

Depth Estimation from Three Cameras Using Belief Propagation 3D Modelling of Sumo Wrestling

Kensuke Ikeya*, Kensuke Hisatomi*, Miwa Katayama*, and Yuichi Iwadate*

*NHK Science & Technology Research Laboratories 1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan

E-mail: {ikeya.k-ec, hisatomi.k-ko, katayama.m-gm, iwadate.y-ja}@nhk.or.jp

Abstract

We propose a method to estimate depth from three wide-baseline camera images using belief propagation. With this method, message propagation is restricted to reduce the effects of boundary overreach, and max and min values and kurtosis of message energy distribution are used to reduce errors caused by large occlusion and textureless areas. In experiments, we focused on scenes of the traditional Japanese sport of sumo and created 3D models from three HD images using our method. We displayed them on a 3D display using the principle of integral photography (IP). We confirmed from the experimental results that our method was effective for estimating depth.

Keywords: Belief propagation, Depth estimation, 3D model

1 Introduction

Recent progress in image-based scene reconstruction techniques has made it possible to generate high quality 3D models. We address the creation of 3D models of sports scenes, e.g., soccer, baseball, and gymnastics, and utilise it for 3DTV content based on the principal of integral photography (IP), which is a stereoscopic display system that reconstructs 3D auto-stereoscopic images in theory. The capture and reconstruction conditions below were assumed, following the conditions used in the production of actual TV programs.

- Multiple camera positions are far from objects.
- The baseline is wide to get disparity.
- The optical axis of multiple cameras cross on the objects.
- Whole objects in the image are reconstructed.

Our research focuses on achieving 3D modelling under these conditions. To create the 3D models, we use a well-known belief propagation algorithm [1] that has been utilised to solve a wide variety of multi-view reconstruction problems because of its strong optimality properties. However, 3D modelling is difficult when errors occur due to boundary overreach and the existence of large occluded and textureless areas. Boundary overreach is the phenomenon in which the boundary position of an object is incorrect due to inaccurate matching of object boundaries. Occlusion errors cause severe visual effects, and errors caused by textureless regions often involve background objects, e.g., walls and floors, and these also result in severe visual effects.

Many efficient belief propagation algorithms have been demonstrated. A good example of the capabilities of current

state-of-the-art algorithms is provided by the Middlebury dataset [2]. In particular, top-level algorithms using belief propagation [3][4][5] produce highly accurate results. However, experiments with these algorithms were conducted under conditions that were different from ours, e. g., using a single camera, narrow-baseline, parallel optical axes, and with target objects and scenes that were not typical for TV content. Errors caused by large occlusion and textureless areas are of little consequence under these conditions. Therefore, it is unclear whether these algorithms would be effective in addressing our target problems.

A number of approaches for reconstructing sports scenes have been proposed. For example, view-interpolation [6][7], the planar billboard technique [8], and 3D modelling using graph cuts [9] have achieved good results. However, none of these studies targeted 3D model reconstruction of entire objects, not only foreground objects but also background ones, in the image. These studies are therefore not suitable for our purpose.

We propose a technique to estimate depth from three wide-baseline camera images using belief propagation. Our method utilises message propagation characteristics and energy distribution to reduce estimation error caused by the effects of boundary overreach, large occlusion, and textureless areas. To reduce the effects of boundary overreach, we first apply the constraint that the message does not propagate between pixels with a large colour difference. Second, we employ max and min values of message energy distribution to remove occlusion area errors. Lastly, we use kurtosis in message energy distribution to reduce errors of textureless areas. We focus on scenes of the traditional Japanese sport of sumo and create a 3D model of it from three HD images. Furthermore, we display the 3D model on the 3D display based on the principal of IP to convert 3D models to an IP image.

