Increasing Realism and Supporting Content Planning for Dynamic Scenes in a Mixed Reality System incorporating a Time-of-Flight Camera

Ingo Schiller, Bogumil Bartczak, Falko Kellner and Reinhard Koch

Multimedia Information Processing
Computer Science Department
Christian-Albrechts-University of Kiel
Hermann-Rodewald-Str. 3
24118 Kiel, Germany
email: {ischiller,bartczak,fkellner,rk}@mip.informatik.uni-kiel.de
www: www.mip.informatik.uni-kiel.de

Abstract

For broadcasting purposes MIXED REALITY, the combination of real and virtual scene content, has become ubiquitous nowadays. Mixed Reality recording still requires expensive studio setups and is often limited to simple color keying. We present a system for Mixed Reality applications which uses depth keying and provides threedimensional mixing of real and artificial content. It features enhanced realism through automatic shadow computation which we consider a core issue to obtain realism and a convincing visual perception, besides the correct alignment of the two modalities and correct occlusion handling. Furthermore we present a possibility to support placement of virtual content in the scene.

Core feature of our system is the incorporation of a TIME-OF-FLIGHT (TOF)-camera device. This device delivers real-time depth images of the environment at a reasonable resolution and quality. This camera is used to build a static environment model and it also allows correct handling of mutual occlusions between real and virtual content, shadow computation and enhanced content planning.

The presented system is inexpensive, compact, mobile, flexible and provides convenient calibration procedures. Chroma-keying is replaced by depth-keying which is efficiently performed on the GRAPHICS PROCESSING UNIT (GPU) by the usage of an environment model and the current ToF-camera image. Automatic extraction and tracking of dynamic scene content is herewith performed and this information is used for planning and alignment of virtual content.

An additional sustainable feature is that depth maps of the mixed content are available in real-time, which makes the approach suitable for future 3DTV productions. The presented paper gives an overview of the whole system approach including camera calibration, environment model generation, real-time keying and mixing of virtual and real content, shadowing for virtual content and dynamic object tracking for content planning.

First presented at the 5th International Conference on Visual Media Production 2008, extended and revised for JVRB

Keywords: 3DTV, Time-of-Flight, 3D-Modeling, Depth keying, PMD, TOF, GPU, Mixed Reality, Shadow mapping

1 Introduction

Mixing of different sources of images is an important part of today's high quality film and TV productions. Many productions rely on virtual or fictive content and especially on the correct mixing of real and virtual scenes. To create video content with coexistence of real and virtual data and to hide this fact from the spectators, multiple challenges have to be met. Correct mixing of real and virtual content requires a consistent temporal and spatial alignment of the two components. Furthermore mutual occlusions have to be detected and handled correctly. Finally lighting conditions and optical effects like shadows have to be consistent between both modalities. An overview of the current techniques for mixed reality applications in TV and film productions is given in [Tho06].

In this paper we present an approach to real-time

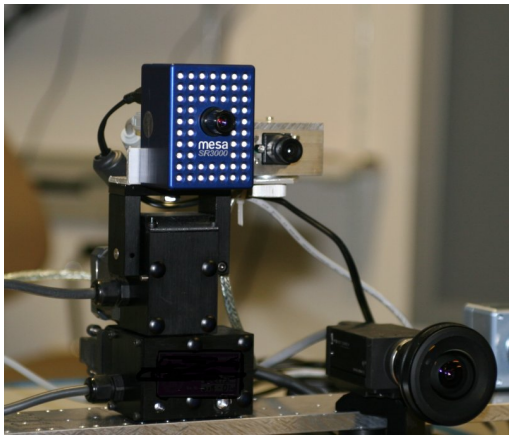

Figure 1: The used setup comprising of a ToF-camera (SwissRanger 3000) mounted together with a CCD firewire camera on a pan-tilt-unit and a fisheye camera mounted at the bottom right.

mixed reality applications based on a background environment model and the use of a Time-of-Flight-(ToF) camera. This work is based on an existing mixed reality system presented in [BSBK08]. We extend this system with mutual shadowing to increase the realism of such a system and with features to simplify the usage of such a system. The placement of virtual objects in the scene can be assisted if information about the movement of dynamic scene content is available. With the presented approach we show a straightforward way to obtain this necessary information. The presented approach is real-time capable, the spatial alignment of the virtual content can therefore be altered during live processing. Other aspects of the system, focussing on interaction and physics simulation, are further elaborated in [KSB+09].

Studios in which mixed content is produced are typically equipped with chroma- or difference-keying facilities and expensive camera tracking installations [TJNU97], or camera mounts with pan-tilt sensors are used. In [GKOY03] a system is presented which also uses a ToF-camera for "depth-keying" and the mixture of real and virtual data. However this approach is based on depth-keying through predefined depth intervals. It only targets the determination of foreground objects and is not suited for mixing of real and virtual content with mutual occlusions, nor does it provide any realistic shadowing.

Determining the translation and rotation of the used cameras is crucial for the alignment of the virtual content. A small inconsistency in the determination of the camera parameters can already destroy the targeted illusion. In [CTB+06, KBK07] systems are proposed, which track a camera pose based on a background environment model.
A precise threedimensional representation of the environment can be used in multiple ways if the precision is sufficiently high. The different methods of generating such a model and their advantages and disadvantages have already been discussed in [BSBK08].

Recently, active depth measuring devices based on the ToF[1] of an intensity modulated infrared light front have reached maturity (cf. [OLK+04]). The effective operation range of these ToF-cameras is only suited for indoor applications, but these devices are capable of capturing dynamic data over an extended field-of-view. (Note that solutions with different modulation frequencies and smaller field-of-view exist which target outdoor applications.) We therefore propose the application of such a ToF-camera in the presented context.

In this work we show how this device can be used for fast large scale background model construction. Furthermore we exploit its depth capturing capabilities during the (online) real data processing. Here the depth information is used in combination with the previously constructed background model to key foreground objects, virtual objects and the background on the fly. This way the need for chroma-keying facilities

[1]Companies providing such devices are represented at: www.pmdtec.com, www.mesa-imaging.ch and www.canesta.com

is made obsolete. Moreover a convenient and correct live handling of mutual occlusions of virtual and real content is provided, which is not possible by means of simple chroma-keying techniques.
This paper is structured as follows. Section 2 introduces the used ToF-camera. Section 3 gives an overview of the whole proposed approach. In the following sections the prerequisites for the live system usage are introduced. These are the calibration (4.1) of the camera rig, shown in figure 1 and the generation of the background model (4.2). Then the central system components are discussed in more detail. Section 5.1 describes how the camera pose is estimated during live processing. In section 5.2 the shadow rendering for the virtual objects is introduced. Section 5.3 describes the transfer of depth images from the ToF-camera to other images as well as the mixing and the segmentation of image content exploiting the processing power of modern graphics cards. Finally in section 5.4 it is shown how dynamic scene content segmentation can be used to support the alignment of the virtual objects. After presenting and discussing the achieved results in section 6 we conclude our work in section 7.

