.e. only determined up to an arbitrary projective transformation). This allows to eliminate matches which are inconsistent with the calibration. The 3D position of a point is restricted to the line passing through its image point and the camera projection center. Therefore the corresponding point is restricted to the projection of this line in the other image. Using this constraint, more matches can easily be found and used to refine this calibration.

The matching is started on the first two images of the sequence. The calibration of these views define a projective framework in which the projection matrices of the other views are retrieved one by one. In this approach we follow the procedure proposed by Beardsley *et al* [1]. We therefore obtain projection matrices (3×4) of the following form:

$$\mathbf{P}_1 = [\mathbf{I}|0] \text{ and } \mathbf{P}_k = [\mathbf{H}_{1k}|e_{1k}] \tag{1}$$

with $\mathbf{H}_{1k}$ the homography for some reference plane from view 1 to view k and e_{1k} the corresponding epipole.

Retrieving the metric framework. Such a projective calibration is certainly not satisfactory for the purpose of 3D modeling. A reconstruction obtained up to a projective transformation can differ very much from the original scene according to human perception: orthogonality and parallelism are in general not preserved, part of the scene can be warped to infinity, etc. To obtain a better calibration, constraints on the internal camera parameters can be imposed (e.g. absence of skew, known aspect ratio, ...). By exploiting these constraints, the projective reconstruction can be upgraded to metric (Euclidean up to scale). In the metric case the camera projection matrices have the following

[2] By *calibration* we mean the actual internal calibration of the camera as well as the relative position and orientation of the camera for the different views with respect to an arbitrary coordinate system.

form:

$$\mathbf{P}_i = \mathbf{K}_i[\mathbf{R}_i|\text{-}\mathbf{R}_i\mathbf{t}_i] \text{ with } \mathbf{K}_i = \begin{bmatrix} f_x & s & u_x \\ & f_y & u_y \\ & & 1 \end{bmatrix} \quad (2)$$

where $\mathbf{R}_i$ and $\mathbf{t}_i$ indicate the orientation and position of the camera for view i and $\mathbf{K}_i$ contains the internal camera parameters: f_x and f_y stand for the horizontal and vertical focal length (in pixels), $\mathbf{u} = (u_x, u_y)$ is the principal point and s is a measure of the image skew.

A practical way to obtain the calibration parameters from constraints on the internal camera parameters is through application of the concept of the absolute quadric [33, 28]. In space, exactly one degenerate quadric of planes exists which has the property to be invariant under all rigid transformations. In a metric frame it is represented by the following 4×4 symmetric rank 3 matrix $\Omega = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$. If $\mathbf{T}$ transforms points $M \rightarrow \mathbf{T}M$ (and thus $\mathbf{P} \rightarrow \mathbf{P}\mathbf{T}^{-1}$), then it transforms $\Omega \rightarrow \mathbf{T}\Omega\mathbf{T}^\top$ (which can be verified to yield Ω when $\mathbf{T}$ is a similarity transformation). The projection of the absolute quadric in the image yields the intrinsic camera parameters independent of the chosen projective basis[3]:

$$\mathbf{K}_i\mathbf{K}_i^\top \propto \mathbf{P}_i\Omega\mathbf{P}_i^\top \quad (3)$$

where $\propto$ means equal up to an arbitrary non-zero scale factor. Therefore constraints on the internal camera parameters in $\mathbf{K}_i$ can be translated to constraints on the absolute quadric. If enough constraints are at hand, only one quadric will satisfy them all, i.e. the *absolute quadric*. At that point the scene can be transformed to the metric frame, which brings Ω to its canonical form. For a detailed analysis of the calibration procedure see [28].

3.2. Dense Correspondence Matching

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. Often some important features are missed during the corner matching and will 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. 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. In particular it is possible to re-map the image pair to standard geometry with the epipolar lines coinciding with the image scan lines. The correspondence search is then reduced to a matching of the image points along each image scan-line.

[3]Using Equation 2 this can be verified for a metric basis. Transforming $\mathbf{P} \rightarrow \mathbf{P}\mathbf{T}^{-1}$ and $\Omega \rightarrow \mathbf{T}\Omega\mathbf{T}^\top$ will not change the projection.

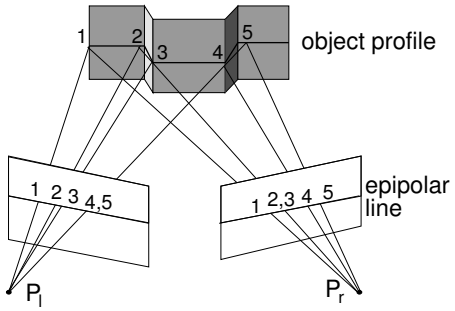

Figure 2. Object profile triangulation from ordered neighboring correspondences.

