Heterogeneous binocular camera-tracking in a Virtual Studio

Matthias Flasko*, Patrick Pogscheba*, Jens Herder* and Wolfgang Vonolfen†

*FH Düsseldorf, University of Applied Sciences
Department of Media
40474 Duesseldorf, Germany
Tel.: +49 (0)211 / 43 51 - 810
Fax: +49 (0)211 / 43 51 - 803
E-Mail: {matthiaschristian.flasko | patrick.pogscheba | herder}
@fh-duesseldorf.de

†Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
53757 Sankt Augustin, Germany
Tel.: (+49) (0)2241 14-3425

E-Mail: wolfgang.vonolfen@iais.fraunhofer.de

Abstract: This paper presents a tracking of parts of a human body in a virtual TV studio environment. The tracking is based on a depth camera and a HD studio camera and aims at a realistic interaction between the actor and the computer generated environment.
Stereo calibration methods are used to match corresponding pixels of both cameras (HD color and depth image). Hence the images were rectified and column aligned. The disparity is used to correct the depth image pixel by pixel. This image registration results in row and column aligned images where ghost regions are in the depth image resulting from occlusion. Both images are used to generate foreground masks with chroma and depth keying. The color image is taken for skin color segmentation to determine and distinguish the actor's hands and face. In the depth image the flesh colored regions were used to determine their spatial position. The extracted positions were augmented by virtual objects. The scene is rendered correctly with virtual camera parameters which were calculated from the camera calibration parameters. Generated computer graphics with alpha value are combined with the HD color images. This compositing shows interaction with augmented objects for verification. The additional depth information results in changing the size of objects next to the hands when the actor moves around.

Keywords: camera calibration, stereo calibration, image registration, image processing, virtual studio, interaction

1 Introduction

It is state of the art in today's virtual studios to use a uniquely colored backdrop such as a blue or green box which defines the background recorded by a camera, being keyed out and replaced by computer generated images. Computer Generated Imagery (CGI) is generated on powerful graphic workstations and combined with correct spatial and perspective relations with the real camera shots. However, a direct interaction between humans and the

virtual content still remains difficult.
The use of optical markers works in position and orientation determination quite well. but also implies a complex image correction to eliminate their optical appearance. Proxy objects could be a better solution. On the other hand, they need to be built up for every new virtual set and thus reduce the advantage of virtual studios to provide a fast and flexible production.
A common method to realize interaction is the analysis of the optical flow in the image supported by camera tracking information to determine the raw direction of object movements in the scene. This, however, creates a noticeable delay (especially disturbing when fed back to the actor) and can only be calculated up to a certain reliability and accuracy in depth.
In any case, the most challenging task is to retrieve a depth information of sufficient resolution and accuracy from two-dimensional images. There are several ways to gain spatial information at TV frame rate like stereo vision or the use of depth cameras (depth from structured light, from pulse modulated light and so on). An actual drawback of many depth cameras are a low resolution and noisy data.
This paper describes an improvement for the interaction with virtual content in virtual studios. We use an additional depth camera to track the actor and implement image processing algorithms working on HD (high-definition) camera pictures and depth images. We describe a method in image registration for a binocular camera setup which requires at least one distance measuring camera. As a result we get two identical images which differ in ghost regions.

2 Related Work

There are several approaches to combine depth data with color information. They mainly use geometrical or image processing methods. Another possibility is the use of optical systems. Lindner et al. [LKH07] use camera calibration to determine the intrinsic and extrinsic camera parameters of a Photonic Mixer Device (PMD) and a color camera. The depth-data is transformed by the external parameters and a perspective projection with the RGB camera intrinsics. The PMD-data is extended by interpolating sub-pixels in order to receive the same resolution. The fusion is done by projective texture maps with an affine transformation of the color image. Finally the occluded parts were eliminated by hidden surface removal. A similar method for image registration is used in [PHW+07]. External camera parameters were determined through calibration and the 3D-points were projected into the 2D-color image. Another approach is to gain super-resolution of depth images from an image sequence [STDT08] while the camera is moving. For comparison, they implemented a joint bilateral upsampling between a high-res color camera and a depth camera. The data fusion was done by a homographic warp of the color image. In this case shadowed regions were wrongly reconstructed. A similar method is an image registration by an enhanced homographic warp which is dependant on the average range to the object. By use of a bilateral filter and a cost function the best fitting is calculated [YYDN07].
A combination of a low-resolution depth data and high-resolution color images with Markov