2 ToF-Camera

Before introducing the system architecture we will briefly introduce the Time-of-Flight-(ToF)camera we use in the presented system. It is a 2.5D camera which delivers dense depth images with up to 25 frames per second. The used camera, the SR3000 offers a resolution of 176×144 pixel, newer cameras such as the CamCube by PMDTec already offer a resolution of 204×204 pixel. The depth measurement is based on the Time-of-Flight principle. The camera actively illuminates the scene by sending out incoherently intensity modulated infrared light with a modulation frequency of 10-60 MHz. In the shown experiments a modulation frequency of 20 MHz was chosen which results in a non-ambiguity range of 7.5m. The light is reflected by 3D scene points and received by the image sensor of the ToF-camera. Different approaches for measuring the ToF with light exist (cf. [LSBS99],[OLK+04]). Depending on the Time-of-Flight a phase shift in the reflected signal is observable. The ToF-camera is able to extract this phase shift in every pixel. For measuring the reflected light the ToF-camera uses a semiconductor structure based on CCD- or CMOS-technology [XSH+98].

The phase difference is measured by cross correlation between the sent and received modulated signal by the camera's image sensor. Due to the used modulation frequency the non-ambiguity range of the ToF-camera is 1.5-7.5 meters. From calibration (see section 4.1) we know that the real depth accuracy of the ToF-camera is around ± 15mm or better. Since the resolution of the phase difference measurement is independent from distance, the achievable depth resolution is independent from scene depth. This is in contrast to stereo triangulation where depth accuracy is proportional to inverse depth.

3 System Overview

In this section we start with an overview of our real-time mixed reality system. Figure 2 shows an outline of our approach.

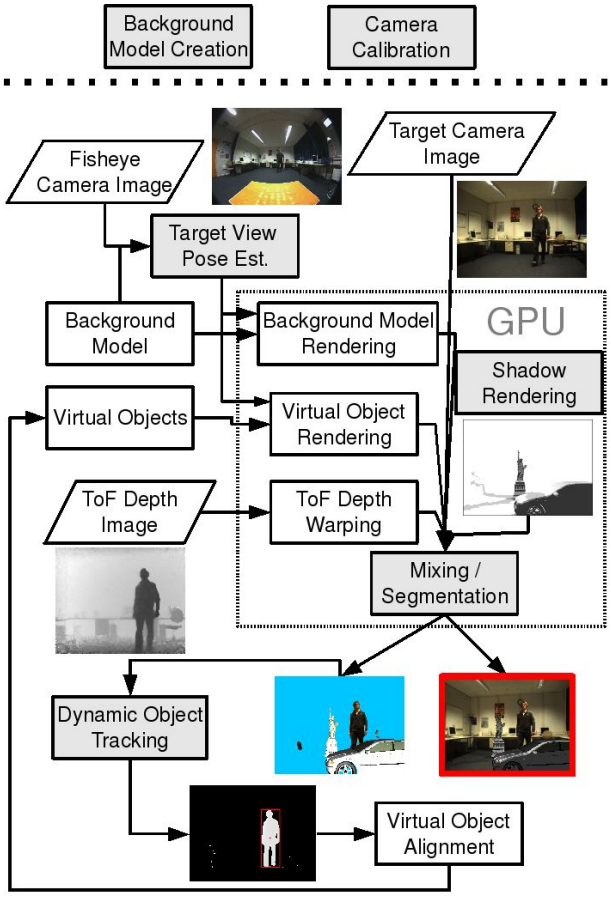

Figure 2: System components and interactions. The grey boxes are discussed in detail in the text.

As input we only use images taken by various cameras and no chroma-keying facility is used to segment the actor as can be seen in the input images shown in

figure 3. In our system we use three cameras which are shown in figure 1, a ToF-camera combined with a CCD-camera, both rigidly coupled and mounted on a Pan-Tilt-Unit, and a spherical CCD-camera. The ToF-camera delivers instantaneous depth images of the scene, as shown at the left-bottom of figure 2, but from a different viewpoint and with different intrinsic parameters than the target view camera. (The target view camera is shown in the top right of figure 2.)

Figure 3: Input images of fisheye-, target- and ToF-camera.

The working flow of our system starts with the calibration of the used camera rig. After that the background model is generated (see fig. 5). For that purpose the ToF-camera and the coupled CCD-camera are systematically moved by a Pan-Tilt-Unit (PTU) to capture the static environment. The retrieval of the environment's geometric and photometric properties is topic of section 4.2.

The background model is used in the following as replacement for chroma-keying technologies. As target camera and ToF-camera do not share the same center of projection and have different perspective camera parameters a warping of the depth measurements into the target view has to be performed. A convenient and very fast depth transfer is possible by means of a graphics processing unit (GPU), which is more closely explained in section 5.3. The GPU is also used for the pixel based depth comparison for segmentation and proper mixing between real and virtual content (see section 5.3). In the same instance shadowing of virtual objects is computed and added to the final mixing result (see section 5.2). In the mixing the depth image of the ToF-camera, the background model and the virtual objects are used. The consequent use of GPU programming allows to perform all computations of the online phase in real-time.

The view dependent data needed for these tasks is rendered from the virtual objects description and the background model. A correct rendering is only possible if the target view camera pose is known in the model's coordinate frame. To determine the current pose of the target camera we use the spherical CCD-camera as a reliable and precise pose sensor. A background model based tracking approach, elucidated in section 5.1, is then used to track the fisheye camera pose in the model's coordinate frame. From the calibration the relative orientation and translation between fisheye- and target camera is known.

Once the background model has been created, the planning and alignment of virtual content can be performed by means of 3D modeling tools. The trajectory of a moving object in the scene can be determined from the segmentation done on the GPU. This is explained in section 5.4. The received trajectory can now be used to assist in the placement of virtual objects in the scene. The subsequent section will discuss the central components of the system (indicated as grey boxes in fig. 2) in more detail.

4 System Prerequisites

4.1 Calibration

The internal and relative calibration of the used cameras is a crucial aspect in this application. The major challenge here arises from the fact that the system comprises of three different camera types. Those camera types are a classical perspective camera, a fisheye camera and a ToF-depth camera.

Well established techniques are available for calibrating classical perspective cameras including radial lens distortion. Those are usually based on multiple images of a calibration pattern. A simple and flexible method based on images of a planar checkerboard pattern has been presented for instance in [Zha99]. For the calibration of fisheye cameras a very similar ap-

proach based on the same planar checkerboard as calibration pattern is available. This has been presented in [SMS06]. Finally, the calibration of the ToF-camera is required. A calibration procedure based on a planar calibration pattern has been presented in [BK08]. ToF-cameras may contain systematic depth errors in the measurements. This can be compensated in different ways, for example with splines [LK06].