The epipolar constraint obtained from calibration restricts corresponding image points to lie in the epipolar plane[4] which also cuts a 3D profile out of the surface of the scene objects. The profile projects onto the corresponding epipolar lines in $\mathbf{I_i}$ and $\mathbf{I_k}$ where it forms an ordered set of neighboring correspondences (see figure 2). For well behaved surfaces this ordering is preserved and delivers an additional constraint, known as 'ordering constraint'. Scene constraints like this can be applied by making weak assumptions about the object geometry. In many real applications the observed objects will be opaque and composed out of piecewise continuous surfaces. If this restriction holds then additional constraints can be imposed on the correspondence estimation. Kochan[21] listed as many as 12 different constraints for correspondence estimation in stereo pairs. Of them, the two most important ones apart from the epipolar constraint are:

- Ordering Constraint: For opaque surfaces the order of neighboring correspondences on the corresponding epipolar lines is always preserved. This ordering allows the construction of a dynamic programming scheme which is employed by many dense disparity estimation algorithms [10], [3], [5].
- Uniqueness Constraint: The correspondence between any two corresponding points is bidirectional as long as there is no occlusion in one of the images. A correspondence vector pointing from an image point to its corresponding point in the other image always has a corresponding reverse vector pointing back. This test is used to to detect outliers and occlusions.

For dense correspondence matching a disparity estimator based on the dynamic programming scheme of Cox

[4]The epipolar plane is the plane defined by the the image point and the camera projection centers

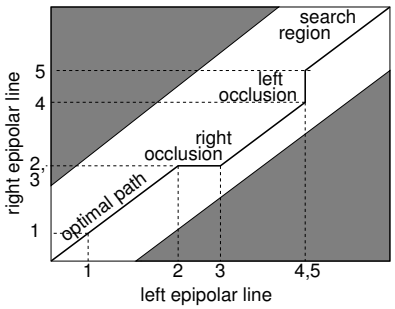

Figure 3. Scanline search for best disparity profile that exploits the constraints.

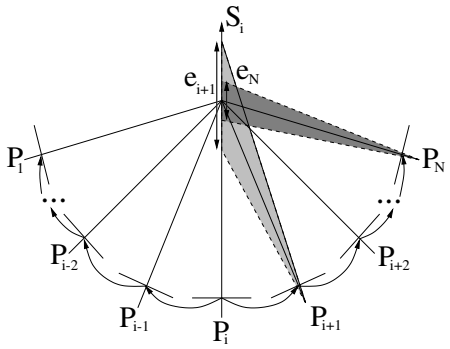

Figure 4. Depth fusion and uncertainty reduction from correspondence linking along the line of sight S_i for a reference image P_i.

et al. [3], is employed that incorporates the above mentioned constraints. It operates on rectified image pairs (I_i, I_k) where the epipolar lines coincide with image scan lines. The matcher searches at each pixel in image I_i for maximum normalized cross correlation in I_k 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 (see Fig. 3). 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 [5].

3.3. Sequence Linking

The pairwise disparity estimation allows the computation of 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 by combining the advantages of small baseline and wide baseline stereo [18].

As *small baseline stereo* we define viewpoints with a baseline much smaller than the observed average scene depth. This configuration is usually valid for image sequences were the images are taken as a spatial sequence from many slightly varying view points. The advantages are an easy correspondence estimation and small regions of occlusion[5] between adjacent images. Disadvantage is clearly the limited depth resolution due to the small triangulation angle between the view points.

[5] As occlusions we consider those parts of the object that are visible in a single image only, due to object self-occlusion.

The *wide baseline stereo* in contrast is used mostly with still image photographs of a scene where few images are taken from a very different viewpoint. Here the depth resolution is superior but correspondence and occlusion problems appear, because the views are very different and large image regions without correspondence may occur.

The *multi viewpoint linking* combines the virtues of both approaches by concatenating corresponding points over multiple images. In addition it will produce denser depth maps than either of the other techniques, and allow for additional features during depth and texture fusion [18]. In the linking process care is taken to deal with occlusions and to check for measurement outliers.

