3D-SCENE MODELING FROM IMAGE SEQUENCES

Reinhard Koch

Institute of Computer Science and Applied Mathematics, Christian-Albrechts-University Kiel, Germany
rk@informatik.uni-kiel.de

KEY WORDS: Computer Vision, Camera Calibration, Structure from Motion, Scene Visualisation, image-based rendering.

ABSTRACT

This contribution gives an overview of automatic 3D scene reconstruction and visualisation from uncalibrated and handheld camera image sequences. We address specifically the problems that are associated with calibration and visual-geometric reconstruction of complex scenes with occlusions and view-dependent surfaces. The scene is then represented by large sets of calibrated real viewpoints with image texture and depth maps. Novel views are synthesized from this representation with view-dependent image-based rendering techniques at interactive rates.

1 INTRODUCTION

3D scene analysis is an important topic for a variety of applications. In visual robot guidance, fast and precise geometric representation of the surrounding environment as well as precise self-localisation is crucial for efficient path planning, obstacle avoidance, and collision-free navigation. The precise visual appearence of the scene is only of secondary importance. Visual appearence is becoming important in visual servoing (6) where the goal is to position a vision-guided robot such that the observed real image matches a stored reference image. Augmented and Mixed Reality (2) on the other hand is a rapidly growing field that aims at seamless integration of virtual objects into live film footage with highest visual quality, while the geometric properties are only of importance to help achieving the primary goal of visual insertion. Here, precise camera tracking and calibration must be achieved to avoid object jitter. While Augmented Reality is mostly concerned with realtime video tracking of predefined markers and direct visual insertion of virtual objects into the live video stream, Mixed Reality goes even further and aims at the intertwining of real and virtual objects in a mixed real-virtual space. Interdependences of occlusions, shading and reflections between both real and virtual objects have to be taken into account. No predefined markers are used but the real scene itself is tracked without markers.

In this contribution we are concerned with Mixed Reality in extended virtual studio film production (1) where both, camera tracking and visual image interpolation, is needed. In this scenario, a real background scene is recorded and virtualized such that virtual views of the real scene can be extrapolated from the prerecorded scene. Real actors that have been recorded in a virtual studio, and computer-generated virtual ojects are then both merged with the background scene in a seamless fashion.

In Mixed Reality applications it is neccessary to reconstruct the 3D background scene with high fidelity. In case of simple scene geometry, few images may suffice to obtain a 3D surface model that will be textured from the real images. Novel views of the scene can then be rendered easily from the model. Typical examples are architectural or landscape models with mostly diffuse and opaque surfaces. In other cases, however, scene geometry and surface properties may be very complex and it might not be possible to reconstruct the scene geometry in all details. In this case one may resort to lightfield rendering (28) by reconstructing the visual properties of the surface reflection. This is possible only in very restricted environments because a very dense image sampling is needed for this approach. We propose a hybrid *visual-geometric modeling* approach where a partial geometric reconstruction (calibrated depth maps) is combined with unstructured lumigraph rendering (16) to capture the visual appearance of the scene.

Visual-geometric reconstruction aims at capturing the visual appearence of a complex scene by first approximating the geometric scene features and then superimposing the precise visual features over the approximation. The real scene is scanned by one or more video or photo cameras. The images from these cameras are termed *real views*. As we may want to capture complex 3D scenes with occlusions and possibly view-dependent surface reflections, we will need to capture very many real view points that cover the complete viewing space. Therefore we have to register real views that span all possible views of a viewing volume to capture all possible scene details.

Virtual views of the scene are novel views that are rendered by extrapolating the visual appearence of the scene from the most similar real views. The local geometry of the scene as seen from a real view is captured by estimating a depth map for each view. Parallax effects between real and novel views are compensated for by warping the real views towards the virtual view according to the local depth map. Thus, for visual-geometric reconstruction and rendering the following three main steps are needed:

1. Estimate 3D position and calibration of each real view in world coordinates,
2. Compute local depth map for each real view,
3. Render novel views from the reconstructed real views.

In the following sections we will describe the different steps of this hybrid approach in more detail. In section 2 we will explain the camera tracking and calibration step. Section 3 deals with dense depth estimation from multiple real view points. In Section 4 different methods to render novel views are discussed.

2 CAMERA TRACKING AND CALIBRATION

This work is embedded in the context of *uncalibrated Structure From Motion* (SFM) where camera calibration and scene geometry are recovered from images of the scene alone without the need

for further scene or camera information. Faugeras and Hartley first demonstrated how to obtain uncalibrated projective reconstructions from image point matches alone (10, 17). Beardsley et al. (3) proposed a scheme to obtain projective calibration and 3D structure by robustly tracking salient feature points throughout an image sequence. Since then, researchers have tried to find ways to upgrade these reconstructions to metric (i.e. Euclidean but unknown scale, see (11, 37, 32)).

