

Augmented Reality for 3D TV using Depth Camera Input

François de Sorbier, Yuki Takaya, Yuko Uematsu, Ismael Daribo and Hideo Saito
Graduate School of Science and Technology, Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, 223-8522, Japan
Email: {fdesorbi,takaya,yu-ko,daribo,saito}@hvrl.ics.keio.ac.jp

Abstract—This paper presents a capture system based on a depth camera that is used for an augmented reality application. Most of depth cameras are unable to capture the color information corresponding to the same viewpoint. A color is then added besides the depth camera and we applied a transformation algorithm to match the depth map with the color camera's viewpoint. Then, using a depth image based rendering (DIBR) approaches, it becomes possible to synthesize new *virtual* views from the 2D-plus-depth data. We also address some research issues in the generation of the *virtual* view is to deal with the newly exposed areas, appearing as holes and denoted as occlusions, which may be revealed in each warped image. The color image and its corresponding enhanced depth image are then combined to produce a mesh representing the real scene. It is used to easily integrate virtual objects in the real scene. Finally, the result can be rendered to create the input image required by a auto-stereoscopic screen.

Index Terms—Depth Video, 3D TV, Augmented Reality, Depth Image Based Rendering, Auto-stereoscopic display

I. INTRODUCTION

3DTV has been recently considered as the next generation of multimedia consumer products. For that purpose, the development of stereoscopic displays [1] has been considerably promoted by several display manufactures.

For such purpose, a lot of efforts have already been performed for generating depth videos from multiple video inputs. Some solutions use as many cameras as the number of input views required by 3D displays, while some others focus on reducing the number of camera by taking advantage of an image based rendering algorithm. One another solution is to use a color image with its corresponding depth image. That technology, named video-plus-depth or 2D-plus-depth, can generate virtual viewpoint videos, free-viewpoint videos, or even change the baseline of the stereoscopic viewing. One approach to quickly capture the depth information from a real environment is to take advantage of a depth camera, such as [2].

This paper is structured as follow. First, we introduce our recent researches related to analysis and synthesis of depth video for 3D TV applications. For generating such 3D TV contents in a flexible manner, such as free-viewpoint display or modifying baseline for the stereoscopic display, the so called 2d-plus-depth data representation is an essential way to convey 3D information of the video. Depth video is not generally rendered for display, but used for rendering color video at virtual viewpoint, which is called as the depth image based

rendering (DIBR). We describe about DIBR for realizing using a 2D a real-time capturing and displaying on auto-stereoscopic display color camera and a depth camera. We also address some research issues in the generation of the *virtual* view is to deal with the newly exposed areas, appearing as holes and denoted as occlusions, which may be revealed in each warped image. Finally, we present our augmented reality application. We explained how we can take advantage of the 2D-plus-depth format to create a mesh. This mesh provides an easy way to add virtual objects and can also be used to compute interactions between virtual objects and the real scene.

II. DEPTH CAMERA FOR AUTO-STEREOSCOPIC DISPLAY

A. System and Data Processing

For real-time capturing of 3D content from live real scene and displaying it on auto-stereoscopic displays, a depth camera can be used to obtain depth information from a real scene. By associating the depth image with the corresponding color image, it is possible to quickly generate the input images for auto-stereoscopic displays as shown in Figure 1.

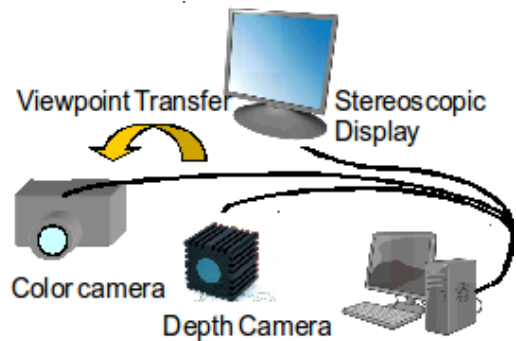


Fig. 1. Depth Camera with Color camera for video capturing to display onto auto-stereoscopic display.

A depth camera is a device that is able to capture and transmit the depth information corresponding to a given environment. The camera of our system is based on the Time Of Flight technology [2] that measures distances in a scene in real time. This camera, also known as TOF camera, emits infrared light that is reflected by the environment and come back to the camera's sensor. The traveling time of the light is then measured for each pixel of the sensor and used for computing the depth of the scene.