3.3.1 Depth fusion

Assume an image sequence with $k = 1 \rightarrow N$ images, and i being an arbitrary image of that sequence. The image i is called reference image and all measurements are related to this image. The goal is to compute a dense depth map for a given reference image i and its view point P_i. Starting from the reference image i, the correspondences between adjacent images $k = \{i+1, ..., N\}$ and $k = \{i-1, ..., 1\}$ are linked in a chain. The depth for each reference image point $\mathbf{x_i}$ is computed from the correspondence linking that delivers two lists of image correspondences relative to the reference, one linking down from $i \rightarrow 1$ and one linking up from $i \rightarrow N$. For each valid pair of corresponding points $(\mathbf{x_i}, \mathbf{x_k})$ we can triangulate a depth estimate $d(x_i, x_k)$ along the line of sight S_{xi} of the corresponding pixel with the range e_k representing the depth uncertainty. Figure 4 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_k decreases with an increasing triangulation angle. 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_k.

3.3.2 Occlusions

If an object region is visible in image i but not in k, we speak of an occlusion. Occlusions are eliminated by incorporating a multi viewpoint matcher that operates symmetrically to a particular viewpoint i. Points that are occluded in the view $i+1$ are normally visible in the view $i-1$ and vice versa. The exploitation of links starting up and down from viewpoint i resolves most of the occlusions and produces a very dense depth map.

4. Surface Modeling

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 consists of dominant planar regions, the local surface normal may be exploited to segment the object into planar parts [17].

The spatially smoothed surface is then approximated by a triangular wire-frame mesh to reduce geometric complexity and to tailor the model to the requirements of Computer Graphics visualization systems.

4.1. Texture Fusion

Texture mapping onto the wire-frame model greatly enhances the realism of the models. As a texture map one could take the texture map of the reference image only 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 are 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. A texture map in this context is defined as the color intensity values for a given set of image points, usually the pixel coordinates. While depth is concerned with the *position* of the correspondence in the image, texture uses the *color intensity value* of the corresponding image point. For each reference image position one may now collect a list of color intensity values from the corresponding image positions in the other viewpoints. This allows to enhance the original texture in many ways by accessing the color statistics. Some features that can be derived naturally from the texture linking algorithm are described below.

Specular reflection and artifact removal. The surface reflectance of the object is modeled by a viewpoint independent diffuse and a viewpoint dependent specular reflection. In this case the color intensity statistics can be modeled as Gaussian noise contaminated with an outlier tail distribution that contains the reflection. By collecting the corresponding color intensities over a series of different viewpoints, one can detect the specular reflectance as outlier and retain the diffuse reflection using median filtering. The same statistics hold if a fast moving object temporarily occludes the observed object, like a pedestrian passing in front of a building to be modeled. The exploitation of a robust mean texture will therefore capture the static object only and the artifacts are suppressed [14].

Super-resolution texture. The correspondence linking is not restricted to pixel-resolution, since each between-pixel-position in the reference image can be used to start a correspondence chain as well. Color intensity values will then be interpolated between the pixel grid. If the object is observed from many different view points and possibly from different object distances, the finite pixel grid of the images for each viewpoint is generally slightly displaced. This displacement can be exploited to create super-resolution texture by fusing all images on a finer resampling grid. The super-resolution grid in the reference image can be chosen arbitrarily fine, but the measurable real resolution of course depends on the displacement and resolution of the corresponding images [25].

4.2 Multiscale Integration

Sometimes it is not possible to obtain a unique metric framework for large objects like buildings since one may not be able to record images continuously around it. For scenes with high complexity we also need an adaptive level of detail since not all scene parts need to be modeled with the same resolution. In that case the different sequences have to be registered to each other. A problem here is to achieve consistency of the models. We follow a coarse-to-fine multiscale approach. The scene is first recorded from

some distance to get an overview of the complete object on a coarse scale. Next we move nearer to the object to record details that can be fitted into the overview model on a finer scale. The registration of the differently scaled models is currently performed semi-automatic with a CAD modeling tool like 3D-StudioMax. Three reference points for each model are selected interactively at each scale and the detail models are then fitted automatically to the overview model, based on the transformations computed between the reference points. In the future we plan to automate this procedure further based on existing surface registration schemes [2].

5. Experiments

In this section the performance of the modeling system is tested on different outdoor sequences.

5.1. Castle sequence

The *Castle* sequence consists of 22 images of 720x576 pixel resolution taken with a standard semi-professional camcorder that was moved freely in front of a building. Figure 5 shows the images 1,8,14, and 22 of the sequence.