When very long image sequences have to be processed there is a risk of calibration failure due to several factors. For one, the calibration as described above is built sequentially by adding one view at a time. This may result in accumulation errors that introduce a bias to the calibration. Secondly, if a single image in the sequence is not matched, the complete calibration fails. Finally, sequential calibration does not exploit the specific image acquisition structure used in this approach to sample the viewing sphere. In our visual-geometric approach, multiple cameras may be used to scan a 2D viewing surface by moving a rigid multi-camera rig throughout the viewing space of interest. We have therefore developed a multi-camera calibration algorithm that allows to actually weave the real views into a connected 2D viewpoint mesh (25, 19).

2.1 Image pair matching

The projection of scene points onto an image by a camera may be modeled by the equation $\mathbf{x} = P\mathbf{X}$. The image point in projective coordinates is $\mathbf{x} = [x, y, w]^T$, where $\mathbf{X} = [X, Y, Z, 1]^T$ is the world point and P is the 3×4 camera projection matrix. The matrix P is a rank-3 matrix. If it can be decomposed as $P = K[R^T | -R^T\mathbf{t}]$ where the rotation matrix R and the translation vector $\mathbf{t}$ represent the Euclidian transformation between the camera and the world coordinate system. The intrinsic parameters of the camera are contained in the matrix K which is an upper triangular matrix

$$K = \begin{bmatrix} f & s & c_x \\ 0 & a \cdot f & c_y \\ 0 & 0 & 1 \end{bmatrix}, \qquad (1)$$

where f is the focal length of the camera expressed in pixel units. The aspect ratio a of the camera is the ratio between the size of a pixel in x-direction and the size of a pixel in y-direction. The principal point of the camera is (c_x, c_y) and s is a skew parameter which models the angle between columns and rows of the CCD-sensor.

The relation between two consecutive frames is described by the fundamental matrix (18) if the camera is translated between these frames. The fundamental matrix $F_{j,i}$ maps points from camera j to lines in camera i. Furthermore the fundamental matrix can be decomposed into a homography $H^{\pi}_{j,i}$ which maps over the plane π and an epipole e which is the projection of the cameraa center of camera j in camera i

$$F_{j,i} = [e]_x H^{\pi}_{j,i}, \qquad (2)$$

where $[\cdot]_x$ is the cross product matrix. The epipole is contained in the null space of F: $F_{i,j} \cdot e = 0$.

$F_{i,j}$ can be computed robustly with the RANSAC (RANdom SAmpling Consensus) method (36). A minimum set of 7 features correspondences is picked from a large list of potential image matches to compute a specific F. For this particular F the support is computed from the other potential matches. This procedure is repeated randomly to obtain the most likely F_{ik} with best support in feature correspondence. From the F we can initialize a projective camera pair that defines a projective frame for reconstruction of the corresponding point pairs (12, 31).

2.2 Multi-viewpoint matching

Once we have obtained the projection matrices we can triangulate the corresponding image features to obtain the corresponding 3D object features. The object points are determined such that their reprojection error in the images is minimized. In addition we compute the point uncertainty covariance to keep track of measurement uncertainties. The 3D object points serve as the *memory* for consistent camera tracking, and it is desirable to track the projection of the 3D points through as many images as possible. This process is repeated by adding new viewpoints and correspondences throughout the sequence.

Although it can be shown that a single camera suffices to obtain a mesh of camera view points by simply waving the camera around the scene of interest in a zig-zag scan (24, 25), a more reliable means is to use an n-camera rig that simultaneously captures a time-synchronized 1D sequence of views. When this rig is swept along the scene of interest, a regular 2D viewpoint surface is generated that can be calibrated very reliably by concatenating the different views in space and time. For each recording time, a number of n simultaneous real views of the scene are obtained and can be used to reconstruct even time-varying scenes. When the camera rig is moved, a sequence of k images is obtained for each of the n cameras. Thus, one may obtain a 2D *viewpoint surface* of $k \times n$ views by simply sweeping the camera rig throughout the scene (26). For each time step, correspondences between adjacent cameras on the rig are searched and fundamental matrices are computed between all n cameras on the rig. Ideally, the fundamental geometry should be identical for each time step, but due to slight vibrations and torsion of the rod during motion, small deviations of the fundamental geometry have to be accounted for. Additionally, the motion of the rod can be tracked by estimating the position of each camera on the rod simultaneously between adjacent time steps. By concatenating the camera motion in time and the different cameras on the rod, a 2D viewpoint surface is built that concatenates all real views.