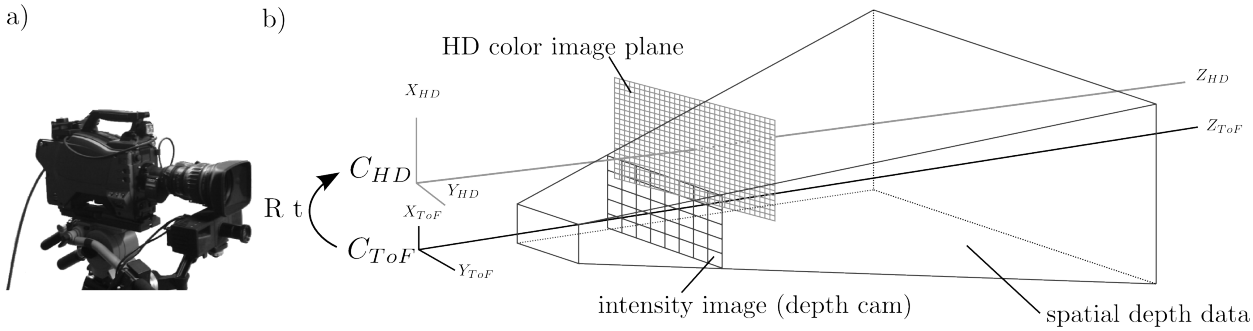

Figure 1: camera model with ToF and HD-TV camera

Random Fields (MRF) are used by Diebel and Thrun [DT06]. The MRF uses discontinuities in color and depth images to co-align the images. They generate high-resolution, low-noise range images with textures.
The Axi-Vision camera was designed at NHK and uses a sensor for color images and one for modulated Infrared-light (IR) pulse for range measurements. The image fusion is done by optics which uses at first a dichromatic mirror [KIA+00] and later a dichromatic prism [KKKI02, KIN+04] to split the light-rays onto a color and an IR depth measuring sensor. The camera produces colored range images and is used in a virtual studio to generate extended masks and combine the images with computer generated content.

3 Setup and Implementation

Our setup uses two cameras. One camera (PMD CamCube 2.0) measures the depth with a pulse modulated distance method and the other camera is a high-resolution HD-TV camera. They are connected to a PC which is used for image processing.

3.1 Hardware

The PMD camera uses an intensity modulated sine wave of IR-light emitting diodes invisible for the human eye and for the HD color camera. The PMD camera has a small resolution (204x204 pixels) while the TV camera was used with 720p (1280x720 pixels) resolution. For a short time we could test a PMD CamCube 3.0 which is able to sync via genlock and received TV frame rates with a Region of Interst (RoI) of 174x142 pixels. The genlock signal was generated by a synchronization unit from IOTracker. However, this paper relies mainly on the CamCube 2.0 which wasn't able to reach a TV frame rate.
Both cameras are fixed to one camera head mounted on a tripod which pans and tilts the cameras conjointly. Figure 1 a) shows the assembly of both cameras with its corresponding camera model (fig. 1 b). The PMD camera uses the intensity image and the spatial depth data which has the principal point C_{ToF}. The HD camera delivers a high-resolution projection of the scene from its principal point C_{HD}. Both principal points differ in a rotation