To judge the geometric and visual quality of the reconstruction, different perspective views of the model were computed and displayed in Figure 6. In the shaded view, the geometric details like the window and door niches are seen. A close-up view from a position that a human observer would take reveals the high visual quality of the model. To demonstrate the texture fusion capabilities of the algorithm, the specular reflection in the upper right window was removed by a texture median filtering and a super-resolution texture with zoom factor of 4 was generated from the image sequence. The bottom of Fig.6 shows the reference image (left) and the generated median super-resolution texture without reflection (right).

Figure 5. Images 1, 8, 14, and 22 of the castle sequence.

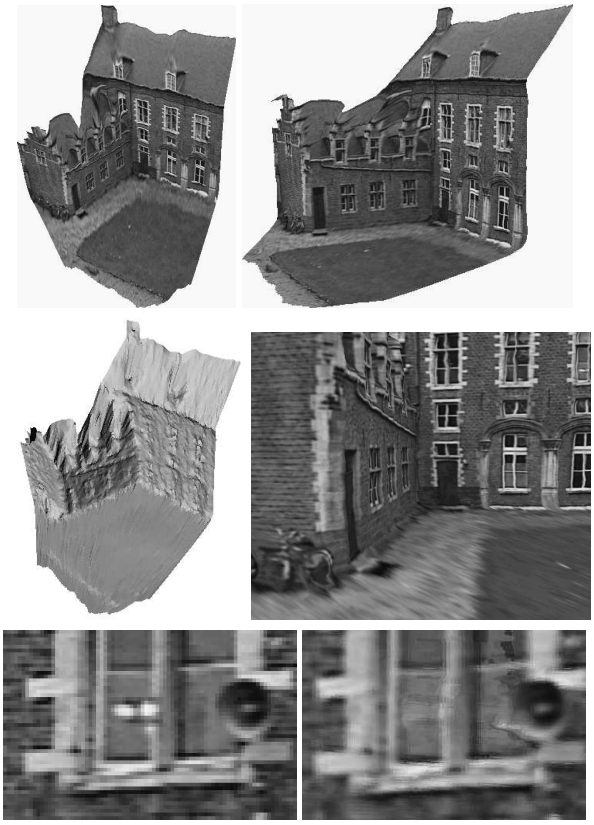

Figure 6. Top: textured views. Middle left: shaded view. Middle right: close-up view. Bottom left: 4x zoomed original region, Bottom right: median-filtered super-resolution texture.

5.1.1 Performance Evaluation

The above reconstructions showed some qualitative results. The quantitative performance of our modeling approach can be tested in different ways. One measure is the visibility V that defines the number of views linked to the reference view. The more views are linked, the higher the reliability of the measurement. Another important feature of the sequence linking algorithm is the density and accuracy of the depth maps. To describe its improvement over the 2-view disparity estimator, we define the fill rate F and the average relative depth error E as additional measures. F [in %] defines the number of estimated pixels vs. all image pixels. E [in %] is the average depth uncertainty of a point w.r.t. the point distance.

The 2-view disparity estimator is a special case of the proposed linking algorithm, hence both can be compared

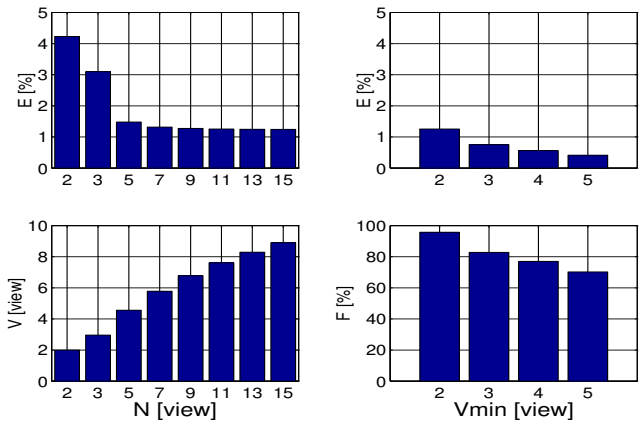

Figure 7. Statistics of the castle sequence. Left: Influence of sequence length N on visibility V and relative depth error E. Right: Influence of minimum visibility V_{min} on depth error E and fill rate F for $N = 11$.

on an equal basis. The 2-view estimator operates on the image pair $(i, i+1)$ only, while the multi view estimator operates on a sequence $1 < i < N$ with $N >= 3$. The above defined statistical measures were computed for different sequence lengths N. Figure 7 displays visibility and relative depth error for sequences from 2 to 15 images, chosen symmetrically around the reference image. The average visibility V shows that for up to 5 images nearly all views are utilized. For 15 images, at average each pixel is linked over 9 images. The amount of linking is reflected in the relative depth error that drops from 5% in the 2 view estimator to about 1.2% for 15 images.