In order to obtain an optimal calibration for the entire system including the relative transformations between the different cameras, an overall adjustment is required. To achieve this goal, we use the analysis-by-synthesis approach presented in [SBK08]. All camera parameters including relative transformations, internal calibration for fisheye as well as for the perspective cameras and systematic depth measurement errors of the ToF-camera are used to render the known calibration pattern into each image (including the depth images of the ToF-camera). Comparing those rendered images with the actual images allows to compute a residual for each pixel. Furthermore, it is possible to compute partial derivatives of the residuals with respect to each of the parameters numerically by rerendering the images with slightly disturbed parameters. This enables the minimization of the residuals using standard adjustment techniques, such as the Levenberg-Marquardt algorithm, yielding a best linear unbiased estimate of all the parameters. Note, that the huge amount of rendering operations can be efficiently performed using graphics acceleration hardware, so that reasonable running times are achieved.

4.2 Background Model Generation

In this section we explain how the background model, which we use for production planning, visual camera tracking, depth-keying and the shadow computation, is generated utilizing the ToF-camera. Since the field of view of the ToF-camera is too small to capture a sufficiently large portion of the environment, the ToF-camera and the target camera were mounted onto a Pan-Tilt-Unit (PTU), as can be seen in figure 1. This way the scene can be scanned with the ToF-camera collecting the necessary data for geometry reconstruction. The colour camera simultaneously captures colour images, which deliver the photometric information needed in the visual camera tracking. The PTU is programmed to do a panoramic sweep covering a field of $270°$ in horizontal and $180°$ in vertical direction. Rather than directly generating a 3D representation of the scanned environment, two cylindrical panorama images are generated, one depth panorama and one colour panorama, as shown in figure 4.

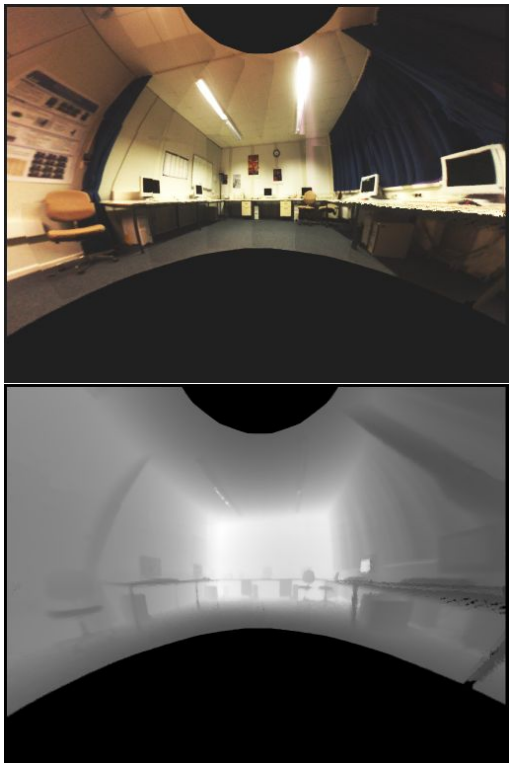

Figure 4: Cylindrical colour and depth image of the scanned panorama.

The creation of these panoramas relies on the previous calibration of intrinsic and extrinsic camera parameters of the ToF-camera, the target camera and the tool center point (TCP) of the PTU. The TCP is the transformation from the origin of the PTU to the optical point of the target camera, consisting of a rotation R and a translation T. The rotation R is known from the angles of the PTU and the translation T is to be estimated. This is done taking several images of a calibration pattern with the target camera with different (known) angles of the PTU. This allows to construct a system of equations which can be solved for T.

Thereafter the correct panorama content can be calculated by the known PTU orientations and the captured depth data in a forward mapping process. The neighborhood relations in the panoramas allow to improve the data by morphological operations and median or bilateral filtering. Moreover these data structures are well suited for a straightforward handling of overlaps in the captured images during the scanning process. The generation of such a panorama on the

basis of 22 captured images needs approximately 60 seconds.
In order to generate a model representation suitable for rendering, the depth- and colour panoramas are converted into one textured triangle-mesh following the proposition described in [BKWK07, p. 93]. The resulting textured triangle-mesh is shown in figure 5. Please observe the fill rate of 100% in the scanned areas, even in the sparsely textured parts, which would not have been obtainable using solely image based approaches. For a reduction of the high amount of data the triangles in the surface mesh can be processed by means of the propositions presented in [SZL92].

Figure 5: 3d-model of the scene corresponding to the panorama depicted in figure 4.

With the presented approach it is also possible to reconstruct the full 360° environment which is shown in figure 6.

In [HJS08] a comparable approach to model building using a ToF-camera is described. In contrast to our work registration has to be done between subsequent measurements using texture and depth information combined with inertial sensor measurements. This registration is not required in our approach as the PTU delivers high precision rotation data. The generated model could however be extended by moving the camera rig to a different location. Thereafter a combination of our approach and the method in [HJS08] could be applied to merge multiple models.
For a qualitative rating of the reconstruction accuracy the original room geometry has been measured with a laser distance measurement device and compared to the environment model. Measurement of the model has been carried out from wall to wall at 100 locations.

Figure 6: Full 360° reconstruction of the environment. Top Left: View of complete environment model. Others: Different views of the interior.

As can be seen in table 1, the mean of one hundred measurements differs 9-12mm from the real geometry and a maximum error of 28-32mm is observable. This is equivalent to an average error of 0.002 - 0.003% and a maximum error of 0.005 - 0.01%.

Reconstruction evaluation:			
Dimensions:(mm)	height	width	depth
room size (ground truth):	2987	5263	4328
model size (mean):	2996	5254	4340
model size (max error):	3019	5291	4302
Reconstruction Error:(mm)			
mean error:	9	9	12
maximum error:	32	28	26

Table 1: Top Lines: Room size measured in real scene, mean model size of 100 measurements taken in the reconstructed model and model size with largest deviation to real model. Bottom two lines: Mean and maximum error between reconstructed model and ground truth room.

The reconstruction of different and smaller objects and different aspects and methods are discussed in [SK09]. The reconstruction of a person, turning on a swivel chair in front of the ToF-camera and the space-time reconstruction of a dynamic scene is investigated.

5 System Components

5.1 Camera Pose Determination

The previously discussed calibration of the used camera rig (see section 4.1) delivers the intrinsic calibration of all cameras in the rig and their relative extrinsic relation. While this fixed calibration is sufficient for the purpose of transferring the depth measurements from the ToF-camera to the target camera, the rendering of the background model and the virtual objects requires the current pose of the target camera in the respective coordinate frame.

Note that the background model and the virtual objects have been aligned in an offline processing step. How this step can be assisted is described in section 5.4. The estimation of the camera pose has to be done with a sufficiently high frame rate and while dynamic objects move through the scenery. It was found that the analysis-by-synthesis camera tracking approach presented in [KBK07] has the necessary properties to meet the demands of this task. We use an approach very similar to this proposed approach, sketched in figure 7.