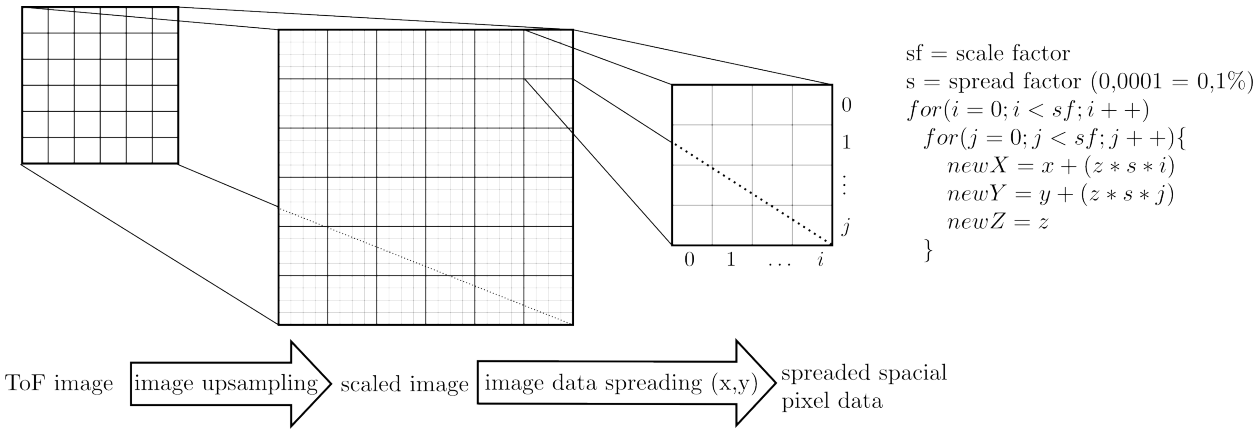

Figure 2: upscaling in image space and spatial pixel value spreading

R and translation t. The HD camera has the ability to change zoom and focus but the implementation doesn't support dynamic changes in optics yet.

3.2 Software

The developed software manages the calibration, image processing and augmentation with tracked objects. The synchronisation of the cameras is done by software triggers which is not perfect but triggers at nearly the same time.

3.2.1 Image Registration

The focus of the image registration step is a matching in pixel-space so that every pixel has an exact corresponding pixel with equal coordinates in either image.
At first both images are scaled to the same size. We use an upsampling method for the smaller time of flight (ToF) data without recovering details. Lindner et al. [LLK08] use a similar method. It simply spreads the known data by purely resizing without use of extrapolation. That means we resize the image by a scale factor assumed the images have the same aspect ratio. Every pixel gets blown up in x and y direction keeping the depth value z. The spread factor, the z-value and the pixel-distance in image space defines the value of the new pixel. So we generate similar pixels in the image grid caused by the spatial data as can be seen in figure 2.
Our rendering engine bases on the openCV library to implement the stereo calibration. The algorithm is based on Bouguets *calibrated stereo* using a checkerboard for the calibration process. By knowing the physical size of the chess grid we get the relative metric coordinates for the rotation and translation between the cameras.
The stereo calibration methods are used for an area based image registration. The aim of the stereo calibration is the rectification which means that the different cameras (HD and range) are row or column aligned depending on vertical or horizontal stereo parallax. The camera setup was adjusted with their optical axes to be as parallel as possible. In the process

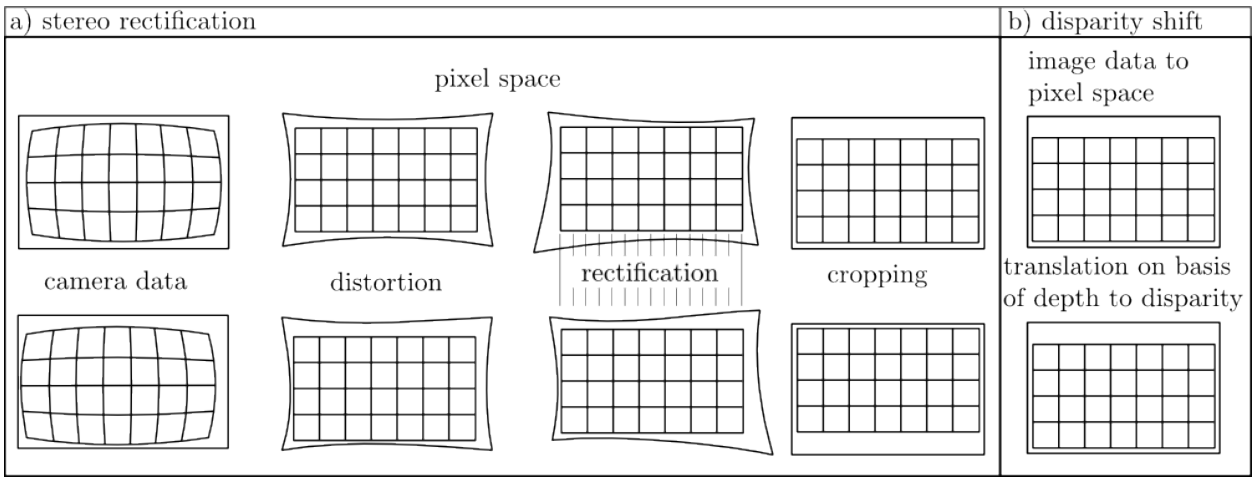

Figure 3: (horizontal) stereo rectification process with translated pixels by disparity