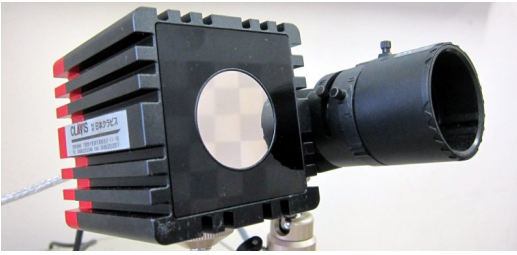


Fig. 2. A color camera is added in the system since our depth camera cannot capture the color information.

However, the depth camera cannot capture color information and its resolution is low (176×144 for example). Another camera is then required in order to capture the color information. Since color is also an important cue for the perception of depth [3], [4], we use a high resolution camera to compensate the low resolution of the depth camera. The color camera is added beside the depth camera as presented in Figure 2.

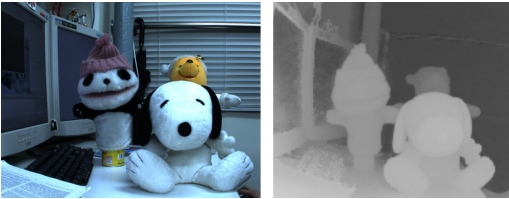


Fig. 3. Color Camera image (left) and depth camera image (right). Each viewpoint is different from each other, so the captured position of the hand is not the same

Since the color camera and the depth camera are not located at the same position, their viewpoints are slightly different as shown in Figure 3. Therefore, a transformation is required to match the depth map with the color image. When mapping color and depth images, a lack of data occurs because of occlusions. Since color is the most relevant visual information perceived by humans, the transformation is then applied on the depth map. The process of our system is presented in Figure 4. The process consists in two parts. First, considering that both cameras are fixed, we pre-compute the intrinsic and extrinsic parameters of the cameras. Second, for each input frame from the depth camera, we apply the view transformation technique to match the color image by using the computed camera parameters. The result is then converted according to the requirement of our auto-stereoscopic screen input format.

The pose estimation of depth and color cameras has to be defined in the same coordinate system in order to be able to project the mesh generated from the depth map onto the color image. For each pixel of the depth image, the depth camera also provides the corresponding 3D coordinate. This set of points is defined in a coordinate system wherein the depth camera's position is the origin. Following this statement, the depth camera is also set as the origin of our capture system. Thus, the calibration stage only requires evaluating the parameters of the color camera.

Using a set of 2D/3D correspondences is a common way to

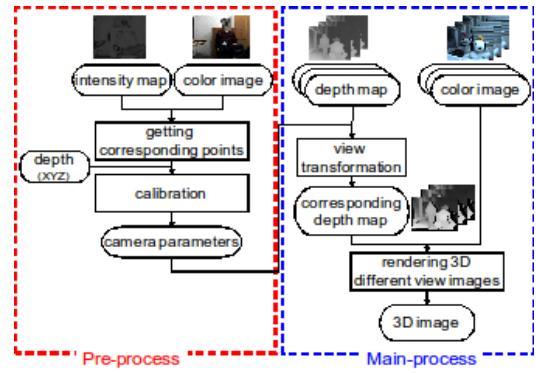


Fig. 4. Flow chart of our method.

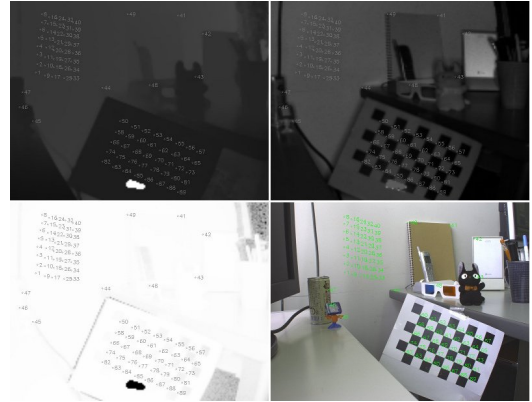


Fig. 5. Four images are used for the calibration. (A) the depth map, (B) the corresponding gray scale image, (C) the confidentiality map and (D) the color image.