Linking two views is the minimum case that allows triangulation. To increase the reliability of the estimates, a surface point should occur in more than two images. We can therefore impose a minimum visibility V_{min} on a depth estimate. This will reject unreliable depth estimates effectively, but will also reduce the fill-rate of the depth map. The graph in Fig. 7 (right) shows the dependency of the depth error and fill rate on minimum visibility for N=11. The fill rate drops from 92% to about 70%, but at the same time the depth error is reduced to 0.5% due to outlier rejection.

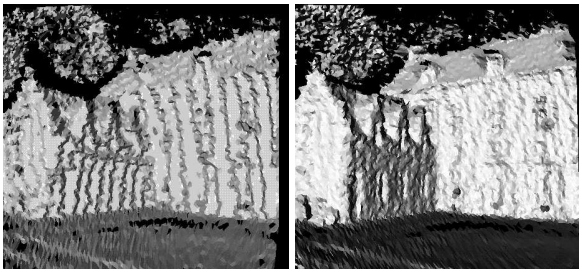

Figure 8. Shaded representation of surface model using 2 views (left) and 11 views (right).

The effects of error reduction with sequence linking can be visualized when the surface models are rendered with shading. Fig 8 shows the shaded models of the castle for the 2-view estimator (left) and the 11-view estimator (right). The shape of the 2-view model is very coarse and no details are visible, due to the quantization artifacts that are inherent to the method. In fact, the complete scene contains only about 20 discrete disparity values. This leads to a very blocky appearance of the model. In the model linked from 11 views the quantization artifacts are reduced greatly due to the increased effective baseline, and many fine details like the window niches and the doors are modeled.

5.2. Pillar sequence

As an example for varying camera parameters, eight images of a stone pillar with curved surfaces were taken. Figure 9 show 2 of the recorded images. While filming and moving away from the object the zoom was changed (2×) to keep the image size of the object constant. In spite of the changes in focal length the metric frame could be retrieved through self-calibration. In Figure 10 some perspective views of the reconstruction are given, rendered both shaded and with surface texture mapping. The shaded view shows that even most of the small details of the object are modeled.

To assess the metric properties for the pillar, 27 different lengths were measured on the real object and compared with the metric model to obtain the scale factor. Averaging all measured distances gave a consistent scale factor of 40.25 with a standard deviation of 5.4% overall. For the interior distances (avoiding the inaccuracies at the boundary of the model), the reconstruction error dropped to 2.3%. These results demonstrate the metric quality of the reconstruction even for complicated surface shapes and varying focal length.

Figure 9. Images 1 and 8 of pillar sequence.

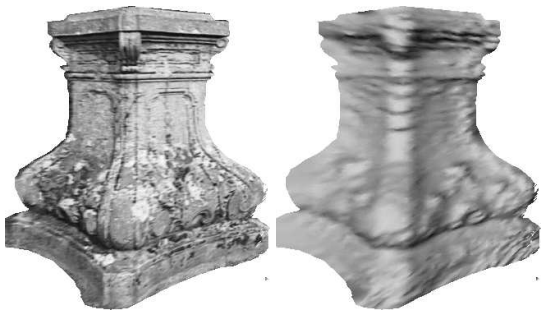

Figure 10. Perspective views of the reconstruction (with texture and shading).

5.3. Sagalassos Virtual exhibition: A Test Case

The proposed system was tested on a large variety of scenes with different cameras of varying quality (35 mm photo camera on Photo-CD, digital still camera, camcorders) and was found to work even under difficult acquisition circumstances. As a special test case, field trials were carried out at the archaeological excavation site of Sagalassos in Turkey. This is a challenging task since the archaeologists want to reconstruct even small surface details and irregular structures. Measurements with highly calibrated photogrammetric workstations failed since those systems could not withstand the high temperatures at the site. The goal of this field test was to prove the feasibility of our approach for a variety of scenes and to model objects for a virtual exhibition that can be presented over the internet.

5.3.1 Sagalassos Site

The *Site* sequence in figure 11 is a good example of a large scale modeling using off-the-shelf equipment. Nine images of the complete excavation site of Sagalassos in Turkey (extension a few km^2) were taken with a conventional photographic camera while walking along the valley rim. The film was then digitized on Photo-CD.