of stereo calibration the image planes are virtually turned so that they become axis-parallel. This is achieved by changing the projections of both image planes minimizing the differences and distortion in the images and maximizing view overlap (fig. 3a).
Normally the rectified images are used for stereo correspondence which is used to calculate the disparity map. In our approach we use the distance data to calculate the disparity for every pixel (fig. 3b). To be more clear we make no feature matching to determine the disparity. By knowing the depth for every pixel we calculate its disparity via triangulation and shift the pixel along the image column in case of horizontal stereo. All pixels are transformed by its disparity within the image grid. The result are two more or less identical images which differ in ghost regions. These regions are a result of the different camera viewpoints which go along with occlusion.

3.2.2 Disparity

A rectified horizontal stereo image pair is column aligned. Corresponding pixels are in the same column but differ in their y-coordinate on the image plane. The horizontal distance in pixels is called disparity and is conventionally used for a disparity map to determine the depth. It is also known that disparity is inversely proportional to depth.
The camera model for the perspective projection uses equation (1) to calculate the projection of world coordinate Y with the distance Z to the image planes which results in y_{ToF} and y_{HD}, respectively. The focal length for both cameras is equal at stereo calibration ($f_y = f_{ToF_y} = f_{HD_y}$). B is the baseline from the translation vector $t = (0, B, 0)$ between the optical centers. The difference of the same projected point on the image planes is the disparity δ [Müh02].

$$y_{Tof} = f_y \cdot \frac{Y}{Z} \ , \ y_{HD} = f_y \cdot \frac{Y - B}{Z} = y - f_y \cdot \frac{B}{Z} = y - \delta \tag{1}$$

The disparity is calculated by:

$$\delta = \frac{B \cdot f_y}{Z} \ . \tag{2}$$

3.2.3 Image Processing

The image processing implements a rough tracking that works for one person in the virtual studio (fig. 5). The actor needs to wear cloths with long arms. The image processing works on the images registered with the method we described before. The masks could be applied in the color and depth image without transformation. So we segment and classify the actor's hands and face in the color image and determine the spatial position in the depth image.
Three masks are combined to get the regions of skin-colored areas (fig. 4). One mask works in the color image and extracts the actor in the foreground. The algorithm used was described in [vdBL99] which is a fast software chroma keying algorithm. For the second mask the depth data is arranged into a histogram. A local maximum in the histogram bins at high values helps finding the background wall of the studio. The bins containing smaller values than the local maximum minus a threshold are used for histogram back-projection to build up the foreground mask.
An adaptive skin color tracker is used to identify the actor's hands and face. The previously explained masks are combined by a logical AND operator. The resulting foreground mask is observed by the skin tracker for the segmentation. The regions of interest are used to calculate the center of mass in the depth data to get the spatial position. The RoIs are also used to classify the face as the biggest area. The remaining regions are assumed to be hands.

4 Experiences

While calibrating the cameras we got blur effects in the color images. As a consequence, the pattern couldn't get recognized correctly and the results of the stereo calibration were of less quality. Moving the pattern slowly improved the pattern recognition. These blurred images have also some drawbacks to the image processing when the actor is blurred.
Another problem is the small resolution of the depth camera which prevented the calibration from delivering proper results and not every calibration process produced an acceptable result. Determining the orientation angle difference of camera mounting didn't provide the expected accuracy and an additional offset was used to correct in yaw and pitch for the

