Billboard Deformation via 3D Voxel by Using Optimization for Free-viewpoint System

Keisuke Nonaka, Qiang Yao, Houari Sabirin, Jun Chen, Hiroshi Sankoh and Sei Naito
KDDI Research, Inc.
Ohara 2-1-15, Fujimino, Saitama, Japan

***Abstract*—A free viewpoint application has been developed that yields an immersive user experience. The free viewpoint approach called the "billboard methodis" suitable for displaying a synthesized 3D view in a mobile device, but it suffers from the limitation that a billboard cannot present an accurate impression of depth for a foreground object, and it gives users an unacceptable impression from certain virtual viewpoints. To solve this problem, we propose the optimal deformation of the billboard. The deformation is designed as a mapping of grid points in the input billboard silhouette to produce an optimal silhouette from an accurate voxel model of the object. We formulate and solve this procedure as a nonlinear optimization problem based on a grid-point constraint and some a priori information. Our results show that the optimal deformation is produced by the proposed method, which generates a synthesized virtual image having a natural appearance.**

I. Introduction

In recent years, free viewpoint navigation systems from a multi camera environment [1], [2] have been among the hottest topics in computer vision. In the navigation system, users can select their viewpoint freely (not limited to actual camera positions), and the selected scene is synthesized by using multi camera videos and several additional items of information. This makes the free viewpoint system very useful to improve user understanding of the scene, and creates an immersive and ultra-realistic user experience, especially when viewing sports. Naturally, the demand for free viewpoint in mobile devices has also increased and this means contents needs to be streamed using a wireless network.

Some conventional famous free-viewpoint synthesis approaches, for example, the "visual hull method [3]–[6]", generate a highly accurate 3D voxel model of the scene by using information from multiple cameras. However, with these methods, viewing the model in a mobile device incurs a high computational cost and they are not suitable for streaming because they have too much fine detail to display and the amount of data is generally huge. In addition, these methods also have the drawback that they require numerous cameras and highly accurate camera parameters because the core idea is utilizing intersection and blending of information from multiple cameras. Even if this was not the case, the synthesized model still might have some undesirable artifacts like a missing part or blurred texture of the target model.

On the other hand, methods based on simple 3D models like the "billboard model method [7]–[9]", have been proposed. The method extracts the texture of a foreground object and uses it as a billboard model (a simple 3D plane), placing the plane model perpendicularly on a 3D ground and rotating it to face towards virtual viewpoints. From this simple framework, the method has the advantage reducing the computational cost of display and the amount of data while keeping quality of the model texture. Furthermore, since the billboard is independently represented on each camera, the method has some robustness in terms of combining multi camera settings. Conversely, the method has difficulty to detect and represent where the billboard model is in 3D space. To be specific, if we can calculate a contact point between an object and the ground in an image, we can find actual position by projecting the point to 3D ground and the texture of object is copied as the model at the position in free viewpoint synthesis. Normally speaking, the contact point is calculated along the bottom line of a bounding box that is used to represent the position information of the object in an image, because we assume the object touches on the ground at a lowest point like feet even in the image. However, in some case where an object (a player) in an image touches the ground at multiple points (e.g. hands and feet) as shown in Fig.1 (a), there will be an ambiguity to calculate the contact point because the bottom line of the bounding box cannot be uniquely determined and it is difficult to represent multiple touch points only by one extracted texture. Consequently, as shown in Fig.1 (b), a synthesized virtual scene in the vertical direction shows unnatural appearance as if the player were flying in the air. We call this the "touch points problem" and conventional billboard methods can neither detect nor solve it.

The "Microfacet-billboarding method [10], [11]" has been proposed as a hybrid of the two methods described above. This method represents a person in the foreground as an aggregate of small billboards that is constructed from a finely subdivided image of the person and his actual position by using the visual hull technique and voxels. This can solve the unnatural touch points problem affecting the billboard model method. Nevertheless, microfacet-billboarding causes undesirable artifacts along the boundaries in the synthesized model due to the large number of small rectangles that comprise it.

