ASIAGRAPH 2010 PROCEEDINGS

# ３次元ＴＶのためのデプスビデオ解析と合成
# Depth Video Analysis and Synthesis for 3D TV

斎藤 英雄, 高谷 優樹, 植松 裕子, ダリボ イズマエル, ドゥソルビエ フランソワ/慶應義塾大学大学院理工学研究科
Hideo Saito, Yuki Takaya, Yuko Uematsu, Ismael Daribo, François de Sorbier / Graduate School of Science and Technology, Keio University
saito@hvrl.ics.keio.ac.jp, takaya@hvrl.ics.keio.ac.jp, yu-ko@hvrl.ics.keio.ac.jp, daribo@hvrl.ics.keio.ac.jp, fdesorbi@hvrl.ics.keio.ac.jp

Abstract: Depth video plays a significant role for generating 3D TV contents. Multiple view video (at least two views) is generally captured for 3D TV based on stereoscopic display. 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 video-plus-depth data representation is an essential way to convey 3D information of the video. In this representation, the depth image based rendering (DIBR) approaches have been recognized as a promising tool which can synthesize some new "virtual" views from the video-plus-depth data representation. In this paper, we introduce our recent researches related to DIBR for realizing a real-time capturing and displaying on auto-stereoscopic display using a 2D 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 disocclusions, which may be revealed in each warped image.

Keywords: Depth Video, 3D TV, Computer Vision, Multiple View Video, Image Based Rendering, Broadcasting

## 1. Introduction

3DTV has been recently considered as the next generation of multimedia consumer products. For that purpose, the development of stereoscopic displays [Dodgson 2005] has been considerably promoted by several display manufactures.

For capturing videos for such stereoscopic displays, stereo video capturing are generally performed using two cameras with some baselines. Such stereo capturing can implicitly capture some sense of 3D structure of the scene, but it is just a two channel of 2D videos. For handling real 3D for flexible purpose of 3DTV, depth is significant information, so the depth video representation is essentially needed for general purpose of 3DTV.

For such purpose, a lot of efforts have already been performed for generating depth videos from multiple video inputs. Some of those technologies can successfully provide depth video so that we can generate virtual viewpoint videos, free-viewpoint videos, or even change the baseline of the stereoscopic viewing. We can also directly capture depth video using depth cameras which are recently been available, such as [Oggier2004].

In this paper, 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 video-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 disocclusions, which may be revealed in each warped image.

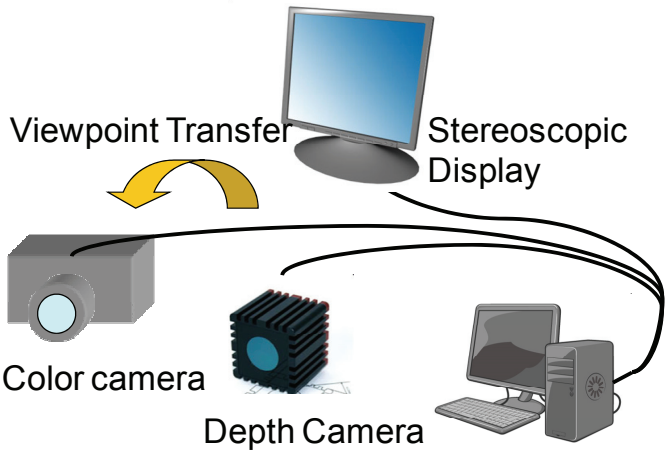



Figure 1: Depth Camera with Color camera for video capturing to display onto auto-stereoscopic display.

## 2. Real-time capturing and displaying on auto-stereoscopic display

### 2.1. 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.

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 [Oggier2004] 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. 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 [Troscianko1991], [Meesters2004], 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.