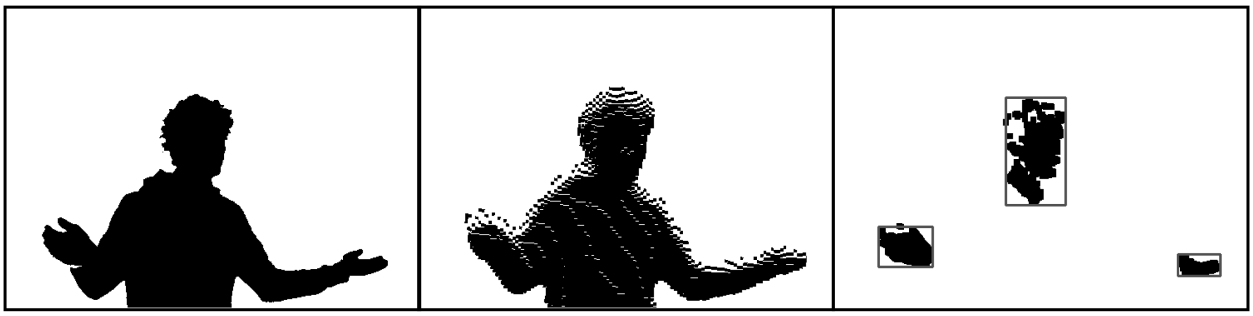

Figure 4: chroma, depth and skin mask

virtual camera.
The chroma keyer shows small holes in the mask at grey values. van Bergh el al. addressed this problem and stated it is quite acceptable because the effect appears and disappears very fast and there is nearly no effect on the segmentation. In addition the depth mask is noisy at edges (fig. 4). This occurs from IR-light reflections at object edges that lead to wrong distance measuring which are so called *flying pixels*. As a consequence the disparity translation is error propagation. Another characteristic are the holes in the depth mask which are caused by the forward transformation of disparity correction. The depth mask has no fine details because of the low resolution of the depth image.
Our method to map the color information onto the point cloud allows a dynamic mapping (fig. 6) that shows an object changing its distance to the cameras. The texture of dark sheet in the actors right hand maps correct without ideally becoming apparent on the background wall. This is an improvement to the homographic warp of the colored image.
Figure 7 shows the spatial camera data which is textured with the color images. Additionally the tracked body parts are marked with boxes and visualizes the scene in a 3D-viewer with a metric grid.
We render simple objects at the tracked positions to an openGL framebuffer. The virtual camera parameters were calculated from the position with the intrinsics of the HD camera [Li01]. The HD image is augmented by the rendered graphic to evaluate the effects of the spatial tracking (fig. 8). The interaction with virtual objects which are attached i.e. to the hands provide coherent mapping. The objects change in size while moving the hands towards the camera. So we could show a plausible interaction in virtual studios.