estimate the pose of a camera. In our case, the 3D coordinates can be easily retrieved since the capture system is composed of a depth camera. First, 2D correspondences are found between the depth and color images by using a chessboard pattern or by clicking pixels. Three kinds of images generated by the depth camera associated with the color image are used to define these correspondences as depicted in Figure 6. The depth and gray-scale images are used to select relevant points whereas the confidence map is used to check the validity of the depth value computed by the camera (white areas represent a high confidentiality). Second, since a 3D coordinate exists for each pixel of the images generated by the depth camera (Figure 6), a list of 2D/3D correspondences between the color image and the 3D space is created.

B. Results and Discussion

We present results to check the availability of our proposed method using the following environment.

- CPU Core 2 Duo : 3.0 GHz
- Memory : 2 GB
- Resolution of color camera : 640×480
- Resolution of depth camera : 176×144

Figure 7 depicts the view transformation result. The depth image shown in the middle is transformed to match the

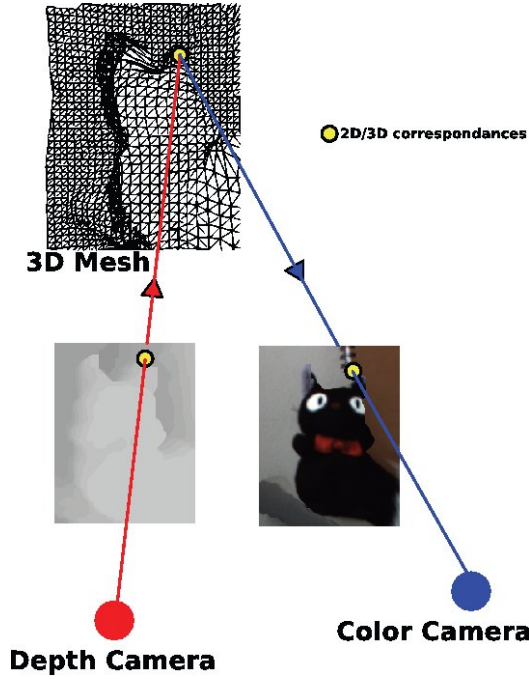


Fig. 6. 3D coordinates corresponding to the depth value provided by the depth camera helps to calibrate the capture system.

upper image. The lower image is the result of our view transformation algorithm. The matching correctness of the result between the depth map and the color image is shown thanks to the red lines.

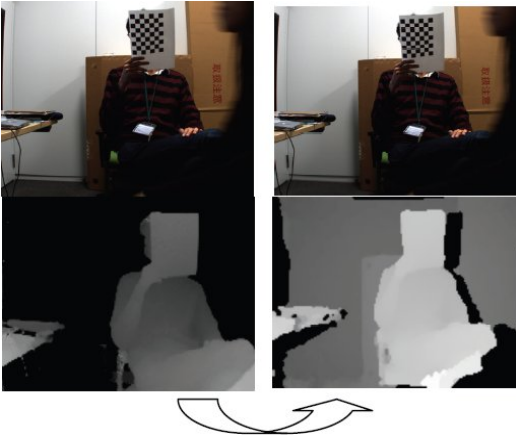


Fig. 7. Example of viewpoint transfer of depth video. The depth image shown in the bottom left is transferred to bottom right, so that the viewpoint can be at the same position as the color image. The color images on the top is the same, but shown for the comparison of the depth images.

Even though the transfer of the depth video to the color camera viewpoint is well known technique in computer vision area, there are some research issues to be considered: rendering speed, hall region in the transferred depth video by occlusion, and resolution difference between color camera and depth camera. The hall region cannot be avoided along the area of the object boundary as shown in the bottom right in Figure 7. This

is a significant issue in such viewpoint transfer from captured depth camera. We will tackle with those issues to improving the quality of the 3D video contents in near future research. Section 3 is also presenting a way of avoiding such hall areas caused by the occlusion.

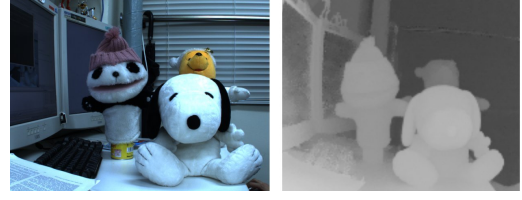


Fig. 8. Result of viewpoint transfer of depth video.