From the above discussion, we propose a hybrid billboard method that uses 3D voxel as support information to solve the touch points problem, preserving the advantages of the billboard. Our method is designed to generate a standard (single plane) billboard model, but appropriately deformed for different virtual viewpoints. The deformation is represented as

ISBN 978-0-9928626-7-1

(a) (b)

Fig. 1. Example of our target "touch points problem" in the conventional billboard method: (a) natural synthesis from the camera viewpoint, (b) unnatural synthesis from a lower virtual viewpoint (the input is shown as Fig.4 Scene 1 (a)).

a mapping of grid points that are obtained by separating the elements of the initial foreground texture. We determine the optimal mapping by using the silhouette of the voxels because we assume that the voxel model can generate the optimal human silhouette from any virtual viewpoint. The optimal mapping of grid points can be obtained using a conventional constrained nonlinear optimization tool, the projected gradient method [12], and the billboard will then be generated according to the points using texture mapping. Our experimental results show that our optimal deformation can generate a billboard representation having a natural appearance and also solve the touch points problem.

II. Proposed method

A. Overview

We assume that the human target (player) is surrounded by K cameras (in our experiments, $K = 8$), each camera position is fixed and the camera parameter should be known. The time synchronization between all cameras is adjusted in advance. Throughout this paper, we discuss how to synthesize a billboard from a virtual viewpoint in only one frame (the time t), but our method can easily be applied to all video sequences because the procedure is independent frame by frame. To simplify the problem, we assume a single target player, but our algorithm can easily be extended to deal with multiple players individually.

The conventional billboard method [7]–[9] is composed of three parts, "silhouette mask extraction", "billboard information creation" and "synthesizing (locating the billboard on the 3D ground in viewer)". Let S be an input image and M be the extracted binary silhouette mask of the player (camera silhouette) obtained by "silhouette mask extraction", and the method simply constructs a billboard (plane) from the texture of the nearest camera image extracted by M. This fact means viewer needs to know only the S and M to generate billboard. Then the view of the billboard model is synthesized by putting the billboard on its 3D ground according to the 3D position that estimated to be the projection of the bottom point of the billboard.

Our method is fundamentally based on the conventional billboard scheme, with the key difference being the "billboard information creation" part. A flowchart of our algorithm is shown in Fig. 2. First, we obtain camera silhouettes in multiple cameras using the same approach as the conventional billboard method. Then, we can generate a voxel model via the conventional visual hull method [3]. When we select a virtual viewpoint, we can get the optimal silhouette of the player M_{opt} by projecting the voxels onto the camera image plane and choose the silhouette of the nearest camera as a target camera silhouette M . In our "billboard information creation" part, we consider M as a set of small regions (patches) separated by a constant interval grid. After obtaining M_{opt} and the grid, the correspondences between M and M_{opt} are calculated by block matching each patch on the edge of the silhouette. Using these correspondences and several a priori assumptions, we can formulate and solve the optimization problem related to the coordinate mapping of the patches. In our viewer, to generate a billboard, we only have to know the mapping, camera image and silhouette M. According to this flow, the key task is to find the mapping of the coordinates that correspond to a selected virtual viewpoint. In this paper, we discuss a case in which the selected viewpoint differs from the camera position only on the vertical axis, because the touch points problem arises only if the viewpoint moves vertically. If the viewpoint moves horizontally, our method simply rotates the generated billboard, just as the conventional method does.

B. Voxel Model Construction and Optimal Silhouette Projection

To find the optimal mapping of patches, we refer to the optimal silhouette M_{opt} generated by the voxel model. The voxel model provides a 3D outline of the player, and is constructed by the visual hull technique using the intersection of the mask of the player in voxel (3D) space [4]. Obtaining the 3D model, we can find the actual shape of the player from any virtual viewpoint. Therefore, by projecting the model back onto the camera image plane, we can get the optimal binary silhouette M_{opt} of the player from a virtual viewpoint (Fig. 3 (b)).

C. Initial Grid Definition and Edge Patch Matching

As a preliminary, we define a patch $\boldsymbol{p}_i = \{\boldsymbol{p}_{i,1}, \boldsymbol{p}_{i,2}, \boldsymbol{p}_{i,3}, \boldsymbol{p}_{i,4}\}^{\mathrm{T}} \in \mathbb{R}^8$ that is the small region in the camera silhouette M, where the vector $\boldsymbol{p}_{i,j}$ is a corner point of $\boldsymbol{p}_i$ and has the coordinates $(x_{i,j}, y_{i,j})$ on the image axes. We can obtain a set of coordinates for the patch $\boldsymbol{p} = \{\boldsymbol{p}_1^{\mathrm{T}}, \boldsymbol{p}_2^{\mathrm{T}}, \ldots, \boldsymbol{p}_m^{\mathrm{T}}\}^{\mathrm{T}} \in \mathbb{R}^{8m}$ by separating the M at periodic intervals on the grid (see Fig.3 (a)), where m is the number of patches in a player. Please note that some $\boldsymbol{p}_i$ share the same corner point because of the structure of the grid, and we call such a set of points a "common vertex set".