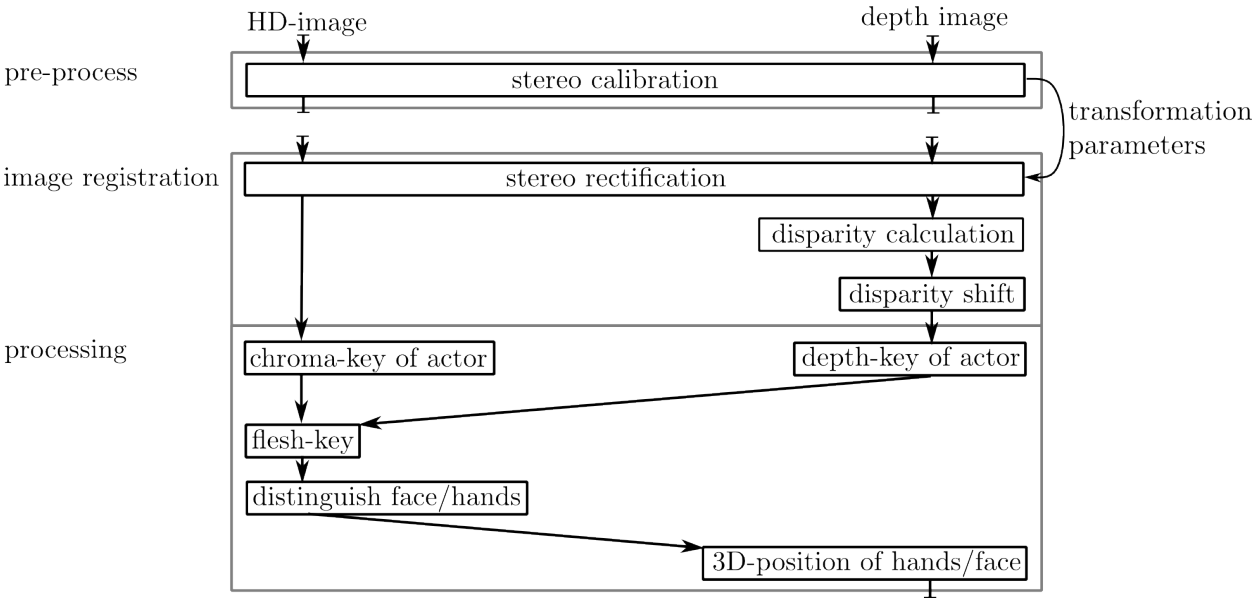

Figure 5: Diagram of the foregoing calibration and the image registration with tracking process for every frame.

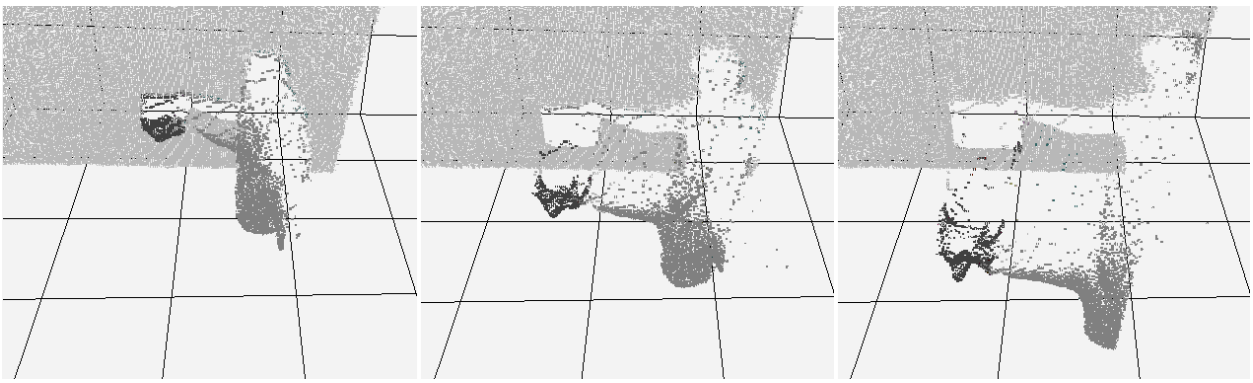

Figure 6: image registration which works in dependence of distance

5 Conclusion and Future Work

In this paper we could show that our approach works for image registration to match a depth camera with another camera. The disparity translation of pixels within the image grid of every pixel enables a dynamic method of image registration with a fast execution. Instead the distance of every pixel must be known for one camera and a previous calibration step needed to be performed. The image registration at changes in zoom and focus of the color camera wasn't done and needs to be solved. The augmented scene shows that the spacial tracking of an actor could be used for interaction with the virtual content.
The use of a camera with a higher resolution anticipate better calibration results and also a more practicable process of calibration. For the future we plan to use a Kinect camera, which is based on structured light technology [PKW11]. It has a higher resolution (640x480 pixels) and less noise as compared to a PMD camera. We expect to abandon a noise filter on depth data and use it for more fine-grained tracking. Measuring will have an insignificant smaller accuracy which doesn't matter in virtual studios. Kinect cameras have a smaller range (0.8-3.5m) and the usage of multiple units is very difficult because of overlapping light patterns. But in the described approach one camera is sufficient. It will be possible to reach TV-frametate but not the synchronisation with genlock. Furthermore we plan to use approved libraries for skeleton or hand tracking.

Acknowledgements

I would like to thank PMDTec for providing us the PMD CamCube. This work was carried out within a 'FHprofUnt' project supported by the Federal Ministry of Education and Research (BMBF), Germany (grant no. 17010X10).