Another result is presented in Figure 3. In this case, we apply a median filter to fill the missing areas. Some occlusion areas represented by black regions in the result image cannot be avoided using the median filter method. Black areas on the bottom and left parts are the consequences of the radial distortion correction applied by the hardware of the depth camera.

The TOF depth camera can generate and transmit the data in real time. In that sense, our system needs computational cost only in the translation phase. The full process of our system runs at an average frame-rate of 10 frames per second. This can be improved by using a GPU framework.

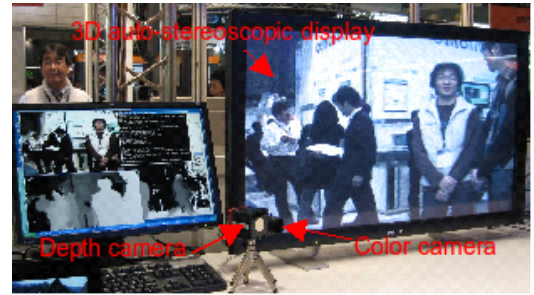


Fig. 9. System for real-time 3D capturing with a depth and a color for auto-stereoscopic display.

In InterBEE 2009 (International Broadcasting Equipment Exhibition), held in November 2009, we demonstrated the prototype system that shows 3D video captured with both color cameras and depth cameras onto an auto-stereoscopic display. Figure 9 shows the picture of the exhibition.

III. PRE-PROCESSING OF THE DEPTH VIDEO FOR FILLING IN THE DISOCCLUSIONS

In this section, we address the problem of filling in the disocclusions within a stereoscopic camera framework, with a small camera inter-distance (around to the human eyes inter-distance). One way to deal with the disocclusion problem is to pre-process the depth video, for example by smoothing, commonly operated with a Gaussian filter. Instead of smoothing the whole depth video, we propose here an adaptive filter taking into account the distance to the edges. The

proposed scheme is summarized in Figure 10. First we apply a preliminary pre-processing stage to extract the edges of the depth map capable of revealing disoccluded areas, that we will refer to in the following as Contours of Interest (CI). This spatial information permits then to compute the distance data, and also to compute the weight information for the proposed filtering operation.

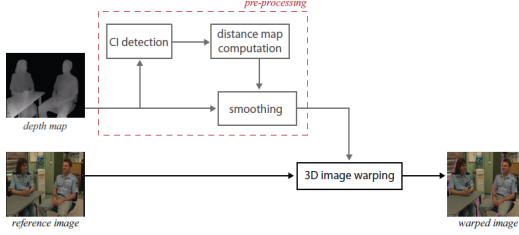


Fig. 10. Pre-processing of the depth map before the 3D image warping.

A. Extraction of regions around the Contours of Interest (CI)

In Figure 11, we can see the resulting warped picture from the 3D image warping process according to the camera set-up such as described in Section 2. The 3D image warping has exposed areas of the scene for which the reference camera has no information (here colored in magenta). These areas are precisely located around the CI of objects and we can identify the location of these regions before the 3D image warping by applying the following preprocessing.

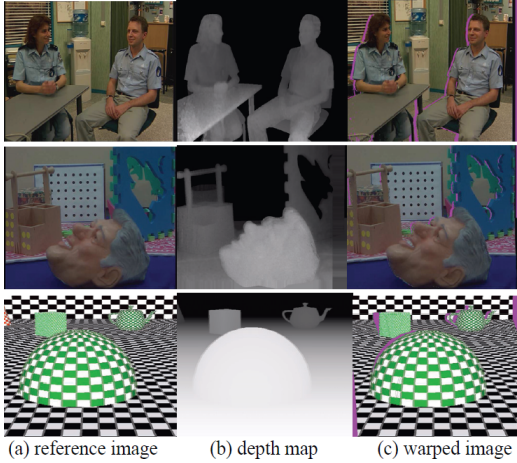


Fig. 11. In magenta: newly exposed areas in the warped picture (from the ATTEST test sequences (up) *Interview*, (middle) *Orbi* and (bottom) *Cg*).