The pose estimation starts with the registration of the current fisheye image to the generated background model and the virtual objects' coordinate frame. Assuming that we are close to the position from which the model was generated, we render the background model with the intrinsic parameters of the fisheye camera and the extrinsic parameters from which the model was generated. We then detect the gradient orientation based SIFT-features [Low04] in the rendered image and generate 3D points for these features. The SIFT-features are matched against features extracted in the current fisheye camera's image and a camera pose is estimated on these 2D/3D correspondences. After this registration the background model is aligned to the current image. In the following we use the proposed (cf. [KBK07]) analysis-by-synthesis tracking approach. Rather than globally optimizing the intensity difference between the current fisheye image and the rendered model image, it is based on tracking interest points ([ST94]) from the model's image to the current fisheye image. From the depth information taken at the positions of the 2D intensity features 3D points are generated.

Based on the established 2D/3D correspondences the current pose can be estimated for each image. Before a feature contributes to the estimation its validity is checked by a robust photometric measure in order to detect features which are occluded by a dynamic object. The fisheye camera's extended FoV always provides sufficient visible features for reliable tracking, even if large parts of the used background model are occluded in the targeted camera view. Using the background model as reference we are free of error accumulation (drift).

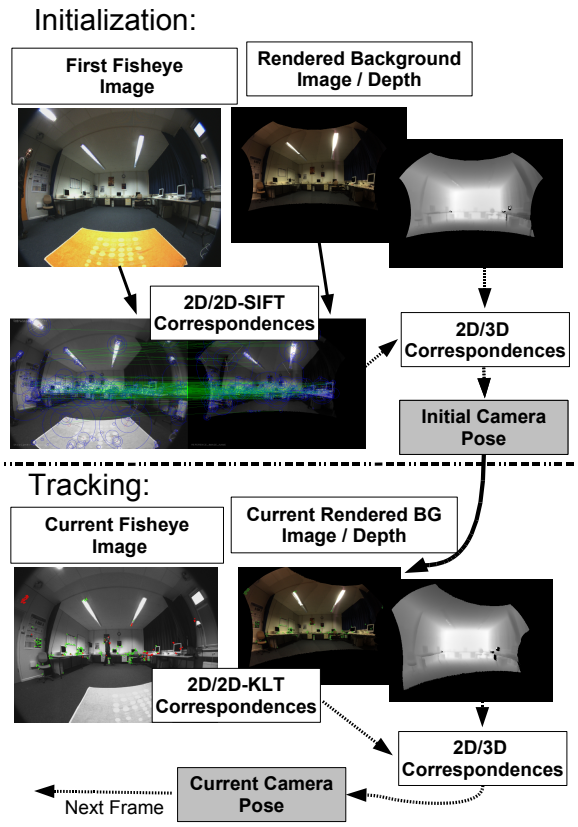

Figure 7: Overview of the camera pose tracking algorithm. In the initialization SIFT-features are computed on the current fisheye image and on the rendered background image. From the rendered depth 2D/3D correspondences are computed and used to estimate the current camera pose. The tracking uses the last available pose to render the model again. KLT-features are detected on the current and on the rendered image and correspondences established. From the depth image 2D/3D correspondences are available for pose estimation.

Utilizing the previously determined fixed relation between target camera and fisheye camera within the used rig, the determined pose can be mapped to the representation needed for rendering.

5.2 Shadow Rendering

Crucial for a plausible augmentation of video data with virtual data is the timely correct and stable placement of the synthetic content in the images. This is addressed by the previously described camera pose tracking. Another issue that significantly rises the quality of the augmentation is the shadow casting of the virtual objects. Take the top images in the right column of figure 8 as an example. The topmost images are showing an augmentation without shadow calculation for the virtual objects. Assuming that the virtual objects were correctly aligned to the floor, the camera's movement should convey this fact to the viewer. However legitimate doubt remains whether the virtual elements are really fixed to the ground or if they are floating slightly over ground. This is resolved when shadows are added, like shown in the lower images of figure 8. In order to add the shadows cast by virtual objects to the real images, so called light maps are calculated for each video frame. These maps basically encode how much light is reaching the part of a scene seen by a particular pixel when virtual content is present. Each pixel in the light map contains a factor $0 \leq s \leq 1$, which is used to scale the RGB colour values in the respective augmented image. A scale factor of 1 corresponds to no shadowing, 0 renders a pixel absolutely black and values in between model partial shadowing. The light maps are generated using the well-known shadow mapping technique [Wil78]. Therefore a depth map is rendered for each light source, containing all objects that should throw shadows. Afterwards the background model and all virtual objects are rendered from the target camera's point of view, 'coloring' the scene with the calculated lights' depth maps using projective texturing. This way for each pixel in the target image the distance values R encoded in the light sources' depth maps can be compared to the distances D between a light source and the 3D point corresponding to the pixel. As the light's depth map provides us with the distance between the light source and the first intersection of the light ray with the scene geometry, we can decide whether the pixel is in shadow ($R < D$) or receives light from the light source ($D = R$). Evaluating all light sources and combining them with an ambient light offset yields the target view dependent light map used for shadow generation as shown in figure 8.

This algorithm automatically adapts to different scene geometries, which allows to take full advantage of the background model's geometric information for shadow rendering. This is demonstrated in the bottom images in figure 8. Observe how consistent results are achieved by not only casting shadows on the floor and the side walls but also on tables, taking the given background geometry into account. The background model can moreover aid in defining the appropriate positions and orientations of light sources in the scene, because the real light sources are visible in its texture. The results presented in the middle row of figure 8 were generated using four point-lights, that were positioned on the real light sources through manual interaction with the background model (see figure 5). This is of course not sufficient to simulate the reality but already increases the augmentation's perceived quality. However, spending more resources for rendering more light sources in conjunction with light map smoothing will already increase the realism without much alteration of the proposed processing scheme.

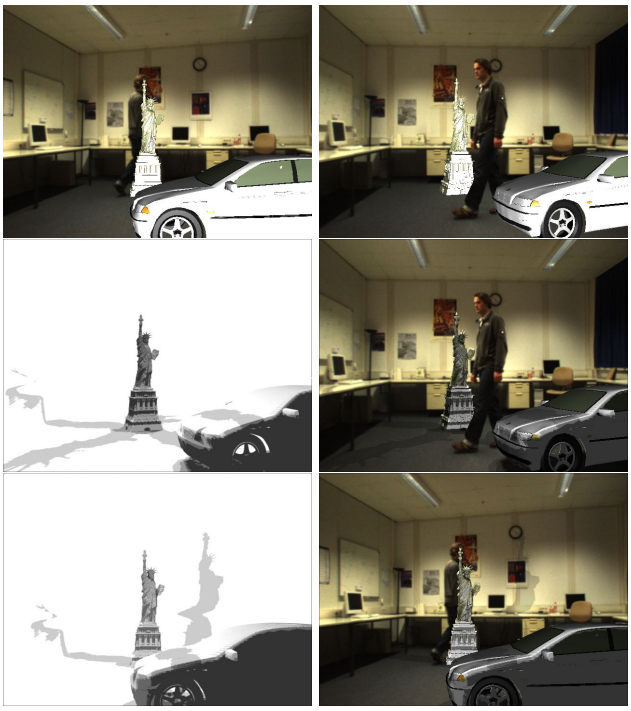

Figure 8: Top row: Mixed images without shadow computation, Middle row: Light map and image with shadows computed from light sources at the ceiling, taken from the background model. Bottom row: Light map and image with altered light source positions.