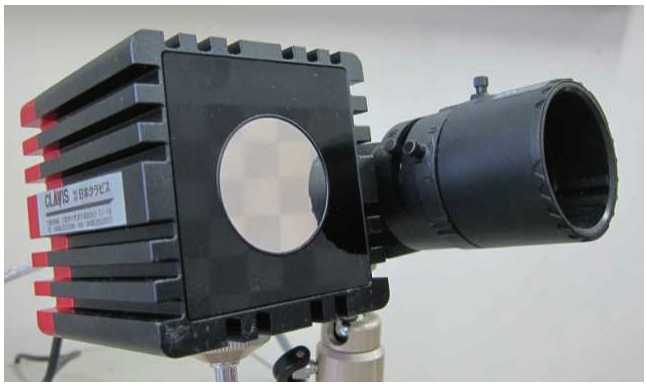
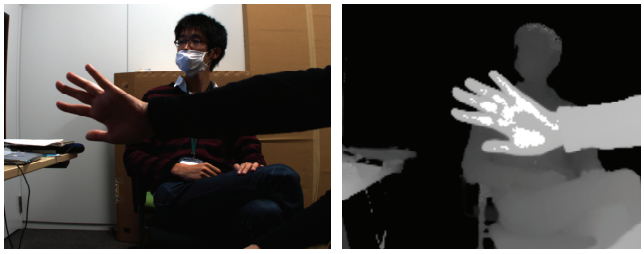


Figure 2: A depth camera is added to the system since the depth camera cannot capture color images.

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.



(a) Color camera image (b) Depth camera image

Figure 3: Color Camera image and depth camera image. Each viewpoint is different from each other, so the captured position of the hand is not the same.

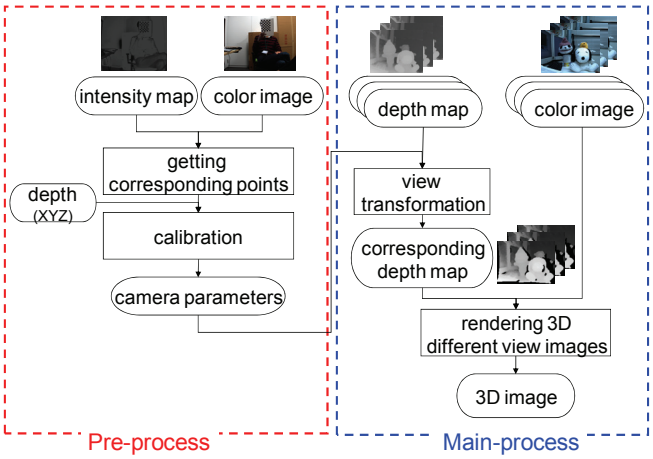



Figure 4: Flow chart of our method.

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 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 5. The depth and grayscale 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.

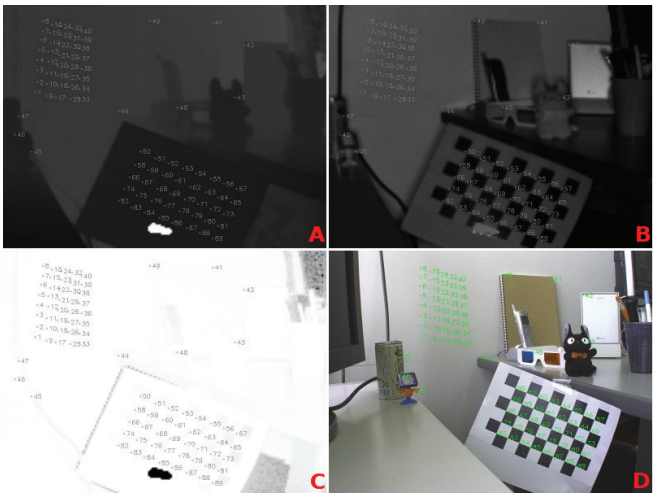



Figure 5: Four images are used for the calibration. (A) the depth map, (B) the corresponding gray scale image, (C) the confidentiality map and (D) the color Image.