The CI are generated from the depth map by applying a directional edge detector, such that only one edge side is detected, as illustrated in Figure 12. To handle the problem of choosing an appropriate threshold, we use an approach by hysteresis1, where multiple thresholds are used to find an edge. The resulting binary map reveals on the depth map areas where displacement is high, and thus, where it is necessary to apply a strong smoothing, leading to a reduction or even an elimination of the dis-occluded areas in the targeted view.

B. Distance map

Discrete distance map computing is commonly used in shape analysis to generate skeletons of objects [5]. Here, we propose to utilize the distance map computation to calculate the shortest distance from a point to a CI. Moreover, we use the distance information as a weight for the filter adaptation.

In a distance map context, a zero value indicates that the pixel belongs to the CI. Subsequently, non-zero values represent the shortest distances from a point to the CI. It is possible to take into account the spatial propagation of the distance, and compute it successively from neighboring pixels with a reasonable computing time, with an average complexity linear in the number of pixels. The propagation of distance relying on the assumption that it is possible to deduce the distance of a pixel from the value of its neighbors, is well suited for sequential and parallel algorithms. One example of distance map is shown in Figure 12.

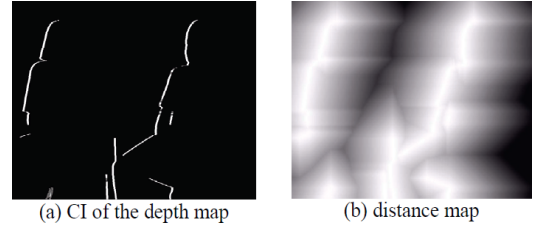


Fig. 12. Examples of distance map derived from the CI (from the 1st frame of the ATTEST test sequences *Interview*).

C. Experimental results

For the experiments, we have considered the three Advanced Three-dimensional Television System Technologies (ATTEST) video-plus-depth test sequences *Interview*, *Orbi* and *Cg* (720×576, 25fps) [6]. We start by comparing our solution with the classical *all blur* solution consisting in applying the Gaussian filter on the whole image. The conditions of the experiments are done in a symmetric and asymmetric fashion way. We can see in Figure examples of resulting preprocessing of the depth map through the means of a Gaussian filtering and the proposed framework. While a Gaussian filtering smooths all the depth map uniformly, our proposed approach focuses on the areas susceptible of being revealed in the warped image. As a consequence, less depth filtering induced distortions are introduced in the warped picture Figure 13, and in the meantime the disoccluded regions are removed in the warped image, as we can see in Figure 14.

However, the depth-filtering-induced distortion may provoke geometric distortion in the warped picture, where vertical line bends (as shown in Figure 14 around the head of the cop). To overcome this issue, we investigate, as proposed by Zhang *et al.* [7], an adaptive asymmetric filtering of the depth map. As we can see in Figure 16, the asymmetric nature of the filter tends to reduce the amount of geometric distortion that might be perceived, and straightens the vertical lines.

In this section, we have introduced a new adaptive filter for 3D image warping, taking into account the distance to

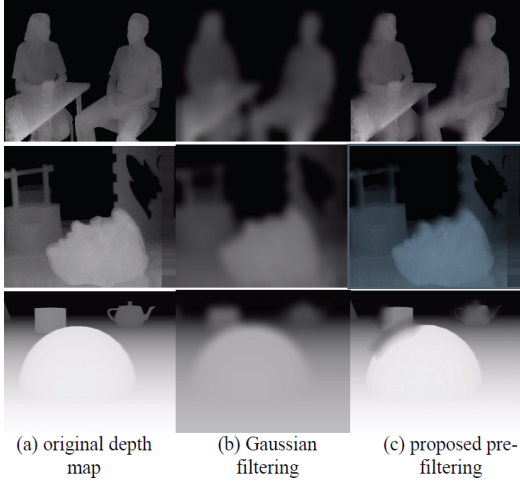


Fig. 13. Examples of symmetric pre-filtering depth map with the proposed framework (from the ATTEST test sequences (top)*Interview*, (middle)*Orbi* and (bottom)*Cg*).