5.3 Depth-Keying and Mixing

In this stage of the processing chain the GPU is programmed to deliver the mixed output images as well

as segmentations of the background and dynamic objects in the current frame. Figure 9 sketches how the input images shown on the left are combined to deliver three output images. The input data is either directly captured by the camera rig, or generated using OpenGL in multiple render-passes. One important aspect is the warping of the depth information captured by the ToF-camera into the target camera view by applying the technique described in [BSBK08]. This is necessary because both cameras have different extrinsic and intrinsic orientations. The warped ToF depth map is on the one hand used to segment the dynamic 'real' foreground objects and on the other hand to correctly handle mutual occlusions between virtual objects and the real scenery. The latter is sketched at the top part of figure 9. Here the depth maps of the ToF camera and the rendering of virtual objects are compared. While the mixed depth map is generated by returning the depth value closest to the target camera, the mixed colour image is generated by selecting the appropriate entries from the corresponding colour images. Colours from the current target camera frame are selected if the ToF-camera measurement is closest to the target camera, otherwise the virtual objects' colour is chosen.

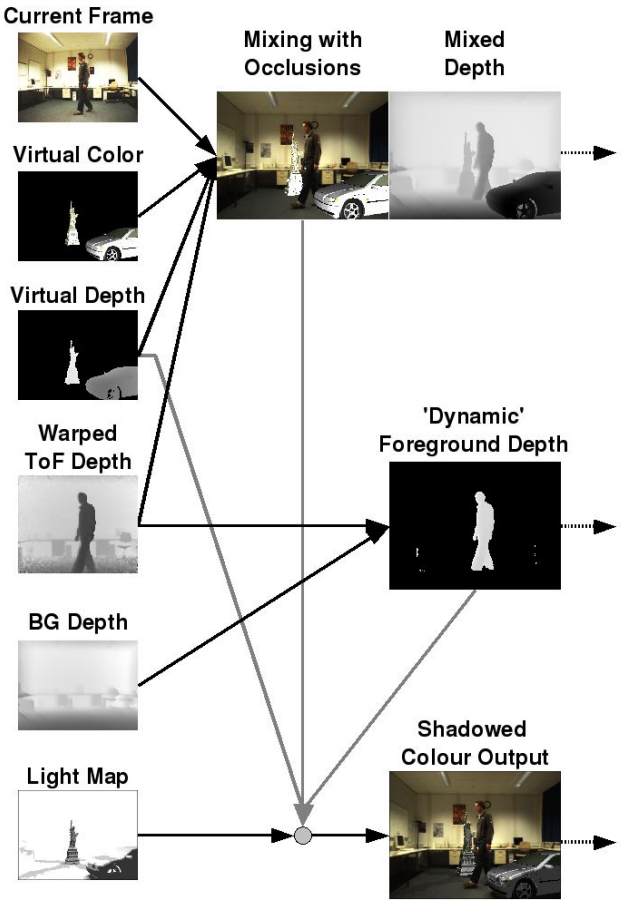

Figure 9: Overview of the mixing and shadowing done on the GPU. On the left hand side all the input images are displayed. Based on the different depth images mutual occlusions can be handled in the augmentation. Moreover foreground segmented depth images and mixed depth images are delivered. The scaling of the augmented image via the light map yields the final colour image output.

The segmentation of foreground and background is depicted in the middle of figure 9. Foreground classification on a per pixel basis is achieved by comparing the rendered depth map of the background model to the current warped ToF depth map. Wherever the difference between the ToF depth map and the rendered background depth map is smaller than a certain threshold, static background is assumed (black colour), otherwise the depth of the ToF measurement is returned. This threshold is dependent on scene size, clipping planes and camera noise. The low resolution of the ToF-camera constitutes the main limitation of the system concerning accuracy. One ToF-camera pixel is approx. 7×7 pixel in the CCD camera if a 1280×1024 pixel CCD-camera is used or even worse for CCD cameras with higher resolution. This leads to visible errors in the segmentation which will be subject to further research.

This segmentation is making a 3D-tracking of the dynamic foreground objects possible as described in section 5.4. Furthermore it is used during the application of shadows to the mixed image, like shown in the bottom of figure 9. Here the segmented depth map is again compared to the depth map of the virtual objects. All pixels of the current image, which are not occluded by virtual objects, are masked from being changed through shadow application. This is done because in these image parts the foreground objects are occluding the geometry considered during light map calculation. The shadows produced by the noisy foreground geometry distort the visual perception of the shadows and are therefore neglected. In these cases the augmentation is less disturbed if only the shadow information already contained in the captured target camera image is used. All other pixels have corresponding shadow information in the light map and are therefore processed as described in section 5.2.

Observe that the generation of all three output images, namely the mixed depth map, the segmented

depth map as well as the shadowed colour image only require simple comparisons, plain texture fetches and assignments for each pixel individually. This is ideal for GPU processing. Moreover the multiple rendering target capabilities of modern graphics cards allow to deliver all three output images in a single rendering pass, making this processing step very fast.

5.4 Tracking Dynamic Foreground Objects

For the planning and virtual content alignment it is helpful to analyse the dynamic objects' movements in the scene. For this task a segmentation of the moving objects and the static background model is needed. In this system the segmentation is inherently done by depth-keying (cf. section 5.3) delivering images containing only scene parts not included in the background model. In the presented example (figure 10) a moving person appears in the segmentation results, which is done for the z-buffer depth map. To detect individual moving objects in the images a simple and fast clustering algorithm, BFS (Breadth-First Search) is used. It is known as a graph search algorithm which can easily be applied to images. Dependent on a given threshold the segmentation result is analyzed. Starting with a single pixel which is above the threshold, all neighboring pixels which are above this threshold are added to the current object. For the sake of robustness to small segmentation errors a minimum desired object size is defined. For each detected object a center-of-mass is calculated by averaging its pixel coordinates. The result of this tracking can be seen in figure 10 where only one object is present. Projecting the detected pixel coordinate with the depth measurement from the ToF-camera to a 3D point results in the 3D point of the center-of-mass of the object. The segmentation sets all background pixel to 0, so for the BFS a threshold >0 is reasonable.