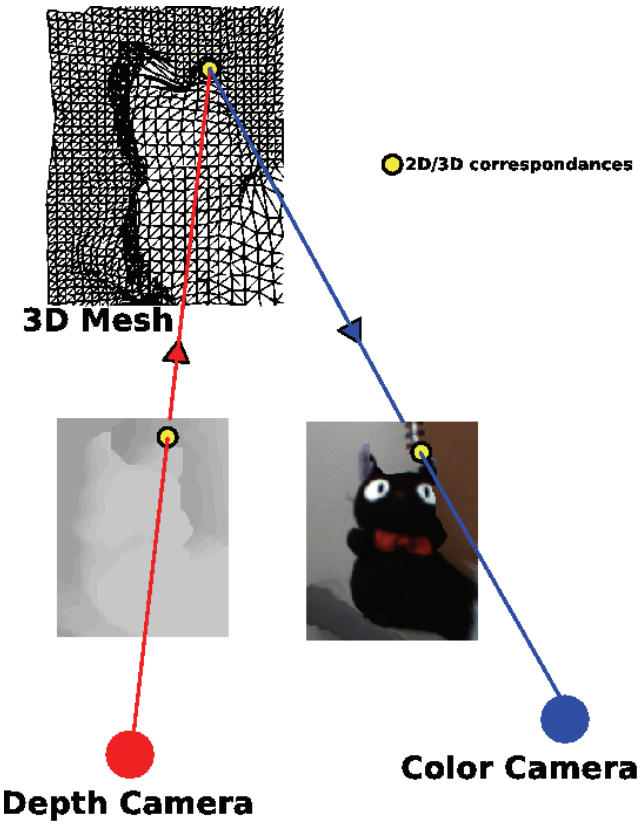



Figure 6: For each pixel of the depth-map corresponds a 3D coordinate. The depth information can be used to create a mesh and to deduce a 3D coordinate for some pixels of the color image.

## 2.2. 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 x 480
- Resolution of depth camera : 176 x 144

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

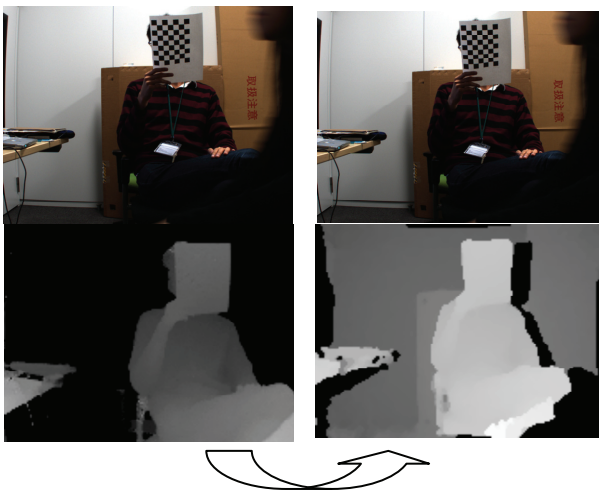


Figure 7: Example of viewpoint transfer of depth video. The depth image shown in the bottom left is transferred to bottom right, so that the viewpoint can be at the same position as the color image. (The color images on the top is the same, but shown for the comparison of the depth images.

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.

Another result is presented in Figure 8. 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 frame / seconds. This can be improved by using a GPU framework.

Figure 8. Result of viewpoint transfer of depth video.

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.

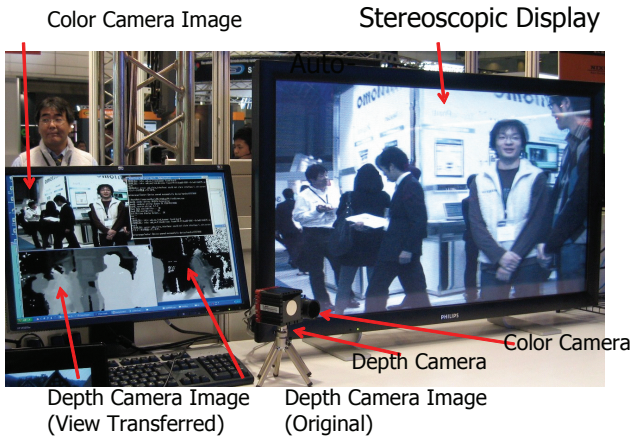



Figure 9: System for real-time 3D capturing with a depth and a color for auto-stereoscopic display.

## 3. 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 interdistance). 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.