Next, we calculate a rough correspondence between M and M_{opt}. Both are specific player regions in the binary image corresponding to each viewpoint. Since the color image is too sensitive to permit finding the correspondence of each patch by block matching, and the voxel model might generate a blurred color image even if the cameras deliver high resolution texture, we use only patches on the edge of each binary camera silhouette (edge patch).

ISBN 978-0-9928626-7-1 © EURASIP 2017

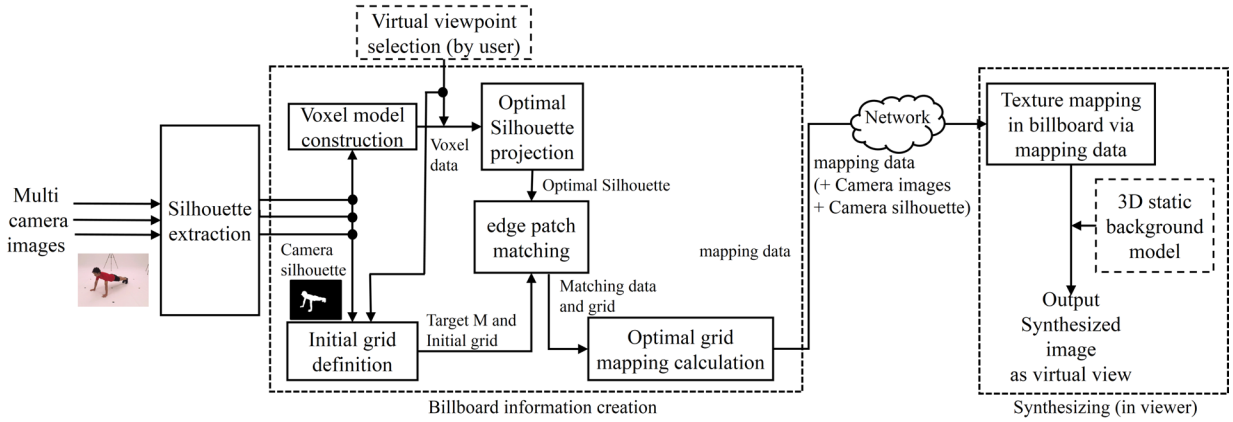

Fig. 2. Flowchart of proposed method.

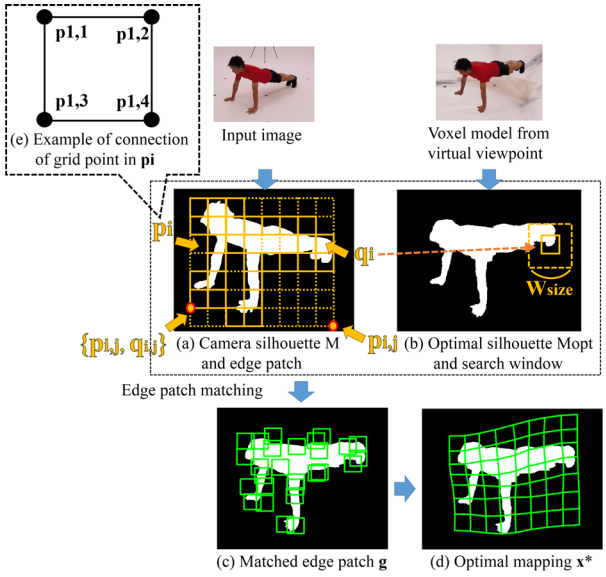

Fig. 3. Edge patch matching to the optimal silhouette and optimal mapping of grid.

An edge patch is defined by the proportion of silhouette pixels in the patch, as in

$$\tau_i = \sum_{p_{pix} \in M(\boldsymbol{p}_i)} p_{pix}/N(\boldsymbol{p}_i), \tag{1}$$

where the functions $M()$ and $N()$ return a set of binary pixel values and the number of pixels in each image patch, respectively. If $th_1 < \tau_i < th_2$, we define the patch $\boldsymbol{p}_i$ as an edge patch $\boldsymbol{q}_i$.