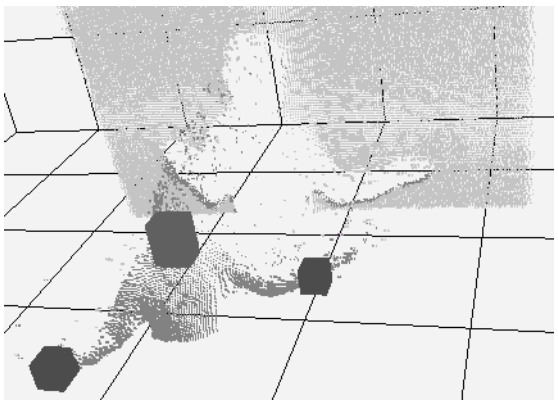

Figure 7: view of textured point cloud with boxes at tracked objects

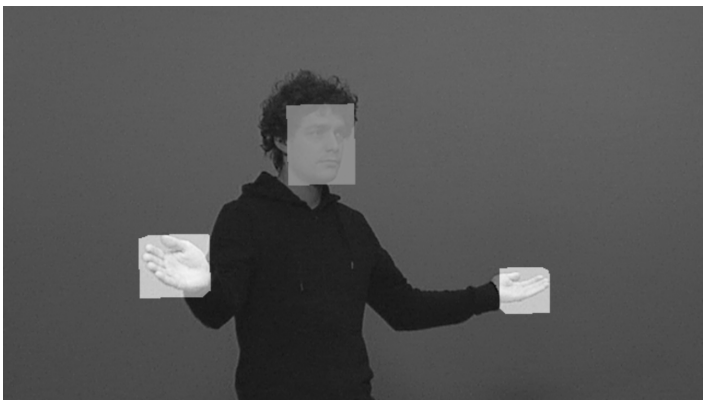

Figure 8: studio scene augmented with rendered virtual objects

References

[DT06] J. Diebel and S. Thrun. An application of markov random fields to range sensing. *Advances in neural information processing systems*, 18:291, 2006.

[KIA+00] M. Kawakita, K. Iizuka, T. Aida, H. Kikuchi, H. Fujikake, J. Yonai, and K. Takazawa. Axi-vision camera (real-time distance-mapping camera). *Applied Optics*, 39(22), August 2000.

[KIN+04] M. Kawakita, K. Iizuka, H. Nakamura, I. Mizuno, T. Kurita, T. Aida, Y. Yamanouchi, H. Mitsumine, T. Fukaya, H. Kikuchi, et al. High-definition real-time depth-mapping tv camera: Hdtv axi-vision camera. *Optics Express*, 12(12):2781–2794, 2004.

[KKKI02] M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue. Hdtv axi-vision camera. In *Proc. of International Broadcasting Conference*, pages 397–404, 2002.

[Li01] Ming Li. Correspondence analysis between the image formation pipelines of graphics and vision. In Sánchez J. Salvador and Pla Filiberto, editors, *Proceedings of the IX Spanish Symposium on Pattern Recognition and Image Analysis*, pages 187–192, Benicasim(Castellón), Spain, May 2001. Universitat Jaume I, Publications de la Universitat Jaume I.

[LKH07] M. Lindner, A. Kolb, and K. Hartmann. Data-fusion of pmd-based distance-information and high-resolution rgb-images. In *Signals, Circuits and Systems, 2007. ISSCS 2007. International Symposium on*, volume 1, pages 1–4. IEEE, 2007.

[LLK08] M. Lindner, M. Lambers, and A. Kolb. Sub-pixel data fusion and edge-enhanced distance refinement for 2d/3d images. *Int. J. Intell. Syst. Technol. Appl.*, 5(3/4):344–354, 2008.

[Müh02] K. Mühlmann. *Design und Implementierung eines Systems zur schnellen Rekonstruktion dreidimensionaler Modelle aus Stereobildern.* PhD thesis, Universität Mannheim, November 2002.

[PHW+07] J. Penne, K. Höller, D. Wilhelm, H. Feußner, and J. Hornegger. Photorealistic 3-d surface reconstructions using tof cameras. *RBC Biomedical Engineering*, 2007.

[PKW11] F. Pece, J. Kautz, and T. Weyrich. Three depth-camera technologies compared. *Beaming Workshop*, 2011.

[STDT08] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. High-quality scanning using time-of-flight depth superresolution. *CVPR Workshop on Time-of-Flight Computer Vision 2008*, 2008.