2.3 Camera selfcalibration

The camera tracking as described above will generate a projective reconstruction with a projective ambiguity. The fundamental matrix is invariant to any projective skew. This means that the projection matrices P_j and P_i lead to the same fundamental matrix $F_{j,i}$ as the projectively skewed projection matrices $\tilde{P}_j$ and $\tilde{P}_i$ (18). This property poses a problem when computing camera projection matrices from Fundamental matrices. Most techniques for calibration of translating and rotating cameras at first estimate the projective camera matrices $\tilde{P}_i$ and the positions $\tilde{X}_k$ of the scene points from the image data with a Structure-from-Motion approach, as described above. The estimated projection matrices $\tilde{P}_i$ and the reconstructed scene points may be projectively skewed by an unknown projective transformation $H_{4\times4}$. Thus only the skewed projection matrices $\tilde{P}_i = PH_{4\times4}$ and the inversely skewed scene points $\tilde{X} = H^{-1}_{4\times4}X$ are estimated instead of the true entities. For uncalibrated cameras one cannot avoid this skew and selfcalibration for the general case is concerned mainly with estimating the projective skew matrix $H_{4\times4}$ e.g. via the DIAC (18) or the absolute quadric (37, 32). Camera selfcalibration from unknown general motion and constant intrinsics has been dicussed in (10, 29, 20). For varying intrinsics and general camera motion the selfcalibration was proven by (37, 21, 32). All these approaches for selfcalibration of cameras only use the images of the cameras themselves for the calibration.

Furthermore there exist approaches for camera calibration with some structural constraints on the scene. For example an in-

teresting approach was recently proposed by Rother and Carlsson (33) who jointly estimate fundamental matrices and homographies from a moving camera that observes the scene and some reference plane in the scene simultaneously. The homography induced by the reference plane generates constraints that are similar to a rotation sensor and selfcalibration can be computed linearily. This approach needs information about the scene in contrast to our approach which applies constraints only to the imaging device.

Only a few approaches exist to combine image analysis and external rotation information for selfcalibration. In (34, 4) the calibration of rotating cameras with constant intrinsics and known rotation was discussed. They use nonlinear optimization to estimate the camera parameters. A linear approach for an arbitrarily moving camera was developed in (14, 13). That approach is able to compute linearily a full camera calibration for a rotating camera and a partial calibration for freely moving camera. More often, calibrated cameras are used in conjunction with rotation sensors to stabilize sensor drift (30). Thus, if external rotation data is available it can be used with advantage to stabilize tracking and to robustly recover metric reconstruction of the scene.

As an example for our multi-camera tracking system a mixed-reality application was developed where the interior of the London Museum of Natural History was scanned with a 4-camera system. 4 cameras were mounted in a row on a vertical rod (see figure 1) and the rig was moved horizontally along parts of the entrance hall while scanning the hallways, stairs, and a large dinosaur sceleton. While moving, images were taken at about 3 frames/s with all 4 cameras simultaneously. The camera tracking was performed by 2D-viewpoint meshing (24) with additional consideration of camera motion constraints. Prediction of potential camera pose is possible because we know that the cameras are mounted rigidly on the rig. We also can exploit the fact that all 4 cameras grab images simultaneously (26). Figure 1 (left) shows the portable acquisition system with 4 cameras on the rod and 2 synchronized laptops attached by a digital firewire connection. Figure 1 (right) gives an overview of parts of the museum hall with the dinosaur sceleton that was scanned. The camera rig was moved alongside the sceleton and 80x4 viewpoints were recorded over the length of the sceleton. Figure 2 displays the camera tracking with the estimated 360 camera viewpoints as little pyramids and the reconstructed 3D feature point cloud obtained by the SfM method. The outline of the skeleton and the back walls is reconstructed very well.

Figure 1: Left: portable image capture system. Right: overview of the scene to be reconstructed.

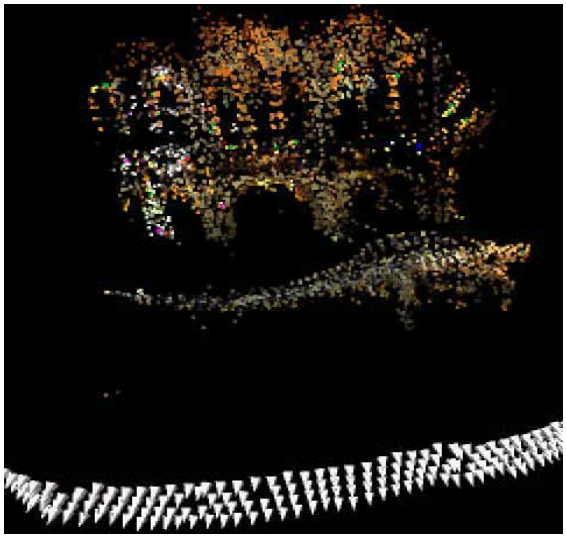

Figure 2: Camera viewpoints and reconstructed 3D feature points of the dinosaur and walls as seen by the 4-camera-rig.