Fig. 14. Error comparison between the original depth map of the pre-processed depth map (from the ATTEST test sequences (top)*Interview*, (middle)*Orbi* and (bottom)*Cg*).

object boundaries. The main advantage consists in limiting any unnecessary filtering-distortion in the depth map. Experiment results have illustrated the high efficiency of the proposed method. Of course, any smoothing filter can be used instead of the used Gaussian one. To deal with the geometric distortion, we applied an asymmetric filter, which is possible since the Gaussian filter is separable. An improvement could be expected by applying according to the direction of the gradient in each part of the image, an adaptive asymmetric filtering to prevent the vertical and horizontal lines from bending.

The depth pre-processing approach is particularly efficient

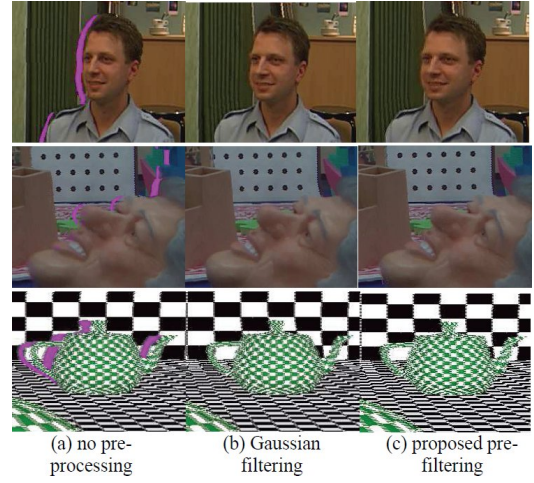


Fig. 15. Warped images issue from the 3D image warping using the different symmetric pre-processed depth map (from the ATTEST test sequences (top)*Interview*, (middle)*Orbi* and (bottom)*Cg*).



Fig. 16. Comparison between symmetric and asymmetric filtering with the proposed framework.

when the baseline is relatively small, for example in the case of a stereoscopic rendering using a baseline equivalent to the average human inter-eye distance. When the baseline becomes larger, this approach would introduce bigger geometric distortions, that will be difficult to handle by just applying an asymmetric filtering strategy. To deal with large baselines, we can consider a post-processing approach on the warped image to fill in the disoccluded regions by inpainting techniques [8].

IV. AUGMENTED REALITY USING A DEPTH CAMERA

According to the previous sections, we can generate the depth map corresponding to the viewpoint of the color camera. Especially, we presented methods to resolve the viewpoint difference and occlusions problems. Color image and its corresponding depth map are very useful information when it is applied to augmented reality. The goal is then to integrate virtual objects in a real scene in a way they appear to be really existing. Moreover, it could be also possible to interact with these virtual objects. However, this task has many constraints even if the depth map reduces the difficulty of the integration of virtual objects.

In our approach, we propose to take advantage of a mesh generated from the depth values. The conversion of depth data into a single mesh requires to know the intrinsic parameters of the color camera, which were already evaluated during the view transformation stage. Then, we compute the normalized vectors that have their origin at the camera's center and are

pointed toward every pixels of the image. We multiply each vector by the corresponding metric depth value of the pixel in order to obtain a cloud of 3D points representing the real scene. The mesh is generated by linking those 3D points with triangles. Finally, the color image is projected onto the mesh using the projective texture approach [9], which produce a 3D model of the real scene that can be used for our augmented reality application.

Our application consists in a virtual ball added in the real scene with which the user can interact. For that purpose, we add a physics engine like Bullet [10] that can manage dynamic meshes. Such a library helps to detect collisions among the triangles of several virtual objects and computes physical effects like bouncing (Figure 17). The main difficulty with this approach is that our mesh is animated, so errors can occurred in collision detection when the mesh is moving too fast.

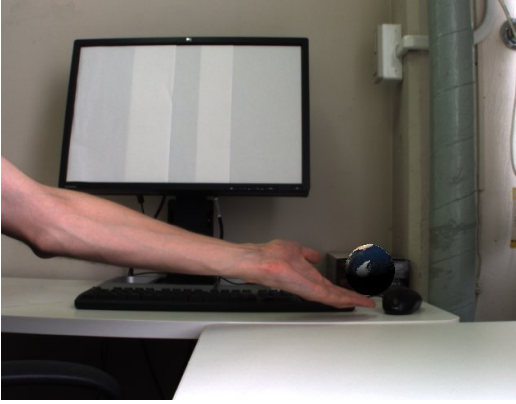


Fig. 17. Our application proposes an interaction with a virtual ball in real-time.