[vdBL99] F. van den Bergh and V. Lalioti. Software chroma keying in an immersive virtual environment. *South African Computer Journal,(24)*, pages 155–162, 1999.

[YYDN07] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

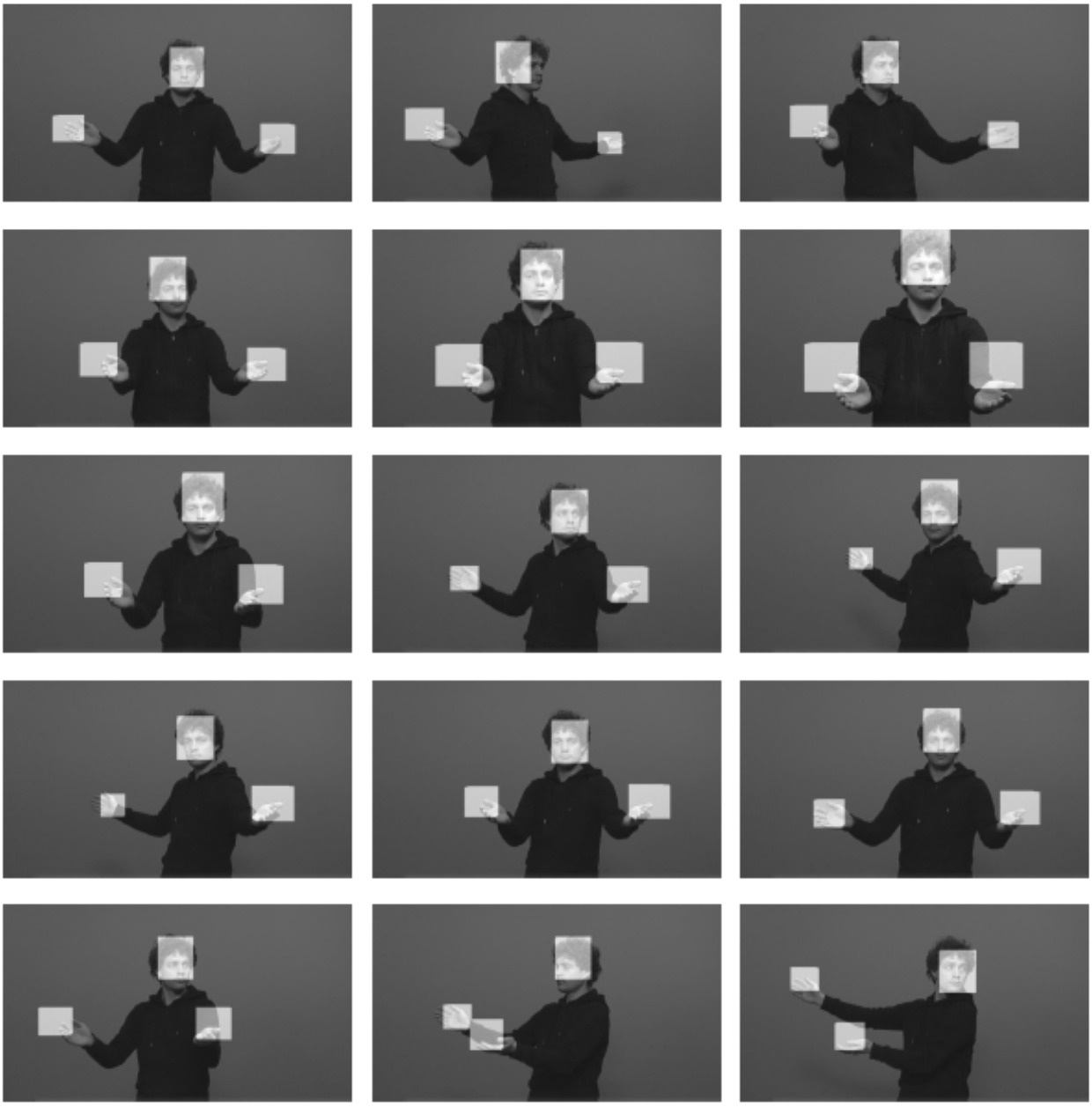

Figure 9: A tracking sequence - The boxes at the hands have the same size in the openGL scene. In the rendering they have different sizes and give the user a depth cue.