3 DEPTH ESTIMATION

Once we have retrieved the metric calibration of the cameras we can use image correspondence techniques to estimate scene depth. We rely on stereo matching techniques that were developed for dense and reliable matching between adjacent views. The small baseline paradigm suffices here since we use a rather dense sampling of viewpoints.

3.1 Stereoscopic disparity estimation

Dense stereo reconstruction has been investigated for decades but still poses a challenging research problem. This is because we have to rely on image measurements alone and still want to reconstruct small details (needs small measurement window) with high reliability (needs large measurement window). Traditionally, pairwise rectified stereo image were analysed that exploit some constraints along the epipolar line as in (15, 38, 5). Recently, generalized approaches were introduced that can handle multiple images, varying windows etc.(35, 27). Also, realtime stereo image analysis has become almost a reality with the exploitation of the new generation of very fast programmable graphical processing units for image analysis (39). We are currently using a hybrid approach that needs rectified stereo pairs but can be extended to multiview depth processing.

For dense correspondence matching an area-based disparity estimator is employed on rectified images. The matcher searches at each pixel in one image for maximum normalized cross correlation in the other image by shifting a small measurement window (kernel size 7x7) along the corresponding scan line. Dynamic programming is used to evaluate extended image neighborhood relationships and a pyramidal estimation scheme allows to reliably deal with very large disparity ranges (9).

3.2 Multi-camera depth map fusion

For a single-pair disparity map, object occlusions along the epipolar line cannot be resolved and undefined image regions (occlusion shadows) remain. The occlusions can be removed with multi-image disparity estimation. The geometry of the viewpoint mesh is especially suited for further improvement with a multi viewpoint refinement (22). For each viewpoint a number of adjacent viewpoints exist that allow correspondence matching. Since the

different views are rather similar we will observe every object point in many nearby images. This redundancy can also be exploited to verify the depth estimation for each object point, and to refine the depth values to high accuracy.

We can further exploit the imaging geometry of the multi-camera rig to fuse the depth maps from neighboring images into a dense and consistent single depth map. For each real view, we can compute several pairwise disparity maps from adjacent views in the viewpoint surface. The topology of the viewpoint mesh was established during camera tracking as described in section 2.2. Since we have a 2D connectivity between views in horizontal, vertical, and even diagonal directions, the epipolar lines overlap in all possible directions. Hence, occlusion shadows left undefined from singe-pair disparity maps are filled from other view points and a potentially 100% dense disparity map is generated.

Additionally, each 3D scene point is seen many times from different viewing directions, and this allows to robustly verify its 3D position. For a single image point in a particular real view, all corresponding image points of the adjacent views are computed. After triangulating all corresponding pairs, the best 3D point position can be computed by robust statistics and outlier detection, eiminating false depth values (22). Thus, reliable and dense depth maps are generated from the camera rig.

As an example, the Dinosaur scene was evaluated and depth maps were generated with different neighbourhoods. Figure 3 shows an original image (top left) and the corresponding depth maps for varying number of images taken into consideration. The depth maps become denser and more accurate as more and more neighboring images are evaluated. For images in an 8-neighbourhood, the fill rate approaches 100% (figure 3, top right). However, some outliers (white streaks) can be observed which are due to the repetitive structures of the ribs. These errors must be eliminated using prior scene knowledge since we know that the scene is of final extent. A bounding box can be allocated that effectively eliminates most gross outliers.