We present our method in this paper. We first give an overview of the proposed method (Section 2). Then we describe the initial depth estimation which restricts message propagation in order to reduce errors caused by boundary overreach (Section 3). We then present a way to reduce errors caused by occlusion and textureless areas using message energy distribution (Sections 4 and 5). The experimental results consisting of depth estimation, 3D modelling of sumo scenes, and their display on 3DTV using integral photography are presented in Section 6. Finally, we highlight the results obtained using the proposed method and briefly touch on our future work.

2 Overview

Our method consists of the following steps, which are illustrated in Fig. 1. With this method, a 3D model is created from three HD captured images.

Step (a) Scene capture: Scenes are captured by multiple synchronised HD cameras separated by a wide baseline. All cameras are calibrated using feature points in the captured images.

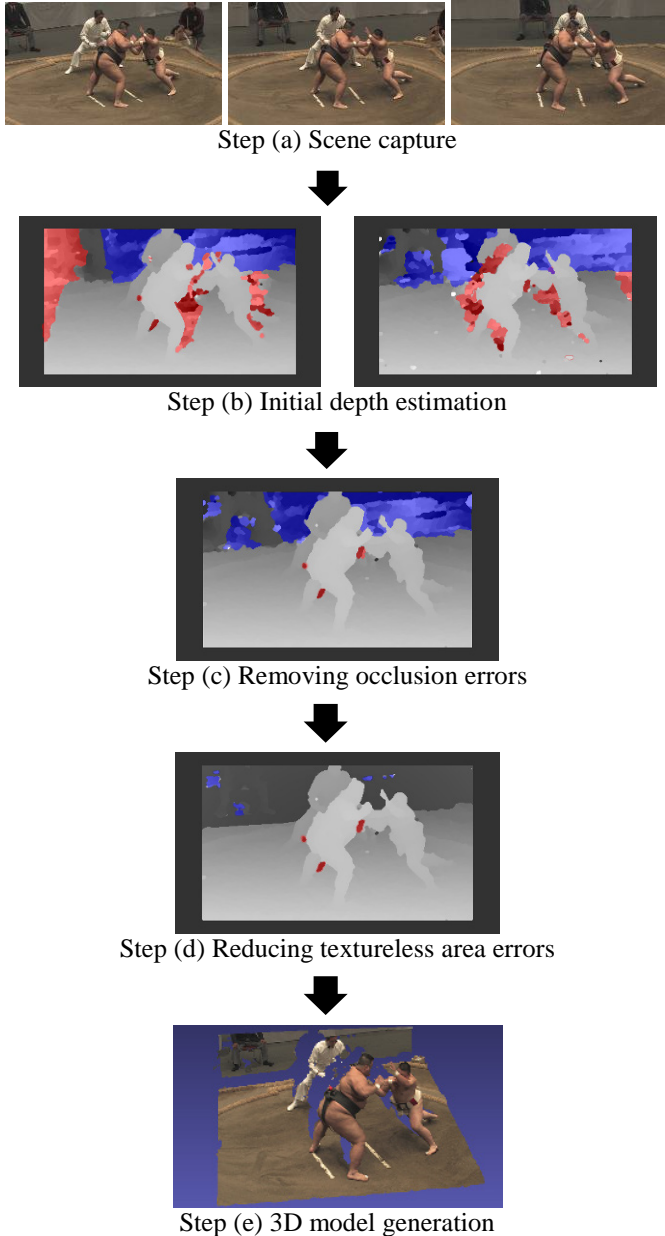


Figure 1: Work flow of proposed method:
Red areas are occlusion errors and blue areas are textureless area errors

Step (b) Initial depth estimation: The initial depth is estimated using belief propagation with three neighbouring camera images. We set up two image pairs, a left pair and a right pair, from three neighbouring camera images (Fig. 2). Initial depth maps have errors due to occlusion and textureless areas. Occlusion errors are indicated in red and textureless area errors are indicated in blue in Fig. 1. These areas are painted manually using Adobe Photoshop to make them easier to recognize.