A virtual light source can be added and its effect applied on the captured environment. For each vertex of our mesh, we compute the corresponding normal and we use a shader to define the new illumination. However, the instability of data captured by the depth camera produces small waves on flat surfaces that decrease the visual quality of the result. The light source can also be used to cast shadows. Thanks to the shadow mapping algorithm [11], virtual objects cast their own shadow on the real scene and *vice-versa* as depicted in Figure 18.



Fig. 18. On the left image: the augmented scene without shadows. On the right image: real objects cast shadows on virtual objects and *vice-versa*.

V. CONCLUSIONS

In this paper, we introduce our recent researches related to analysis and synthesis of depth video for augmented reality and 3D TV applications.

First, we presented a capture system based on a color camera and a depth camera in order to generate input images for auto-stereoscopic displays. In particular, our approach can generate the color image with its corresponding depth map that are required by the well known 2D plus depth format. However, the TOF depth camera does not provide color information and is low resolution. It means that captured data cannot be used directly with the auto-stereoscopic display. We use the high resolution color camera located besides the depth camera. A high resolution color camera is added beside the depth camera. The difference of viewpoint is corrected by applying a view translation techniques that transform the depth information to match the color camera viewpoint.

Second, we present a way of filling hole areas that are caused by occlusion in generating the virtual view images via depth image based rendering (DIBR). We demonstrate that the pre-filtering of the depth map for reducing the discontinuity in the depth map is efficient for filling the hole areas. However, this approach can be applicable only if the baseline of the input views is small. The way of filling the hole area in case of larger baseline should be the research issue in near future.

Finally, we described an application of our capture system in order to augment a real scene with virtual objects. We explained how to convert the depth map into a mesh and use it to easily add virtual elements as for a normal computer graphics application. By integrating a physics engine to our system, it becomes possible to interact with virtual objects. We also presented how to add a virtual lighting in order to generate shadows on virtual objects and also on the real scene. However, the mesh suffers of a capture instability that, for example, avoid virtual objects to completely lay on a flat surface. We plan, in future works, to detect such surfaces and replace it with a stabilized mesh.

ACKNOWLEDGMENT

This work is partially supported in part by a Grant-in-Aid for the GCOE for high-Level Global Cooperation for Leading-Edge Platform on Access Spaces from MEXT, Japan. This research is partially supported by National Institute of Information and Communications Technology, Japan.

REFERENCES

- [1] N. Dodgson, "Autostereoscopic 3D displays," *Computer*, vol. 38, no. 8, pp. 31–36, 2005.
- [2] T. Oggier, M. Lehmann, R. Kaufmann, 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)," in *Proc. SPIE*, vol. 5249, no. 65, 2004, pp. 534–545.
- [3] T. Troscianko, R. Montagnon, J. L. Clerc, E. Malbert, and P.-L. Chanteau, "The role of colour as a monocular depth cue," *Vision Research*, vol. 31, no. 11, pp. 1923 – 1929, 1991.

- [4] L. Meesters, W. IJsselsteijn, and P. Seuntjens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, 2004.
- [5] Y. Ge and J. Fitzpatrick, "On the generation of skeletons from discrete Euclidean distance maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 11, p. 1055, 1996.
- [6] C. Fehn, K. Schuur, I. Feldmann, P. Kauff, and A. Smolic, "Distribution of ATTEST test sequences for EE4 in MPEG 3DAV," in *MPEG Meeting-ISO/IEC JTC1/SC29/WG11, MPEG02/M*, 2002.
- [7] W. Liang Zhang Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Transactions on Broadcasting*, vol. 51, no. 2, 2005.
- [8] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, 2004.
- [9] C. Everitt, "Projective texture mapping," *White paper, NVidia Corporation*, 2001.
- [10] Bullet Physics Library, "<http://bulletphysics.org/>," website, 2010.
- [11] L. Williams, "Casting curved shadows on curved surfaces," *ACM SIG-GRAPH Computer Graphics*, vol. 12, no. 3, pp. 270–274, 1978.