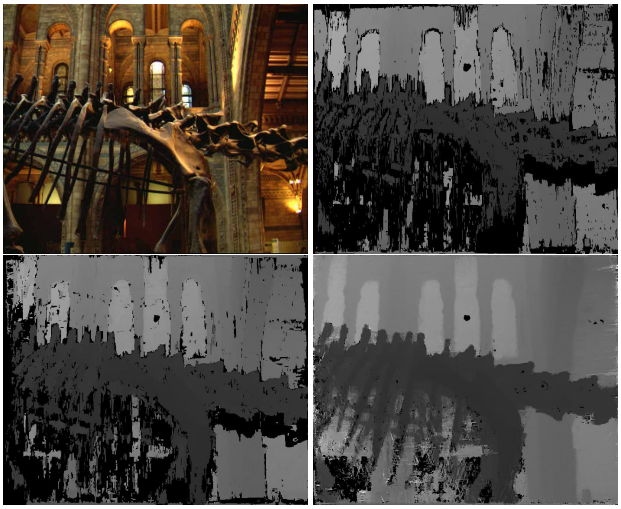

Figure 3: Original image (top left) and depth maps computed from the Museum sequence. Top right: Depth map from single image pair, vertically rectified (light=far, dark=near, black=undefined). Bottom left: 1D sequential depth fusion from 4 vertically adjacent views. Bottom right: Depth fusion from a 2D neighbourhood (8-connectivity) with 8 adjacent neighbours (horizontally, vertical, diagonal).

4 IMAGE-BASED INTERACTIVE RENDERING

The calibrated views and the preprocessed depth maps are used as input to the image-based interactive rendering engine. The user controls a virtual camera which renders the scene from novel viewpoints. The novel view is interpolated from the set of real calibrated camera images and their associated depth maps. During rendering it must be decided which camera images are best suited to interpolate the novel view, how to compensate for depth changes and how to blend the texture from the different images. For large and complex scenes hundreds or even thousands of images have to be processed. All these operations must be performed at interactive frame rates of 10 fps or more.

We have to address the following issues for interactive rendering:

- Selection of best real camera views,
- Multiview depth fusion,
- viewpoint-adaptive texture blending.

4.1 Camera ranking and selection

For each novel view to render, it must be decided which real cameras to use. Several criteria are relevant for this decision. We have developed a ranking criterion for ordering the real cameras w.r.t. the current virtual view(7). The criterion is based on the *normalized distance* between real and virtual camera, the *viewing angle* between the optical axes of the cameras and the *visibility*, which gives a measure of how much of the real scene can be transfered to the virtual view.

All three criteria are weighted and combined into one scalar value which represents the ability of a particular real camera to synthesize the new view. After calculating the quality of each camera, the list of valid cameras is sorted according to quality. During rendering it is finally decided how many of the best suited cameras are selected for view interpolation.

4.2 Multiview depth fusion

The ranked cameras are now used to interpolate novel views. Since the novel view may cover a field of view that is larger than any real camera view, we have to fuse views from different cameras into one locally consistent image. To efficiently warp image texture from different real views into the novel viewpoint we generate warping surfaces that approximates the geometry of the scene. Two different approximations were tested: an interpolating connected triangular surface mesh and unconnected planar quadrangles (Quads). The interpolating surface mesh guarantees that for each image points in the virtual view some texture is mapped, however the mapping might be distorted at object boundaries (7). The unconnected quads handle occluding boundaries better and distortions are much less visible, but at occlusion boundaries there might be some textureless regions in the virtual image that appear as holes (8). In both approaches, we merge the data from the best ranked l views to adaptively build the novel view.

Rendering from triangular patches: For interpolation from the connected triangular surface, a regular triangular 2D-grid is placed in the image plane of the virtual camera. This warping surface will be updated for each novel virtual viewpoint. The spacing of this grid can be scaled to the complexity of the scene. For each triangular surface patch of the grid we test which real camera gives the least distorted mapping according to distance,

visibility and viewing angle of the real views. The approximating 3D surface depth is than mapped from the depth values of the best real view, if available, and the surface patch is textured from one or more best real views, depending on the texturing mode (see section 4.3) Grid points that have no valid 3D depth value associated are interpolated from adjacent valid points.

Rendering from Quads: Quads are generated just the inverse way. We start from the real views and generate a quadtree of the depth map that represents the depth map with boundary-adaptive cells. Therefore, we automatically build a hierarchical depth representation that segments the depth map well at depth boundaries and efficiently represents large continuous or planar depth regions. We start with a preselected coarse quadtree subdivision of the depth map. For each cell we compute the approximation error to the mean surface plane and subdivide it further if the approximation threshold is exceeded. The quadtree is stored in a Level of Detail hierarchy file and can be rendered very efficiently with OpenGL rendering hardware. For each novel virtual view, all Quadtrees of the selected cameras are projected and rendered into the virtual view.

4.3 Texturing

The texturing step effectively maps the image texture of real cameras into the virtual view with the help of the viewpoint-adaptive geometry. Several slightly different methods for texturing are considered. The most simple one is to chose the best ranked camera as texture source. If this real camera is not too far away from the virtual camera and both have a similar field of view, the results are good. This is the fastest texturing method since switching between different textures in one render cycle is not necessary and each triangle has to be drawn only once. Problems arise when parts of the mesh are not seen from the selected camera. These parts remain untextured.