Furthermore, we find matched edge patches for the optimal silhouette by block matching as follows:

$$\boldsymbol{g}_i^* \in \underset{\boldsymbol{g}_i \in W_{M_{opt}}(\boldsymbol{q}_i)}{\arg\min} MAD(M_{opt}(\boldsymbol{g}_i), M(\boldsymbol{q}_i)), \tag{2}$$

where the function $W_{M_{opt}}()$ returns a set of patches in M_{opt} included in a search window which center is $\boldsymbol{q}_i$'s center position (the window size is defined as W_{size}), and the size of returned patch $\boldsymbol{g_i}$ is the same as $\boldsymbol{q}_i$. Therefore the $\boldsymbol{g}_i$ means a patch which should be compared with $\boldsymbol{q}_i$. In addition, the function calculate Mean Absolute Difference (MAD) between $M_{opt}(\boldsymbol{g}_i)$ and $M(\boldsymbol{q}_i)$ by block matching (the $M_{opt}()$ returns binary pixel value in the same manner as $M()$). Finally, we obtain a set $\boldsymbol{g} = \{\boldsymbol{g}_1^*, \boldsymbol{g}_2^*, \ldots, \boldsymbol{g}_k^* : k$ is the number of edge patches $\boldsymbol{q}_i\}$, as the matched edge patches in M_{opt}. It should be noted that the corner point in a common vertex set has different coordinates in this state, as in Fig.3 (c).

D. Optimal Grid Mapping Calculation

To find the optimal mapping of $\boldsymbol{p}$, we define a constrained energy function as follows:

$$\boldsymbol{x}^* \in \underset{\boldsymbol{x} \in \mathbb{R}^{8m}}{\arg\min} \frac{\lambda_1}{2} \|E\boldsymbol{x} - E\boldsymbol{p}\|_2^2 + \frac{\lambda_2}{2} \|H\boldsymbol{x} - \boldsymbol{g}\|_2^2 \text{ s.t. } \boldsymbol{x} \in C, \tag{3}$$

where the function $\|\boldsymbol{x}\|_2$ is L_2 *norm* for a vector $\boldsymbol{x} \in \mathbb{R}^n$.

The first term is a "structure coherence term" that works to preserve the entire structure of $\boldsymbol{p}$, and the matrix E is designed to calculate the distance between the corner points of the $\boldsymbol{p}$. For example, if the $\boldsymbol{p}$ has only one patch $\boldsymbol{p}_1 = \{\boldsymbol{p}_{1,1}, \boldsymbol{p}_{1,2}, \boldsymbol{p}_{1,3}, \boldsymbol{p}_{1,4}\}^{\mathrm{T}}$ and these points are connected as Fig.3 (e), the matrix E is designed as follows:

$$E = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \tag{4}$$

Please note this example describes only one (x) dimension of $\boldsymbol{p}_1$ for the sake of brevity, but the matrix is simply extended in the case of two dimensions (x, y). The second term is a "edge patch term", and the matrix H just extracts the coordinates of $\boldsymbol{x}$ corresponding to the edge patches $\boldsymbol{g}$. In other words, the matrix H shorten the length of $\boldsymbol{x}$ from $8m$ to $8k$ ignoring non-edge patch coordinates in $\boldsymbol{x}$. This term has the effect of making the camera silhouette similar to the optimal one. The set C is a set of vectors $\boldsymbol{x}$ in which the corner points in a common vertex

ISBN 978-0-9928626-7-1 © EURASIP 2017

set have the same coordinates, so this constraint guarantees connectivity between patches.

To solve Eq.(3), we use a conventional constrained nonlinear optimization tool, the projected gradient method [12]. The procedure is described in Algorithm 1. Since $\mathrm{proj}_C \boldsymbol{x}$ is a projection onto the closed convex set C from $\boldsymbol{x}$, it is easily calculated by averaging the coordinates of vertices that occupy the same position in the initial state. Clearly, the averaged vertices also belong to the common vertex set. In this paper, we continue to iterate the procedure until $|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t| < 0.1$. Deformed silhouette and texture are acquired by using texture mapping according to $\boldsymbol{x}^*$ (Fig.3 (d)).