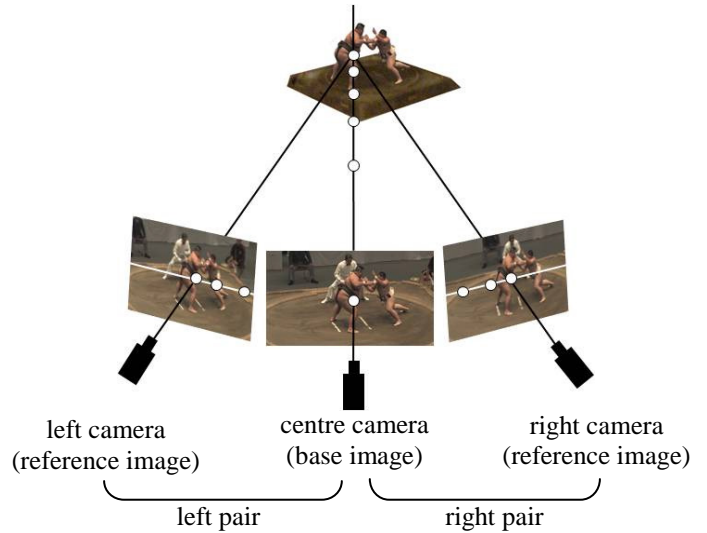


Figure 2: Initial depth estimation with three camera images

Step (c) Removing occlusion errors: Initial depth maps contain severely occluded areas because the objects are captured by wide-baseline cameras whose optical axes cross on an object. The initial depth maps are used to complement each other in order to remove occlusion errors in the base image.

Step (d) Reducing textureless area errors: Depth maps created using the above steps may contain errors due to textureless areas, especially in the background. In this step, areas with such errors are evaluated, and processing is applied to reduce the errors.

Step (e) 3D model generation: The 3D model data format is VRML and is generated from a depth map of the base image.

In this method, there are three novel points in the workflow. The first point is the message propagation restriction in step (b). The initial depth map contains errors caused by the effects of boundary overreach. Therefore, the message propagation is restricted depending on the colour difference between neighbouring pixels in order to reduce these errors. The second point is the use of message energy distribution in step (c). We employ max and min values in this distribution to determine which areas have occlusion errors in the initial depth maps for each camera pair, and we then remove these

errors by comparing them. The third point is the use of a kurtosis of message energy distribution to reduce errors caused by textureless areas in step (d). We determined which areas had textureless errors in the depth maps by using kurtosis of message distribution and carried out a process to reduce these errors. We describe below the details of the proposed method.

3 Initial depth estimation

We employ belief propagation [1] to estimate the initial depth with three camera images, and then restrict the message propagation to reduce the effects of boundary overreach.

We define the energy function of a message as follows:

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} V(f_p - f_q) \quad (1)$$

where P is the set of pixels in an image, p is the pixel, q is the neighbouring pixel of p , f is the depth label, and N are the edges in the four-connected image grid graph. The first term D is the data energy and is defined below:

$$D_p(f_p) = \begin{cases} \lambda_{data} \Delta_{f_p} & \text{if } \Delta_{f_p} \leq T_{data} \\ T_{data} & \text{if } \Delta_{f_p} > T_{data} \end{cases} \quad (2)$$

$$\Delta_{f_p} = \left(\sum_{c \in \{r,g,b\}} |I_c(p) - I'_c(p + d_p)| \right) / 3 \quad (3)$$

where c is the colour, I is the intensity of the base image, I' is one of the reference images, d is the disparity corresponding to f , λ_{data} is the weight value, and T_{data} is the limiter of the data term.

The second term V is the smooth energy and is defined as follows:

$$V(f_p - f_q) = \begin{cases} |f_p - f_q| & \text{if } |f_p - f_q| \leq T_{smooth} \\ T_{smooth} & \text{if } |f_p - f_q| > T_{smooth} \end{cases} \quad (4)$$

Here, T_{smooth} is the limiter of the smooth term. Three camera images (left, centre, right) are divided into a left pair (left and centre) and right pair (centre and right). The centre camera image is defined as the base image, and the other images are defined as reference images (Fig. 2). We first convert depth to disparity and calculate the data term and the smooth term using (2), (3) and (4) for each pair.