Projecting the detected 3D point of the center-of-mass to the ground plane of the background model directly yields the trajectory of the moving person on the floor, as shown in figure 11 at the top. This information can be used during live processing to place the models in the scene. It can additionally be used in an offline step to plan the placement of the virtual models. After this placement step the processing chain can be repeated with corrected model alignment. In the exemplary results shown in figure 11 the person entered the model from the left and walked round the room two and a half times, turned around and walked in the

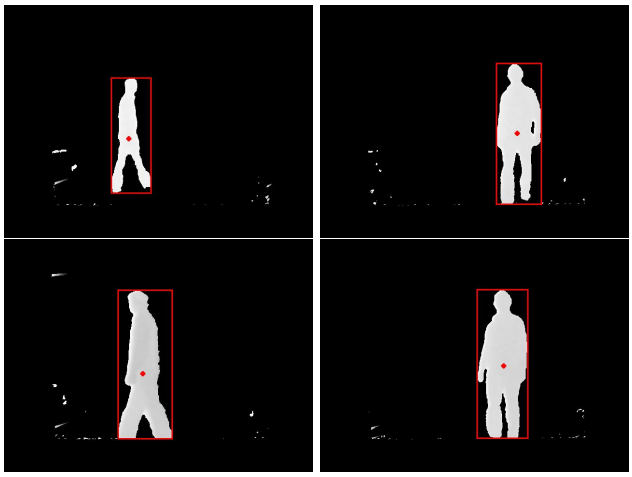

Figure 10: Tracking of dynamic content. The moving person is detected and marked with a rectangle, the center-of-mass is marked with a dot.

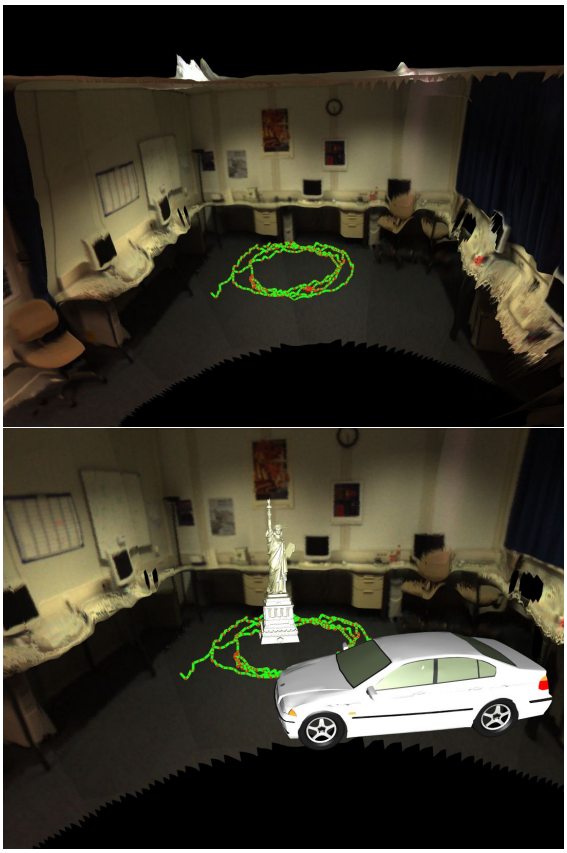

Figure 11: Top: Trajectory of the moving person projected to the ground plane. Points show the positions, connections the path of the person. Bottom: Background model with path and aligned virtual models.

opposite direction. Note that this tracking of moving scene content can also be used to simulate interaction, e.g. collisions between real and virtual content.

6 Results and Discussion

Evaluating the system with synthetic data produces reasonable results. Occlusion detection and mixing of perfect models and depth measurements is not a challenge. The actual challenges are the handling of noise during the segmentation and mixing, the alignment of the generated model to the current image frame and the generation of a real-time system which conveys adequate results. We will therefore discuss the whole system qualitatively on real data. The data used in the shown case consists of 420 images taken simultaneously with each of the three cameras, the ToF-, the perspective CCD- and the fisheye-camera. Some exemplary input images of the cameras are shown in figure 3. The top row shows two images taken with the fisheye camera which is used as a pose sensor. The second row shows images taken with the perspective camera, also called the target camera in previous sections and the third row shows two depth images taken with the ToF-camera.
The background model generation is realized by the combination of the ToF-camera, the perspective (target)-camera and a Pan-Tilt-Unit. A sweep is performed with the two cameras and a cylindrical depth and texture panorama is generated capturing a FoV of $270^\circ \times 180^\circ$. From these panoramic images a 3D model of the environment is created by meshing the depth panorama and applying projective texturing. Accuracy is crucial in this step and is bound by the accuracy of the ToF-camera. We found that the model fits the real environment with a mean error of approximately 12mm.

As our scenarios include movement of the camera rig and dynamic scene content the reliable determination of the camera pose is mandatory for consistent alignment of virtual content in the scene. The proposed camera pose estimation approach using a fisheye camera evaluated as being well chosen for this purpose. The fisheye camera is, due to its large FoV, robust to dynamic content. The analysis-by-synthesis tracking approach incorporating the background model prevents drift, of which standard frame-to-frame approaches suffer from. The effect of the camera movement and correct determination of it is visible in figure 13. Observe the correctly aligned car in the front while the camera is moving from right to left and back again.
Shadowing of virtually introduced objects is mandatory for a believable impression of the composed content. We proposed to use a simple but effective technique using the well-known shadow mapping algorithm. The light source positions can be determined with help from the background model, to achieve better results additional light sources can easily be introduced. To review the effect of shadowing refer to figure 8. In the top row the scene is shown without shadowing of the virtual objects. The other two rows show the computed light maps and the composed images with shadows. The shadowing between the virtual objects and between the virtual objects and the background, for example the shadow of the statue on the floor, the walls, the tables and on the car, is very important for a realistic perception. This consistent shadow casting on the floor, on the walls and on the tables is only possible due to the usage of the background model. However since the full 3D shape of the dynamic foreground object is unknown to the system, it is unable to deliver shadowing information between the dynamic objects and the background geometry.

The warped depth maps are mixed on the GPU as described in section 5.3 with the background model and the virtual objects which are to be rendered in the final augmented scene. The result of the mixing can be seen in figure 13. Due to the integration time of the ToF-camera and motion of scene parts, errors in segmentation and mixing infrequently occur. Though this is happening infrequently the effect is visible in figure 12 in the right image at the silhouette of the person.

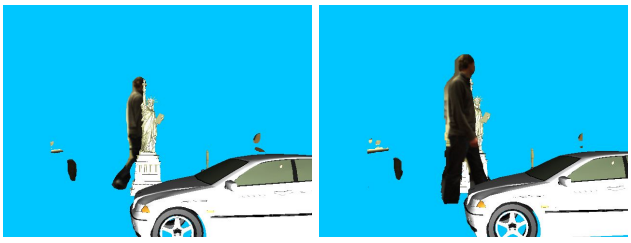

Figure 12: Mixing of real and virtually added data. In order to make the consequences of motion blur in the depth images more visible the background has been removed by depth-keying.