To texture all triangles properly it is necessary to select the texture according to the cameras where the geometry originated from. The triangle vertices are depth sample points where the originating real camera is known. However, since each vertex is generated independently, a triangle may have vertices from up to three cameras. Here one may decide to either select the best-ranked camera (single-texture mode) or to blend all associated camera textures on the triangle (multi-texture mode). Proper blending of all textures will result in smoother transition between views but with higher rendering costs for multi-pass rendering.

Sharp edges between textures can be avoided by multi-texturing and blending. Each triangle is drawn three times using the textures associated with the three cameras (multi-camera, multi-texture mode). On modern graphics hardware it is also possible to use single-pass multi-texturing. Different texture units are loaded with the three textures and then the triangle is drawn only once. This gives a speed-up of approximatly 30% compared to the multi-pass texturing. The performance gain is not factor 3 as one would expect, because for each triangle the texture units have to be reloaded which is quite expensive. A detailed comparison of the different texture modes can be found in (7).

As an example, the Dinosaur scene was rendered with the proposed methods. The scene is particular difficult since it contains very many small and repetitive structures with occlusions.

Figure 4 (top left) shows the camera track for 100 real views. To compare the image synthesis with ground truth, one real view was taken out of the sequence and re-synthesized with the triangular mesh interpolation. Figure 4 (top right) shows the original ground truth reference view. The synthesized reference view is shown on the bottom part of figure 4. It can be seen that visible artifacts occur mostly near the occlusion boundaries of the ribs. The synthesized ribs are clearly distorted due to boundary distortion effects.

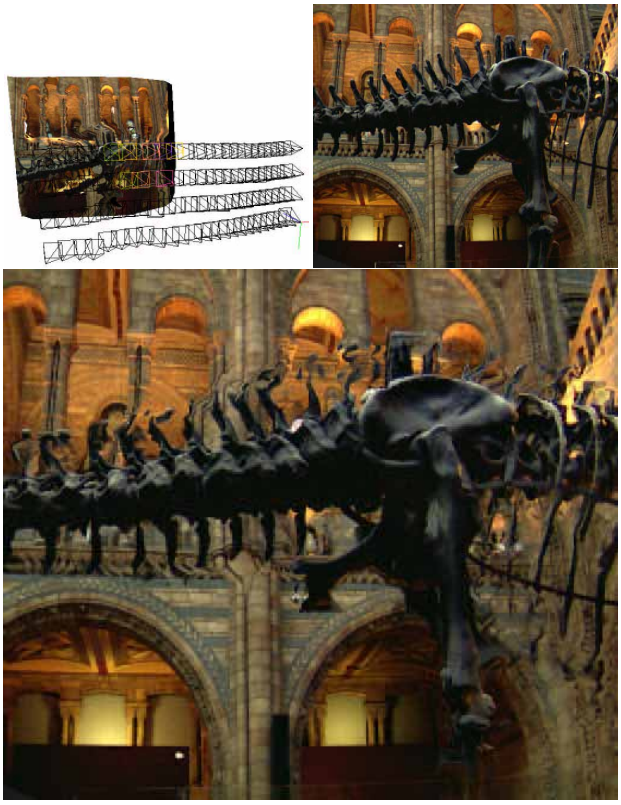

Figure 4: Ground truth comparison between real and synthesized views. Top left: Position of real camera views. Top right: selected real image as visual ground truth. Bottom: synthesized reference image, showing image distortions in the rib region.

The quads perform much better in regions near occlusions since they fit better to the occluding boundary. Figure 5 shows a synthesized view of the dinosaur with the quad method. Here one can see that the ribs are reproduced much better, however one can still see visual artifacts from spurious quads that were generated from the depth outliers as described in section 3.2.

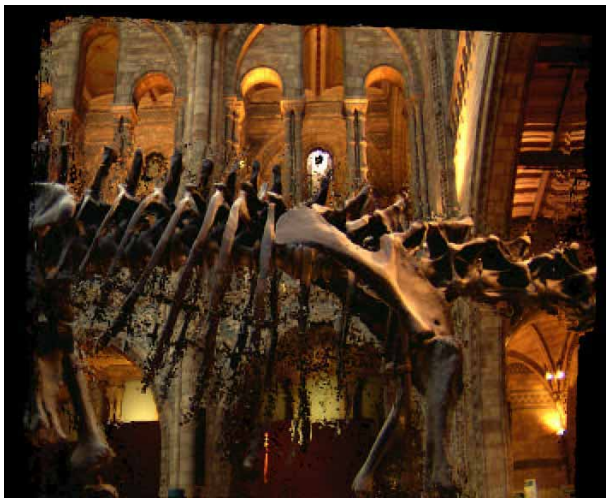

Figure 5: Synthesized view rendered with quads.