Then, we update the message and propagate it. The message update is defined below:

$$m_{p \rightarrow q}^t(f_p) = \begin{cases} \min_{f_p} \left(V(f_p - f_q) + D_p(f_p) + \sum_{s \in N(p) \setminus q} \lambda_{s,p} m_{s \rightarrow p}^{t-1}(f_p) \right) & \text{if } |I_c(p) - I_c(q)| \leq T_{message} \\ m_{p \rightarrow q}^{t-1}(f_p) & \text{if } |I_c(p) - I_c(q)| > T_{message} \end{cases} \quad (5)$$

$$\lambda_{s,p} = \begin{cases} 1 & \text{if } |I_c(s) - I_c(p)| \leq T_{message} \\ 0 & \text{if } |I_c(s) - I_c(p)| > T_{message} \end{cases} \quad (6)$$

where $m_{p \rightarrow q}^t$ is the message that pixel p sends to pixel q at iteration t . The term $m_{p \rightarrow q}^0$ is initialised to zero. The s is a pixel, and $N(p) \setminus q$ denotes the neighbours of p other than q . $T_{message}$ is a threshold, and $\lambda_{s,p}$ is the weight values of message propagation. After T iterations, a belief vector b is computed for each pixel,

$$b_q(f_q) = D_q(f_q) + \sum_{p \in N(q)} \lambda_{p,q} m_{p \rightarrow q}^T(f_q) \quad (7)$$

$$\lambda_{p,q} = \begin{cases} 1 & \text{if } |I_c(p) - I_c(q)| \leq T_{message} \\ 0 & \text{if } |I_c(p) - I_c(q)| > T_{message} \end{cases} \quad (8)$$

where $\lambda_{p,q}$ are the weight values of message propagation. Finally, the depth label f_p that minimises $b_q(f_q)$ individually at each pixel is selected.

In the proposed method, when the message is updated and propagated, the message flow is restricted depending on the difference in colour values between neighbouring pixels. Fig. 3 shows examples where the message flow is restricted by the colour difference between bright pixels and dark pixels. If colour differences between neighbouring pixels are lower than $T_{message}$, the message flow is not restricted (Fig. 3 (a)). If the colour difference between pixel p and pixel s is higher than $T_{message}$, the message from pixel s is not used in the message update (Fig. 3 (b)). Similarly, if the colour difference between pixel p and pixel q is higher than $T_{message}$, an updated message is not propagated (Fig. 3 (c)).

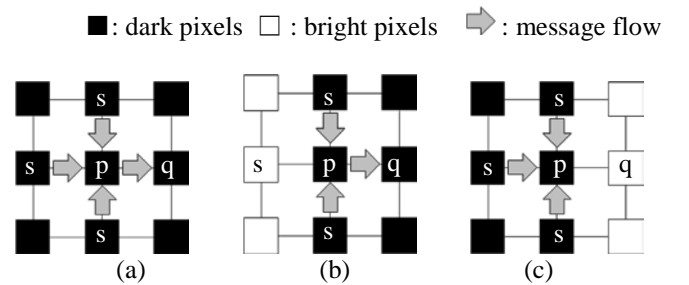


Figure 3: Example of message propagation restriction:
(a) when there is no colour difference between neighbouring pixels;
(b) when there is a colour difference between pixels s and p ;
(c) when there is a colour difference between pixels p and q .

4 Removing occlusion errors

The initial depth maps have occlusion errors. At this stage, we remove the errors from the depth maps using energy distribution.