The *Site* reconstruction in figure 11 (bottom) gives a good overview of the valley relief. Some of the dominant objects like the Roman Bath and the Market place, as well as landmarks like big trees or stones are already modeled at this coarse scale but without any detail.

5.3.2 Detail models of the Roman bath

This sequence is a typical example of a detailed model. It consists of one part of the Roman bath that was modeled with high resolution from six images. Figure 12 shows 3 of the original images and the fused depth map. The relative depth error was estimated to 0.8% and the depth map is

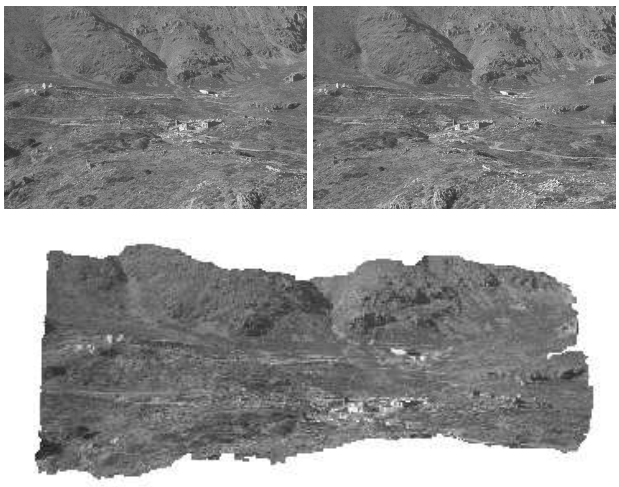

Figure 11. Images 2 and 9 of the site sequence (top) and overview model of the complete site (bottom).

very dense. Figure 13 reveals the high reconstruction quality which gives a realistic impression of the object. The close-up view confirms that each stone is modeled, including relief and small indentations. The indentations belong to erosion gaps between the stones.

5.3.3 Multiscale Integration of different Level of Detail

The models acquired at different scales are registered to each other for adaptive level of detail. For the example of the Roman bath we reconstructed models at three different scales: the site overview model, the detailed part of the bath model, and an intermediate model that overlooks the complete area of the baths. These reconstructions thus naturally fill in the different levels of details (LOD) which should be provided for optimal rendering. In Figure 14 reconstruc-

Figure 12. Images 1, 3 and 6 of Roman Bath sequence. Lower right: estimated depth map (dark = near, light = far).

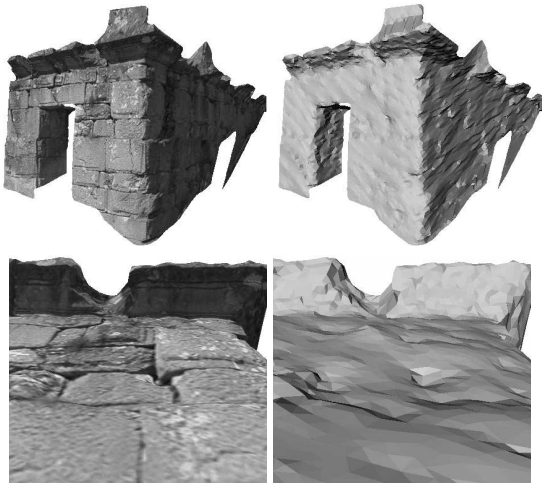

Figure 13. Textured and shaded views of Roman bath model. The close-up view shows that even small details like single stones are modeled.

tions of the Roman baths are given for the three different levels of details. These models are then registered to each other and inserted in the overview model. For scene visualization, the viewer may automatically switch between the different LOD depending on object distance, thus allowing an efficient scene representation. Fig 15 shows the integration result. On of the major problems here is the seamless integration of the surface textures from the different levels of detail. The different texture resolution of overview and detailed model leads to a severe blurring in the foreground.

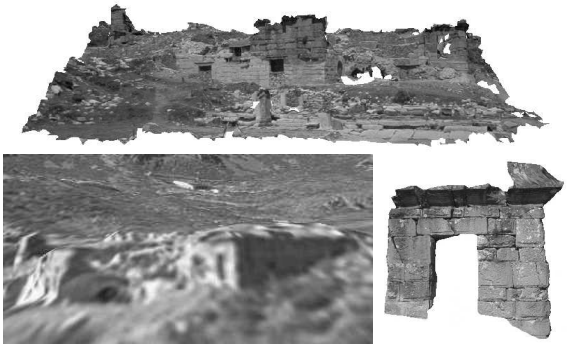

Figure 14. Models of the Roman baths at different scales: complete baths on intermediate level (top), zoom onto the baths in the overview model of Figure 11 (bottom left), detailed right corner of the baths from fig 13 (bottom right).