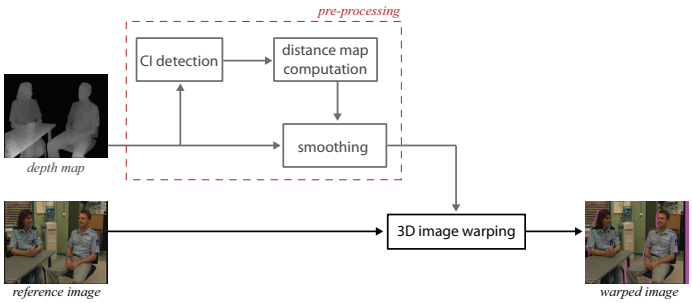
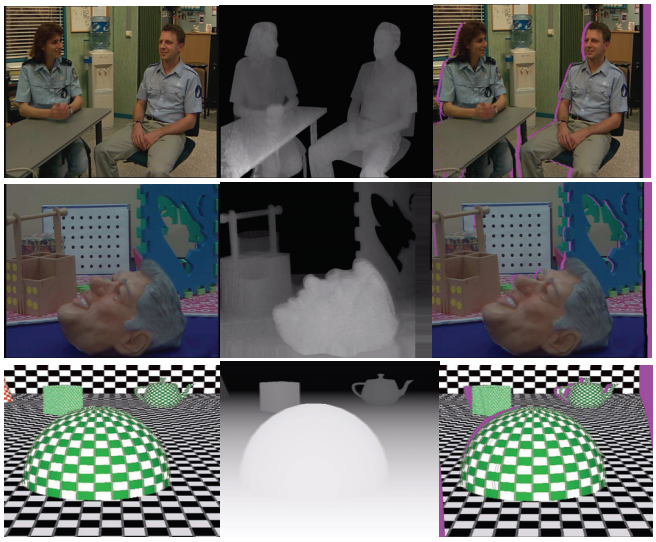



Figure 10: Pre-processing of the depth map before the 3D image warping.

### 3.1. 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.



(a) reference image (b) depth map (c) warped image

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

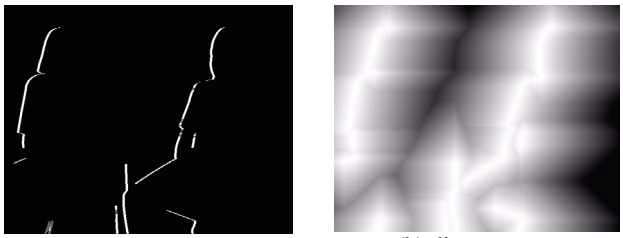
The CI are generated from the depth map by applying a directional edge detector, such that only one edge side is detected, as illustrated in Figure 10. To handle the problem of choosing an appropriate threshold, we use an approach by

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 disoccluded areas in the targeted view.

### 3.2. Distance map

Discrete distance map computing is commonly used in shape analysis to generate skeletons of objects [Ge1996]. 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.



(a) CI of the depth map (b) distance map

Figure 12: Examples of distance map derived from the CI (from the 1st frame of the ATTEST test sequences "Interview").

### 3.3. 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) [Fehn2002].

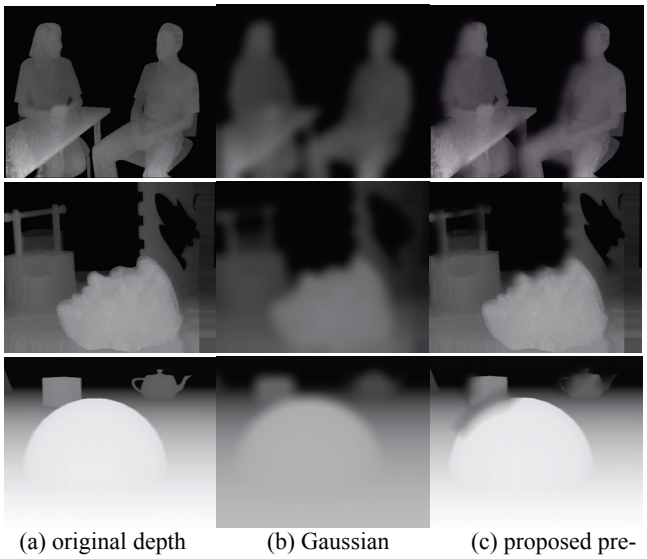
We start by comparing our solution with the classical "all blur" solution consisting in applying the Gaussian filter on the whole image. The conditions of the experiments are done in a symmetric and asymmetric fashion way.

We can see in Figure 13 examples of resulting pre-processing 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 14, and in the meantime the disoccluded regions are removed in the warped image, as we can see in Figure 15.

However, the depth-filtering-induced distortion may provoke geometric distortion in the warped picture, where vertical line bents (as shown in Figure 14 around the head of the cop). To overcome this issue, we investigate, as proposed by Zhang et al. [Zhang2005], 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.