Figure 6: Original views number 3, 25,

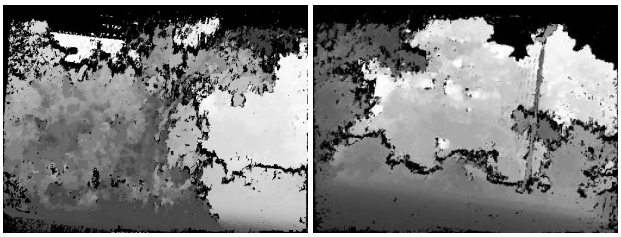

Figure 7: Depth maps number 3, 25,

5 EXPERIMENTS WITH OUTDOOR SCENE

We used a 2-camera rig in an unstructured outdoor environment. The cameras were mounted horizontally on the rod about 30 cm apart and 106 image pairs were taken while walking through a parking lot with cars, bushes, trees, and lamp posts. The camera trajectory is about 70 m long and the scene has a horizontal extension of about 100x80 m. Figure 6 and 9 show 4 original views, covering the length of the scene, and Figure 7 and 10 the associated estimated depth maps. Figure 8 shows the camera track with some of the tracked 3D feature points and the camera positions as little pyramids. Figure 11 shows a synthesized novel view of the scene. The synthesized image was generated with triangular mesh interpolation.

6 CONCLUSIONS

We have discussed an approach to render novel views from large sets of real images. The images are calibrated automatically and dense depth maps are computed from the calibrated views using multi-view configurations. These visual-geometric representations are then used to synthesize novel viewpoints by interpolating image textures from nearby real views. Different rendering techniques were developed that can handle occluded regions and large amounts of real viewpoints at interactive rendering rates of 10 fps and more.

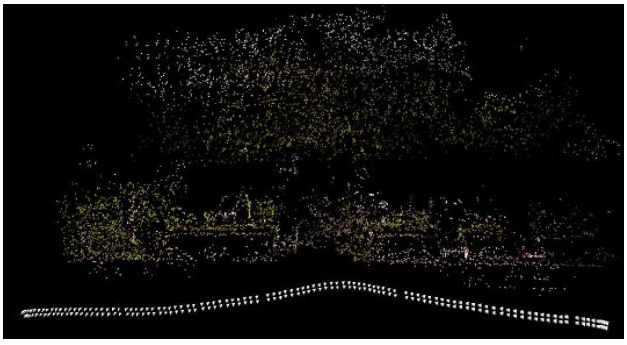

Figure 8: Camera tracks and tracked 3D feature points

Figure 9: original views number 70, 100.

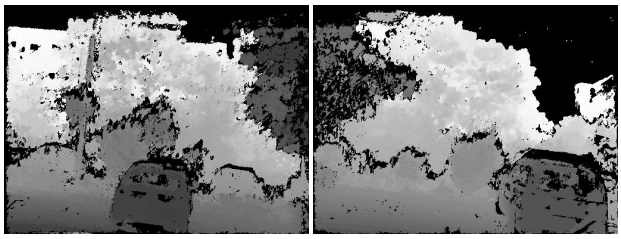

Figure 10: depth maps number 75, and 100.

REFERENCES

G. Bazzoni, E. Bianchi, O. Grau, A. Knox, R. Koch, F. Lavagetto, A. Parkinson, F. Pedersini, A. Sarti, G. Thomas, S. Tubaro: The ORIGAMI Project – advanced tools and techniques for high-end mixing and interaction between real and virtual content. IEEE Proceedings of 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT'02) June 19 - 21, 2002 Padova, Italy.

R. Behringer: Proceedings of IEEE and ACM International Symposium on Mixed and Augmented Reality Sept. 30 - Oct. 1, 2002 in Darmstadt, Germany. Further contacts at The Augmented Reality homepage: www.augmented-reality.org

P. Beardsley, P. Torr and A. Zisserman: 3D Model Acquisition from Extended Image Sequences. *ECCV 96*, LNCS 1064, vol.2, pp.683-695.Springer 1996.

F. Du and M. Brady: Self-calibration of the intrinsic parameters of cameras for active vision systems. *Proceedings CVPR*, 1993.

Figure 11: Synthesized novel view of parking lot scene.

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.

Peter I. Corke: Visual Control of Robots: High Performance Visual Servoing. Research Studies Press (John Wiley), 1996.