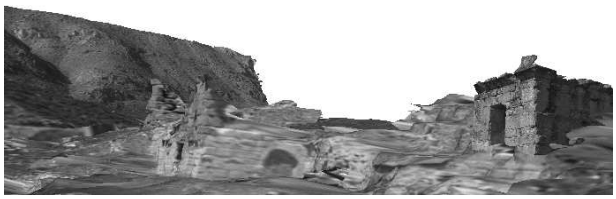

Figure 15. Multilevel model of Roman bath with three LOD.

On way to handle this problem would be to synthesize texture for the overview model with higher resolution to improve the appearance.

5.3.4 Augmenting the site with CAD models

Another interesting approach is the merging of the reconstructed real site model with hypothesized buildings that were constructed from archaeological findings. This technique allows the reconstruction of the site as it once might have been. In the case of Sagalassos some building hypothesis were translated to CAD models [23] and integrated with our reconstructions. The result can be seen in Figure 16. This scene augmentation is a powerful visualization tool that fits naturally with our modeling approach.

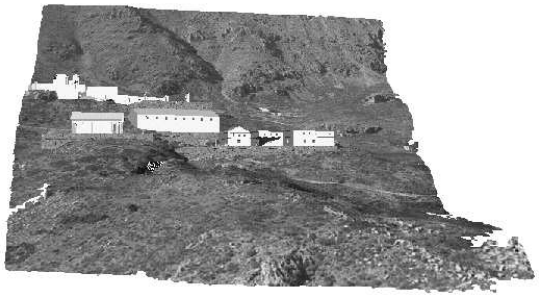

Figure 16. Virtualized landscape of Sagalassos combined with CAD-models of reconstructed monuments.

6. Conclusion

An automatic 3D scene reconstruction system was described that is capable of building metric textured 3D models from images of a freely moving, uncalibrated camera. The technique extracts metric surface models without prior knowledge about the scene or the camera other than assuming rigidity of the objects. The approach was tested with different off-the-shelf camera types and for scenes of varying scale and complexity. The algorithms estimate very dense depth maps and achieve a depth accuracy of typically 1% of scene depth. The approach was tested on a variety of

real outdoor sequences and has proven its robustness. The high quality of the reconstructed objects, the different scene types, and the use of off-the-shelf equipment demonstrate the versatility and flexibility of the proposed scene reconstruction approach.

Work remains to be done in constructing multiscale models from varying LOD. The surface texture resolution of the different levels need to be adapted and the models need to be integrated geometrically into a consistent surface model.

Acknowledgments

We would like to thank Andrew Zisserman and his group from Oxford for supplying us with robust projective reconstruction software. I. Cox and L. Falkenhagen supplied versions of their disparity matching code.

References

[1] P. Beardsley, P. Torr and A. Zisserman: 3D Model Acquisition from Extended Image Sequences. In: B. Buxton, R. Cipolla(Eds.) Computer Vision - ECCV 96, Cambridge, UK., vol.2, pp.683-695. Lecture Notes in Computer Science, Vol. 1064. Springer Verlag, 1996.

[2] Y. Chen and G. Medioni: Object Modeling by Registration of Multiple Range Images. Proc. Int. Conf. on Robotics and Automation, Sacramento CA, pp. 2724-2729, 1991.

[3] I. J. Cox, S. L. Hingorani, and S. B. Rao: A Maximum Likelihood Stereo Algorithm. Computer Vision and Image Understanding, Vol. 63, No. 3, May 1996.

[4] P.E. Debevec, C.J. Taylor, J. Malik: Modeling and Rendering Architecture from Photographs: A Hybrid Geometry-and Image-Based Approach. Proceedings SIGGRAPH '96, pp 11-20, ACM Press, New York, 1996.

[5] L.Falkenhagen: Hierarchical Block-Based Disparity Estimation Considering Neighborhood Constraints. Intern. Workshop on SNHC and 3D Imaging, Rhodes, Greece, Sept. 1997.

[6] O.Faugeras: Three-Dimensional Computer Vision - a geometric viewpoint. MIT-Press, 1993.

[7] O. Faugeras: What can be seen in three dimensions with an uncalibrated stereo rig. *Proc. ECCV'92*

[8] O. Faugeras, Q.-T. Luong and S. Maybank: Camera self-calibration - Theory and experiments. *Proc. ECCV'92*, pp.321-334.