From the segmentation of the dynamic content in the warped ToF-measurements, as shown in figure 10, the independently moving objects in the scene are de-

tected applying a clustering algorithm. From these clusters the center-of-mass is detected and projected to the ground plane of the background model. This is done in real-time and as a result, the paths of the moving objects can be exploited to place the models in the scene, avoiding unwanted overlaps.

Figure 13: The final mixing result with the augmentation of virtual objects in the target camera, occlusion handling and shadow rendering.

In figure 13 the final result of the system is shown. The models are rendered in the target camera while mutual occlusions are taken into account and shadowing between virtual objects and the background is added. Note the high quality of the occlusion detection, for example the segmentation between the legs in the bottom left image, despite the low resolution of the ToF-camera. Since the camera pose is estimated by the system the camera can move in the scene without destroying the illusion of the virtual objects really being present in the room. Observe how more and more of the car becomes visible when the camera moves to the left. Figure 13 also shows the corresponding mixed depth image for every augmented frame. Both data sets together can be used to generate the required information needed by todays auto-stereoscopic 3D displays, making the presented approach suitable for future 3DTV productions.

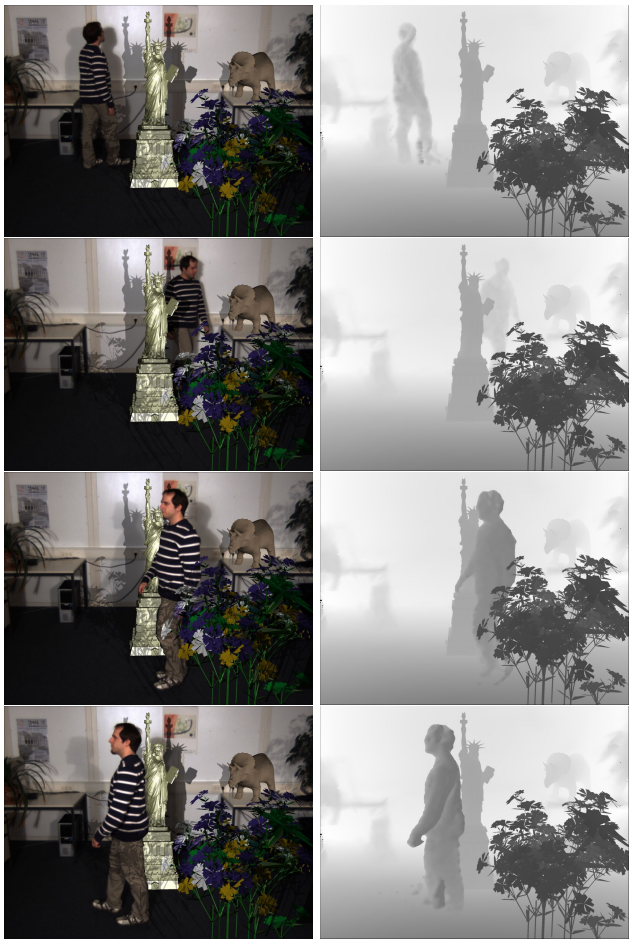

Figure 14: Final mixing result with shadow mapping. In this example two light sources are placed left and right of the camera. The real light sources have been simulated in the shadow mapping and shadow casting between virtual content and real dynamic content is visible. (Note the shadow on the legs of the person.)

Figure 14 shows another example for the mixing results. In this example cameras with a higher resolution have been used and two lightsources have been placed

on either side of the camera rig. This setting of light-sources is also used in the shadow mapping and both shadows, the real ones and the virtual ones, are visible in the scene. The person casts shadows on the walls, as well as the virtual objects. Additionally, the virtual elements cast shadows on the dynamically captured geometry of the person.

Figure 15: Closeup on a final composited image. Note the mutual shadowing between virtual content and background and between virtual object and dynamic real content. (e.g. the shadows of the Statue of Liberty on the person.)

7 Conclusions and Future Work

In this work two significant extensions to the mixed reality system presented in [BSBK08] were addressed. Central to this system is the usage of a ToF-camera which is used to generate an environment model and provides the necessary information for occlusion detection and segmentation by depth. With the combination of three different cameras, a fisheye camera as pose sensor, a ToF-camera as depth sensor and a perspective camera as target camera for the mixing result a complete system is presented without the need of a special studio surrounding with for example chroma-keying facilities. Moreover solutions to several important challenges are provided. These are the real-time camera pose determination in presence of dynamic scene content, the correct handling of mutual occlusions and the computation of depth maps for 3DTV broadcasting. However the system lacks realism due to missing shadows of virtual objects. Therefore the first contribution of this work is the integration of shadow casting of the virtually introduced objects in an intuitive way. The second contribution is the capability to automatically detect the trajectories of moving objects and the application of these trajectories for interactive and post-productive alignment of virtual objects in the scene.

The proposed shadow calculation approach is quite effective, although not perfect. The mutual shadowing between dynamic and virtual objects suffers from the incompleteness of the dynamic scene content due to the fact that the scene is just observed by a single ToF-camera, which is only delivering 2.5D information from a single viewpoint. Therefore it is only possible to calculate shadows which cast on the visible surfaces of the dynamic objects as shown in figure 14. Future extension of the system's shadow rendering will deal with this aspect. One approach would be to add further (static) ToF-cameras into the scene to supplement the 2.5D information of an object ([GFP08]). The usage of multiple ToF-cameras is possible if the cameras are operated using different modulation frequencies. This way not only the shadow computation could be improved but also the segmentation and object tracking.

It would be more sophisticated to combine the system with a motion capture approach like proposed by [RKS$^+$05], where a virtual model or proxy of a dynamic and deformable object is tracked. This approach would require substantial a priori knowledge of the observed objects. Nevertheless it could help in detecting and reducing the (spurious) motion blur effects, affecting segmentation and correct occlusion handling at discontinuities. At this point a less complex attempt might also consider silhouette methods suited for cluttered backgrounds in order to integrate colour information into the segmentation ([GHK06]).

Acknowledgments

This work was partially supported by the German Research Foundation (DFG), KO-2044/3-2 and the Project 3D4YOU, Grant 215075 of the ICT (Information and Communication Technologies) Work Programme of the EU's 7^{th} Framework program.

References

[BK08] C. Beder and R. Koch, *Calibration of focal length and 3d pose based on the reflectance and depth image of a planar object*, International Journal of Intelligent Systems Technologies and Applications (IJISTA), Issue on Dynamic 3D Imaging **vol. 5** (2008), no. 3-4, 285 – 294, ISSN 1740-8865.

[BKWK07] B. Bartczak, K. Koeser, F. Woelk, and R. Koch, *Extraction of 3D Freeform Surfaces as Visual Landmarks for Real-Time Tracking*, Journal of Real Time Image Processing **vol. 2** (2007), no. 2-3, 81–101, ISSN 1861-8200.

[BSBK08] Bogumil Bartczak, Ingo Schiller, Christian Beder, and Reinhard Koch, *Integration of a Time-of-Flight Camera into a Mixed Reality System for Handling Dynamic Scenes, Moving Viewpoints and Occlusions in Real-Time*, Proceedings of the 3DPVT Workshop (Atlanta, GA, USA), June 2008.