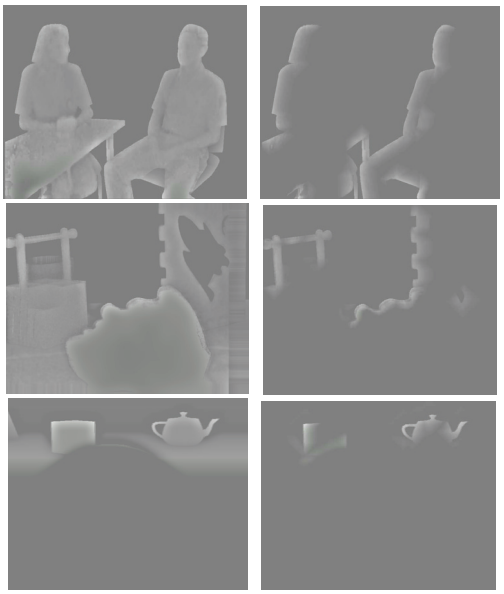
In this section, we have introduced a new adaptive filter for 3D image warping, taking into account the distance to object boundaries. The main advantage consists in limiting any unnecessary filtering-distortion in the depth map. Experiment results have illustrated the high efficiency of the proposed 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 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 [Criminisi2004].



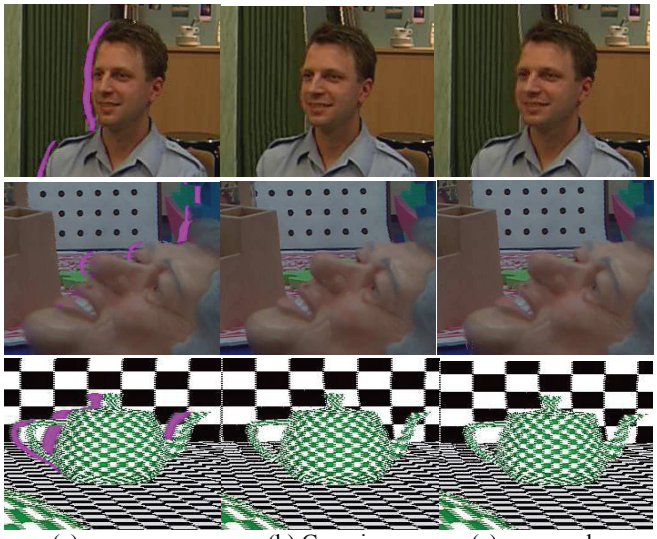
(a) original depth map (b) Gaussian filtering (c) proposed pre-filtering

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



(a) Gaussian filtering (b) proposed pre-filtering

Figure 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”).



(a) no pre-processing (b) Gaussian filtering (c) proposed pre-filtering

Figure 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”).

(a) no pre-processing (b) symmetric pre-filtering (c) asymmetric pre-filtering

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

## 4. Conclusions

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

First, we present a system for auto-stereoscopic display based on a color camera and TOF depth camera. Auto-stereoscopic display can render different view images from a single color image with its corresponding depth map. 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. A high resolution color camera is added beside the depth camera. The color camera and the depth camera are not located at the same position which means that viewpoints are slightly different. We apply view translation techniques to depth information and generate the corresponding depth map.

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.

## Acknowledgements

This research was supported by National Institute of Information and Communications Technology, Japan.

## References

A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” IEEE Transactions on Image Processing, vol. 13, no. 9, pp. 1200–1212, 2004.

N. A. Dodgson, “Autosteroscopic 3D Displays” , Computer, pp. 31-36, August, 2005.

C. Fehn, K. Sch¨u¨ur, I. Feldmann, P. Kauff, and A. Smolic, “Distribution of ATTEST test sequences for EE4 in MPEG 3DAV,” ISO/IEC JTC1/SC29/WG11, M9219 doc., Dec. 2002.

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, pp. 1055–1066, 1996.

L. Meesters, W. IJsselsteijn, and P. Seuntiens, “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.

T. Oggier, et al., “An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger)”, Proc. SPIE, Vol. 5249, 534 (2004).

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.

L. Zhang and W. Tam, “Stereoscopic image generation based on depth images for 3D TV,” IEEE Transactions on Broadcasting, vol. 51, no. 2, pp. 191–199, June 2005.