[9] A.W. Fitzgibbon, G. Cross, A. Zisserman: Automatic 3D Model Construction for Turntable Sequences. Proceedings SMILE, LNCS 1506, Springer 1998.

[10] G.L.Gimel'farb: Symmetrical approach to the problem of automating stereoscopic measurements in photogrammetry. Cybernetics, 1979, 15(2), 235-247; Consultants Bureau, N.Y.

[11] S. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen: The Lumigraph. Proceedings SIGGRAPH '96, pp 43–54, ACM Press, New York, 1996.

[12] R. Hartley: Estimation of relative camera positions for uncalibrated cameras. *Proc. ECCV'92*, pp.579-587.

[13] A. Heyden and K. Åström: Euclidean Reconstruction from Image Sequences with Varying and Unknown Focal Length and Principal Point. *Proc. CVPR'97.*

[14] M. Irani and S. Peleg: Super Resolution from Image Sequences. Tenth International Conference on Pattern Recognition (Atlantic City, NJ, June 16-21, 1990), IEEE Catalog No. 90CH2898-5, 1990, subconference C, 115-120.

[15] M. Irani, P. Anandan, S. Hsu: Mosaic-based representations of video sequences and their applications. Proc. ICCV'95, Boston, USA, 1995.

[16] R. Koch: 3-D Surface Reconstruction from Stereoscopic Image Sequences, *Proc. ICCV'95*, Cambridge, USA, June 1995.

[17] R. Koch: Surface Segmentation and Modeling of 3-D Polygonal Objects from Stereoscopic Image Pairs. *Proc. ICPR'96*, Vienna 1996.

[18] R. Koch, M. Pollefeys, and L. Van Gool: Multi Viewpoint Stereo from Uncalibrated Video Sequences. *Proc. ECCV'98*, Freiburg, June 1998.

[19] R. Koch, B. Heigl, M. Pollefeys, L. Van Gool, H. Niemann: Calibration of Hand-held Camera Sequences for Plenoptic Modeling. Proc. ICCV99, Corfu, 1999.

[20] R. Koch, M. Pollefeys, L. Van Gool: Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images. Proc. DAGM Symposium 99, Bonn, Germany, 1999.

[21] A. Kochan: Eine Methodenbank zur Evaluierung von Stereo-Vision-Verfahren. Ph.D. Thesis, TU Berlin, June 1991.

[22] M. Levoy, P. Hanrahan: Lightfield Rendering. Proceedings SIGGRAPH '96, pp 31–42, ACM Press, New York, 1996.

[23] F. Martens, P. Legrand, J. Legrand, L. Loots and M. Waelkens, Computer Aided Design and Archeology at Sagalassos: methodology and possibilities of 3D computer reconstructions of archaeological sites. *Virtual Reality in Archaeology*, Eds. J.A. Barcelo, M. Forte and D. Sanders.

[24] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system", *Proc. SIGGRAPH'95*, pp. 39-46, 1995.

[25] E. Ofek, E. Shilat, A. Rappoport, M. Werman: Highlight and Reflection-Independent Multiresolution Textures from Image Sequences. IEEE Computer Graphics and Applications vol. 17 (2), March-April 1997.

[26] M. Pollefeys, L. Van Gool and M. Proesmans: Euclidean 3D Reconstruction from Image Sequences with Variable Focal Lengths. *Proc. ECCV'96*, vol.1, pp. 31-42.

[27] M. Pollefeys and L. Van Gool: Self-calibration from the absolute conic on the plane at infinity. *Proc. CAIP'97.*

[28] M. Pollefeys, R. Koch and L. Van Gool: Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. Proc. ICCV'98, Bombay, India, Jan. 1998. Also accepted for publication in: International Journal on Computer Vision, Marr Price Special Issue, 1998.

[29] H. Shum, R. Szelisky, S. Baker, M. Han, P. Anandan: Interactive 3D modeling from multiple images using scene regularities. Proceedings SMILE, LNCS 1506, Springer, 1998.

[30] R. Szeliski and H. Shum: Creating full view panoramic image mosaics and texture-mapped models. Computer Graphics (SIGGRAPH'97), 1997.

[31] S. Teller: Automated Urban Model Aquisition: Project Rationals and Status. Proc. DARPA IU Workshop 1998.

[32] P.H.S. Torr: Motion Segmentation and Outlier Detection. PhD thesis, University of Oxford, UK, 1995.

[33] B. Triggs: The Absolute Quadric. *Proc. CVPR'97.*