J.-F. Evers-Senne, R. Koch: Image Based Rendering with View Dependent Geometry from an Uncalibrated Mobile Multi-Camera System. Computer Graphics Forum. Proceedings Eurographics 2003, Granada, Spain, Sept. 2003.

J.-F. Evers-Senne, R. Koch: Interactive Rendering with View-Dependent Geometry and Texture. Presented at Siggraph Sketches and Applications, SIGGRAPH 2003, San Diego, July 2003.

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

O. D. Faugeras and M. Herbert: The representation, recognition and locating of 3-D objects. *Intl. J. of Robotics Research*, 1992.

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

A. Fitzgibbon and A. Zisserman: Automatic Camera Recovery for Closed or Open Image Sequences. Proceedings ECCV'98. LNCS Vol. 1406, Springer, 1998.

J.-M. Frahm and Reinhard Koch: Robust Camera Calibration from Images and Rotation Data. In Proceedings of DAGM, 2003.

J.-M. Frahm and Reinhard Koch:Camera Calibration with Known Rotation. Proceedings of IEEE Int. Conf. Computer Vision ICCV'03,Nice, France, Oct. 2003.

G. Gimel'farb: Symmetrical approach to the problem of automatic stereoscopic measurements in photogrammetry. Cybernetics, 1979, 15(20, 235-247; Consultants Bureau, N.Y.

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

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

R. Hartley and A. Zisserman: Multiple View Geometry in Computer Vision. *Cambridge university press, Cambrige, 2000*

B. Heigl, R. Koch, M. Pollefeys, J. Denzler, L. Van Gool: Plenoptic Modeling and Rendering from Image Sequences taken by a Hand–Held Camera. Proc. DAGM 99, Bonn, Germany, 1999

A. Heyden and K. Aström: Euclidian Reconstruction from constant intrinsic parameters. *Intl. Conf. PR*, 1996.

A. Heyden and K. Aström: Euclidian Reconstruction from image sequences with varying and unkwon focal length and principal point. *CVPR*, 1997.

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

R. Koch, B. Heigl, M. Pollefeys, L. Van Gool, H. Niemann: A Geometric Approach to Lightfield Calibration. Proceedings CAIP'99, Ljubljana, Slovenia, Sept. 1999.

R. Koch, M. Pollefeys, B. Heigl, L. Van Gool, H. Niemann: Calibration of Hand-held Camera Sequences for Plenoptic Modeling. Proc. of ICCV'99, Korfu, Greece, Sept. 1999.

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

R. Koch, J.-M. Frahm, J.-F. Evers-Senne, J. Woetzel: Plenoptic Modeling of 3D Scenes with a Sensor-augmented Multi-Camera Rig. Proceedings Tyrrhenian International Workshop on Digital Communications, Capri, Italy, Sept. 2002.

V. Kolmogorov, R. Zabih: Multi-camera Scene Reconstruction via Graph Cuts. Proceedings ECCV 2002, Kobenhagen, DK, May 2002.

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

S.J. Maybank and O. Faugeras: A therory of self-calibration of a moving camera. *Int. J. of Computer Vision*, 1992.

L. Naimark, E. Foxlin: Circular Data Matrix Fiducial System and Robust Image Processing for a Wearable Vision-Inertial Self-Tracker. ISMAR'02 IEEE Symposium on Mixed and Augmented Reality, Darmstadt, Germany, 2002.

M. Pollefeys, R. Koch, M. Vergauwen and L. Van Gool: Metric 3D Surface Reconstruction from Uncalibrated Image Sequences. In: 3D Structure from Multiple Images of Large Scale Environments. LNCS Series Vol. 1506, pp. 139-154. Springer-Verlag, 1998.

M. Pollefeys, R. Koch and L. Van Gool: Selfcalibration and metric reconstruction in spite of varying and unknown internal camera parameters. *ICCV*, 1998.

C. Rother and S. Carlsson: Linear multi view reconstruction and camera recovery using a reference plane. International Journal of Computer Vision IJCV 49(2/3):117-141.

G. Stein: Accurate internal camera calibration using rotation, with analysis of sources of error. *ICCV*, 1995.

D. Scharstein and R. Szeliski: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV 47(1/2/3):7-42, April-June 2002.

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

B. Triggs: Autocalibration and the Absolute Quadric. Proceedings Conference on Computer Vision and Pattern Recognition, pp. 609-614, Puerto Rico, USA, June 1997.

M. Okutomi and T. Kanade: A Locally Adaptive Window for Signal Processing. International Journal of Computer Vision, 7, 143-162, 1992.

R. Yang, M. Pollefeys: Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware. Proceedings IEEE Conf. Computer Vision and Pattern Recognition CVPR03, Madison, WISC., HSA, June 2003.