We first calculate occlusion value O_q from the message energy distribution at each pixel. O_q is defined below:

$$O_q = (\max(b_q(f_q)) - \min(b_q(f_q))) / \max(b_q(f_q)) \quad (9)$$

O_q at the pixel in the occluded area is lower than that of other pixels in the area. Fig. 4 shows initial depth maps with occlusion errors indicated in red, and occlusion maps in which high value areas of O_q are represented by bright pixels and low value areas are represented by dark pixels. It is obvious that the red areas in the depth maps correspond to the dark pixels in the occlusion maps. O_q indicates occluded areas in the depth maps correctly. We compare O_q at each pixel between the left and right initial depth maps, and an initial depth of higher O_q pairs is assigned to the depth map.

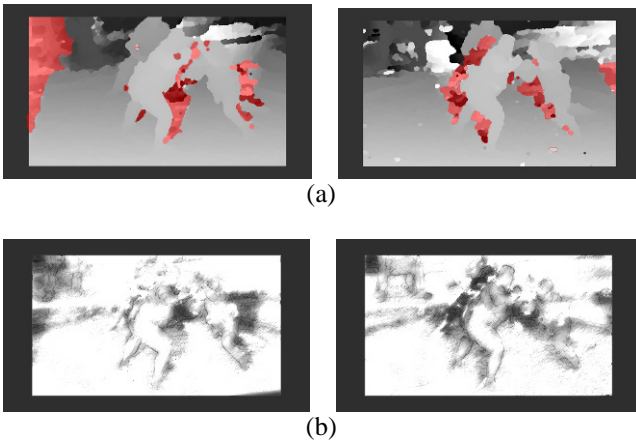


Figure 4: Initial depth map and occlusion value:
(a) initial depth map for left and right pairs. Occlusion errors are indicated by red areas;
(b) occlusion map. High value areas of O_q are represented by bright pixels, and low value areas are represented by dark pixels.

5 Reducing textureless area errors

The depth map created through the above steps includes errors of textureless areas, which in this case are the floor and walls with poor texture in the background. We employ kurtosis of message energy distribution to reduce errors in these areas.

Fig. 5 shows an example of message energy distribution at pixels assigned with a correct and an incorrect depth label due to textureless area errors. The horizontal axis shows depth labels at the pixel and the vertical axis shows the amount of message energy stored at each depth label. The energy

distribution at the pixel assigned the correct depth label is sharper than that at the pixel assigned the incorrect depth label. The sharpness of energy distribution is expressed by the kurtosis of energy distribution. Kurtosis K_q is defined below:

$$K_q = \frac{n_f(n_f+1)}{(n_f-1)(n_f-2)(n_f-3)} \sum_{f_q} \left(\frac{b_q(f_q) - \bar{b}_q}{\sigma(b_q)} \right)^4 - 3 \frac{(n_f-1)^2}{(n_f-2)(n_f-3)} \quad (10)$$

Here, n_f is the number of labels f . The more peaks there are in the energy distribution, the higher K_q becomes. We identify textureless areas by K_q , and reduce the errors in these areas.

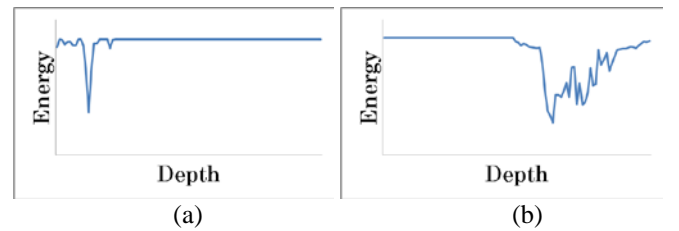


Figure 5: Example of message energy distribution:
(a) distribution at pixel assigned correct depth label;
(b) distribution at pixel assigned incorrect label

Fig. 6 shows the process in this step; (a) is the input image and (b) is the depth map that includes textureless area errors, which are indicated in blue. Fig. 6(c) shows the kurtosis map in which high value areas of K_q are represented by bright pixels and low value areas are represented by dark pixels. It is obvious that the kurtosis in the textureless areas is lower than that in the other areas. Then, we employ the pyramid segmentation algorithm implemented in Open CV to segment the base image with colour (Fig. 6(d)). We calculate the average kurtosis in each segmentation area and assign it to each area. We divide these segmented regions into two areas: a textureless area and the “other” area, by setting a threshold at the kurtosis average. Fig. 6(e) shows the two areas: textureless areas, which are indicated in black, and the other areas which are indicated in white.