[CTB+06] J. Chandaria, G. Thomas, B. Bartczak, R. Koeser, K. Koch, M. Becker, G. Bleser, D. Stricker, C. Wohlleber, M. Felsberg, J. Hol, T. Schoen, J. Skoglund, P. Slycke, and S. Smeitz, *Real-time Camera Tracking in the MATRIS Project*, Proceedings of International Broadcasting Convention (IBC) (Amsterdam, The Netherlands), 2006, pp. 321–328.

[GFP08] L. Guan, J.-S. Franco, and M. Pollefeys, *3D Object Reconstruction with Heterogeneous Sensor Data*, 4th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT) (Atlanta, GA, USA), June 2008.

[GHK06] D. Grest, D. Herzog, and R. Koch, *Monocular Body Pose Estimation by Color Histograms and Point Tracking*, Proceedings of DAGM Symposium, Pattern Recognition (Berlin, Germany), LNCS, vol. 4174, Springer, Sept. 2006, pp. 576–586.

[GKOY03] R. Gvili, A. Kaplan, E. Ofek, and G. Yahav, *Depth keying*, Stereoscopic Displays and Virtual Reality Systems X (Andrew J. Woods, Mark T. Bolas, John O. Merritt, and Stephen A. Benton, eds.), vol. 5006, SPIE, 2003, ISBN 0-8194-4806-0, pp. 564–574.

[HJS08] B. Huhle, P. Jenke, and W. Strasser, *On-the-Fly Scene Acquisition with a Handy Multisensor-System*, International Journal of Intelligent Systems Technologies and Applications (IJISTA), Issue on Dynamic 3D Imaging **vol. 5** (2008), no. 3-4, 255 – 263, ISSN 1740-8865.

[KBK07] K. Koeser, B. Bartczak, and R. Koch, *Robust GPU-Assisted Camera Tracking using Free-form Surface Models*, Journal of Real Time Image Processing **vol. 2** (2007), no. 2-3, 133–147, ISSN 1861-8200.

[KSB+09] Reinhard Koch, Ingo Schiller, Bogumil Bartczak, Falko Kellner, and Kevin Koeser, *MixIn3D: 3D Mixed Reality with ToF-Camera*, Dynamic 3D Imaging Workshop at DAGM 2009, Dyn3D, LNCS 5742 (Jena, Germany), September 2009, pp. 126–141.

[LK06] M. Lindner and A. Kolb, *Lateral and Depth Calibration of PMD-Distance Sensors*, Lecture Notes in Computer Science Vol. 4292, Advances in Visual Computing, International Symposium on Visual Computing (ISVC06), vol. 2, Springer, 2006, ISBN 978-3-540-48626-8, pp. 524–533.

[Low04] D. G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision (IJCV) **vol. 60** (2004), no. 2, 91–110, ISSN 0920-5691.

[LSBS99] R. Lange, P. Seitz, A. Biber, and R. Schwarte, *Time-of-flight range imaging with a custom solid state image sensor*, Proc. SPIE Vol. 3823, p. 180-191, Laser Metrology and Inspection (H. J. Tiziani and P. K. Rastogi, eds.), 1999.

[OLK+04] T. Oggier, M. Lehmann, R. Kaufmannn, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger, and N. Blanc, *An all-solid-state optical range camera for 3D-real-time imaging with sub-centimeter depth-resolution (SwissRanger)*, Optical Design and Engineering, Proc. SPIE Vol. 5249 (Laurent Mazuray, Philip J. Rogers, and Rolf Wartmann, eds.), 2004, pp. 534–545.

[RKS+05] B. Rosenhahn, U. Kersting, D. Smith, J. Gurney, T. Brox, and R. Klette, *A System for Marker-Less Human Motion Estimation*, Proceedings of DAGM Symposium, Pattern Recognition (Vienna, Austria), LNCS, vol. 3663, Springer, Sept. 2005, ISSN 0302-9743, pp. 230–237.

[SBK08] Ingo Schiller, Christian Beder, and Reinhard Koch, *Calibration of a PMD camera using a planar calibration object together with a multi-camera setup*, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (Beijing, China), vol. Vol. XXXVII. Part B3a, 2008, XXI. ISPRS Congress, pp. 297–302.

[SK09] Ingo Schiller and Reinhard Koch, *Datastructures for Capturing Dynamic Scenes with a Time-of-Flight Camera*, Dynamic 3D Imaging Workshop at DAGM 2009, Dyn3D, LNCS 5742 (Jena, Germany), September 2009, pp. 42–57.

[SMS06] D. Scaramuzza, A. Martinelli, and R. Siegwart, *A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion*, Proceedings of International Conference of Vision Systems (ICVS), IEEE, January 2006, p. 45, ISBN 0-7695-2506-7.

[ST94] J. Shi and C. Tomasi, *Good features to track*, Conference on Computer Vision and Pattern Recognition (CVPR) (Seattle), IEEE, June 1994, ISB 0-8186-5825-8, pp. 593–600.

[SZL92] William J. Schroeder, Jonathan A. Zarge, and William E. Lorensen, *Decimation of triangle meshes*, SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques (New York, NY, USA), ACM, 1992, ISBN 0-89791-479-1, pp. 65–70.

[Tho06] G. A. Thomas, *Mixed reality techniques for TV and their application for on-set and pre-visualization in film production*, International Workshop on Mixed Reality Technology for Flimmaking, 2006.

[TJNU97] G.A. Thomas, J. Jin, T. Niblett, and C. Urquhart, *A versatile camera position measurement system for virtual reality TV production*, International Broadcasting Convention **vol. 2** (1997), no. 447, 284–289, ISSN 0537-9989.

[Wil78] L. Williams, *Casting curved shadows on curved surfaces*, SIGGRAPH '78: Proceedings of the 5th annual conference on Computer graphics and interactive techniques **vol. 12** (1978), no. 3, 270–274, ISSN 0097-8930.

[XSH+98] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck., *Smart pixel - photonic mixer device (PMD)*, M2VIP - International Conference on Mechatronics and Machine Vision in Practice, 1998, pp. 259 – 264.

[Zha99] Z. Zhang, *Flexible Camera Calibration by Viewing a Plane from Unknown Orientations*, Proceedings of the International Conference on Computer Vision (ICCV) (Corfu, Greece), vol. I, 1999, pp. 666–673, ISBN 0-7695-0164-8.

Citation

Ingo Schiller, Bogumil Bartczak, Falko Kellner and Reinhard Koch, *Increasing Realism and Supporting Content Planning for Dynamic Scenes in a Mixed Reality System incorporating a Time-of-Flight Camera*, Journal of Virtual Reality and Broad-, casting, 7(2010), no. 4, August 2010, urn:nbn:de:0009-6-25786, ISSN 1860-2037.