Finally, we apply processing to the identified textureless area to reduce errors. Fig. 7 depicts the processing. We set a sphere in 3D space, the centre of which is the cross point of the optical axes of multi-view cameras. The radius of the sphere is set to include all objects inside the sphere. We calculate the cross point of the base camera’s optical axis and the sphere surface. Depth in the textureless areas is computed from the distance between the base camera position and the cross point. Fig. 6(f) shows a depth map in which calculated depth is assigned to the textureless area.

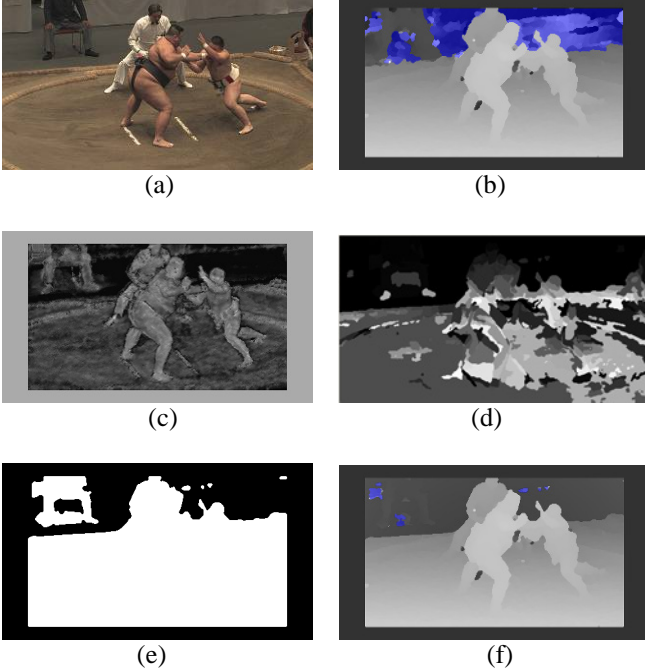


Figure 6: Overview of reducing textureless area errors: (a) input image; (b) depth map containing textureless area errors; (c) kurtosis map of message energy distribution (low kurtosis areas are represented by dark pixels; high kurtosis areas are represented by bright pixels); (d) colour segmentation; (e) textureless areas and other areas (textureless areas are represented by dark pixels; other areas are represented by bright pixels); (f) depth map showing reduced textureless area errors

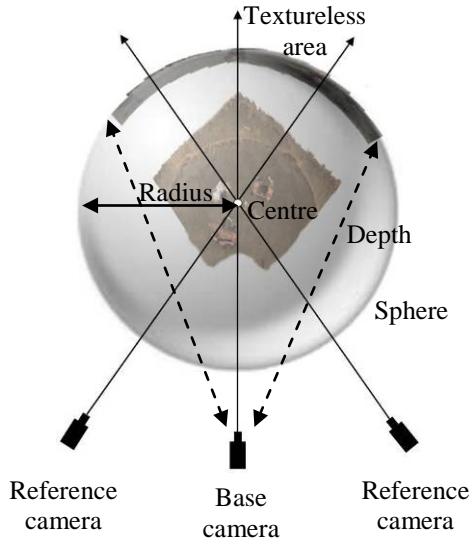


Figure 7: Assignment of depths to textureless areas

6 Experimental results

We evaluated our method by conducting the following experiments. First, we evaluated the accuracy of our method using the Middlebury dataset [2]. We compared depth maps obtained using our method and another method called the sum of squared differences (SSD). Next, we evaluated the proposed method using a multiple-view sequence of sumo, which is a popular traditional sport in Japan. We generated 3D models from these estimated depth maps. Lastly, we displayed a 3D model of the sumo scene on a 3Ddisplay using IP to convert the 3D model to an IP image.

6.1 Depth estimation using the test data set

We evaluated our method using Middlebury datasets. For this evaluation, we compared the depth map estimated by the SSD method and the proposed method. We used three images: Aloe, Dolls, and Bowling 2, and we calibrated the cameras using feature points in the images before estimating the depth. The SSD block size was 3 x 3 pixels. The parameters λ_{data} , T_{data} , T_{smooth} , $T_{s,p}$, $T_{p,q}$ and T , which were used for the belief propagation, were set to 0.07, 15, 1.7, 64, 64 and 10, respectively. To compare the results fairly, three cameras were used with the SSD method to reduce occlusion errors, similarly to our method.

In the SSD method, depth was estimated at each camera pair as shown in Fig. 2 and the error energy of the assigned depth was compared at each pixel. If an error energy value was twice as high as the other one, the depth of the smaller error energy was assigned. Otherwise, the average depth at each camera pair was assigned. Fig. 8 shows the input image and the depth maps obtained by SSD and the proposed method. In addition, we introduced the ground truth images in Fig. 8 for reference; these images have a disparity label.

Fig. 8 shows that the proposed method is effective for estimating depth in several types of images. The depth maps obtained by SSD contain a lot of fine noise. The proposed method can inhibit these errors and estimate smooth depth. The boundary position between the foreground and background in the depth maps estimated using the proposed method nearly corresponds to the boundary in the ground truth images.

6.2 Capture of sumo scene by multi-view cameras

The sumo sequence was captured by 11 HD cameras surrounding one corner of the sumo ring, as shown in Figs. 9 and 10. The sequence was obtained for use in the multi-view HDTV system [10], a video production system. The cameras were set in the front row on the first floor, about 25 m away from the centre of the sumo ring; covering about 90 degrees of the ring. Therefore, the angle between three cameras was about 18 degrees and the actual distance between the cameras

was approximately 2 m. They were synchronised using the reference signal provided by a signal generator and calibrated using the feature points extracted from the images. The depth map was estimated using 3 cameras, side by side, which were selected from the 11 cameras.



Figure 9: Sumo arena

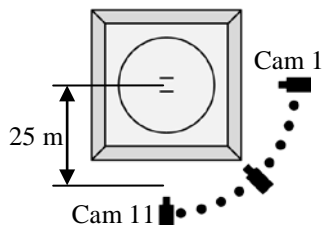


Figure 10: Camera positions

6.3 3D modelling

We compared the results of depth estimation using three different methods, i.e. SSD, standard belief propagation (standard BP) [1], and the proposed method. The SSD and standard BP methods used three cameras to reduce occlusion errors similarly to our method. The SSD block size, the parameters used for the belief propagation, and the way of assigning depth in the SSD method were the same as that in the experiment in Section 5.1. In standard BP, we changed the SSD block size to 1 x 1 pixels and utilized its error energy for data energy. Figure 11 shows the depth estimation results using Cam 2. The luminance of the depth images was normalised and it differed between the different sequences. Then, we generated 3D models from depth maps estimated using standard BP and the proposed method and compared the accuracy of depth estimation. Figure 12 shows the 3D models estimated using cam 2.

Fig. 11 shows that the estimation error obtained with our method is much lower than in the other depth maps. A lot of salt and pepper noise is observed, and the estimation fails at the border of the foreground and background in the SSD results. These errors are restrained in the results using standard BP and the proposed method. Both methods using belief propagation were successful in capturing the ground of the sumo ring, which is presented as a smooth gradation pattern in the depth map. The standard BP, however, failed to remove the occlusion errors indicated in red, while the proposed method succeeded in doing this. The accuracy of estimation in the textureless areas, such as the one indicated in blue in the base image in Fig. 12, was also improved using the proposed method. Fig. 12 shows that estimation errors in the textureless areas using the standard BP cause visually intense noise when the 3D model is projected onto the desired viewpoint. However, the proposed method prevented these errors. The boundary position between the back of the sumo wrestler and the referee is not in the right boundary position due to the effects of boundary overreach (Fig. 12 first row, third column). This error is prevented and the accuracy at the

border of the foreground and background was also improved using the proposal method.

6.4 Display 3D models on 3D display based on IP

We converted the 3D model to an integral image and displayed it on an integral 3D display. This system consists of a high-resolution liquid-crystal panel and a lens array. It enables users to obtain a perspective view of 3D auto-stereoscopic images from any direction. Fig. 13 shows our IP display, which is a 4Kx2K liquid-crystal panel display with 160x118 lens arrays. We used an algorithm [11] to convert the 3D model to an integral image.

Fig. 14 shows the images on the display captured by a digital still camera from different viewpoints, i.e., a high angle, low angle, left angle, and right angle. In the experimental results, we confirmed that the subject displayed on 3DTV had a high depth sense. The sumo wrestlers' muscles are expressed with high presence, and the sumo ring continues from front to back smoothly on 3DTV. We can also perceive motion parallax by the different movement of the fore- and backgrounds.



Figure 13: 3D display using integral photography

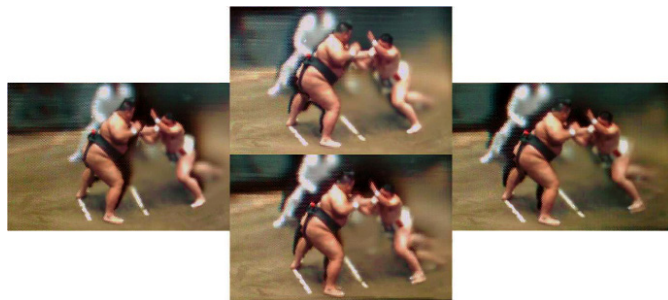


Figure 14: View on 3D display using IP

7 Conclusion

We presented a technique to estimate depth from multi-view images using belief propagation. The method enabled us to reduce estimation errors caused by the effects of boundary overreach, occlusion, and textureless areas. We focused on scenes of the traditional Japanese sport of sumo and created 3D models of the scenes and displayed them on 3DTV using IP. In the future, we will create 3D models of various sport scenes and improve the accuracy of the 3D modelling algorithm.

Acknowledgements

The author gratefully acknowledges the support of the National Institute of Information and Communications Technology.

References

- [1] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. CVPR, pages I: 261–268, 2004.
- [2] D. Scharstein and R. Szeliski, Middlebury Stereo Vision Research Page, <http://vision.middlebury.edu/stereo/eval/>, 2011.
- [3] A. Klaus, M. Sormann and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. ICPR, page 15-18, 2006.
- [4] Z. Wang and Z. Zheng. A region based stereo matching algorithm using cooperative optimization. CVPR, page 1-9, 2008.
- [5] Q. Yang, L. Wang, R. Yang, H. Stewénus, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. PAMI, 31(3):492–504, 2009.
- [6] K. Kimura and H. Saito. Player viewpoint video synthesis using multiple cameras. CVMP, page 112–121, 2005.
- [7] N. Inamoto and H. Saito. Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. IEEE Transactions on Multimedia, 9(6), 1155–1166, 2007
- [8] Y. Ohta, I. Kitahara, Y. Kameda, H. Ishikawa, T. Koyama. Live 3D video in soccer stadium. IJCV, 75(1), 173–187, 2007.
- [9] J-Y. Guillemaut and A. Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications, IJCV, page 1-28, 2010.
- [10] K. Tomiyama, K. Hisatomi, M. Katayama, and Y. Iwadate. Advanced video image technologies for live sports TV productions. Proceedings of NAB, page 193-198, 2008
- [11] M. Katayama and Y. Iwadate. A method for converting three-dimensional models into auto-stereoscopic images based on integral photography. Proceedings of SPIE, 6805, 2008.