Algorithm 1 Specialized projected gradient method for solving Eq.(3)

1: Set $t = 0$, and choose $\gamma > 0, \boldsymbol{x}_0 \in \mathbb{R}^{8m}$.
2: **repeat**
3: $\quad \boldsymbol{x}_{t+1} = \mathrm{proj}_C(\boldsymbol{x}_t - \gamma(\lambda_1 E^{\mathrm{T}} E(\boldsymbol{x} - \boldsymbol{p}) + \lambda_2 H^{\mathrm{T}}(H\boldsymbol{x} - \boldsymbol{g})))$
4: $\quad t \leftarrow t + 1$
5: **until** some stopping criterion is satisfied.

All the optimal mappings can be calculated before transmission by network in advance (Fig. 2). If it is difficult to calculate the mapping from all viewpoints, it can be obtained by linear interpolation from discrete sample mapping. Therefore, the computational cost to display these billboards is almost the same as in the conventional billboard method because the only added overhead is the cost of texture mapping each patch. Moreover, the total amount of data in the generated result is also close to that in the conventional billboard method because the only additional information is the mapping data of each grid point from all viewpoints.

III. Experiments

We examined the performance of the proposed method by comparing it to the conventional billboard method [9]. The results are shown in Fig. 4. The background images without the subject camera image are synthesized from the same virtual viewpoint, and the ground truth is given by using the standard visual hull method [4]. In this experiment, we use Full HD input images and the size of the patch is 50×50 [pixel]. The parameters of our method are defined as $th_1 = 0.5, th_2 = 0.95, \mu = 1.0, \lambda_1 = 0.5, \lambda_2 = 0.5, W_{size} = 50$ [pixel]. Our method can generate views that are better able to represent the ground truth than the conventional method, confirming that our proposed method solves the "touch points problem". For example, in scene 1-1, the conventional method generates a view that makes it appear as if he is standing only his hands (his feet are over horizontal line), but our method can deform the shape of the man in a way that makes it similar to the outline of the ground truth and his feet in our synthesis appear to be under the horizontal line. However, the model can still be considered a billboard. In scene 2, the shape of women's head is slightly different from the ground truth. This is because the 3D voxel is represented only by 2D information and the balancing of the weights, λ_x. Even in this case, compared with the conventional method, it seems our method can preserve the outline of the ground truth and solve the touch points problem.

To examine the similarity between the silhouette of the ground truth and the product of each method objectively, we define the similarity κ by

$$\kappa = \sum_{(x,y)\in\{X \cup M_{opt}\}} 1 - \frac{|X(x,y) - M_{opt}(x,y)|}{N(X \cup M_{opt})}, \qquad (5)$$

where the matrix X is an input binary silhouette of the comparison, namely, the camera silhouette produced by the conventional method and the silhouette output of the proposed method, and the (x, y) returns pixel value in each image region (if the pixel is not detected, the value defined as 0). In the case where the κ is 1, the input silhouette is identical to the optimal one. The score κ is shown in Table I, which shows that the silhouette generated by our method is similar to the ground truth, according to an objective evaluation. Furthermore, we examine the quality of the synthesized image. The most important feature of our method is its ability to preserve both the player's shape and structure in the synthesis process. For the sake of verifying this feature, we use a well-known structure evaluation method related to human perception, Mean Structured SIMilarity (MSSIM) [13]. The input synthesized image for this evaluation is also shown in Fig. 4 and we calculate MSSIM as for luminance between the ground truth (the voxel synthesis) and each method. A higher value is better in MSSIM, and the range of that is from 0 to 1. The results of MSSIM are also shown in Table I and the score shows our method can preserve not only the silhouette of the player but also the structure of the synthesized image.

TABLE I
COMPARISON BY OBJECTIVE SCORE

	Silhouette similarity κ		MSSIM	
	[9]	Proposed	[9]	Proposed
Scene 1-1	0.81	**0.95**	0.69	**0.84**
Scene 1-2	0.86	**0.95**	0.68	**0.84**
Scene 2	0.90	**0.95**	0.75	**0.88**

Since our method is dependent on the accuracy of the voxel model, we show result of ours in a case where the generation of the voxel fails as shown Fig. 5. Being different from Fig. 4, the voxel result is generated by a mask threshold value that is too high in several cameras (except the nearest one from a virtual viewpoint) and it causes part of the left hand to be missing. Even in a case like this one, our method is not affected by the accuracy of other cameras and can recover the missing part and improve synthesis. This means our method is robust against some voxel artifacts, because we use a simple silhouette feature, a connected grid and formulate a structure coherence term in our optimization.

IV. Conclusion

We have proposed an optimal billboard deformation method to overcome the "touch points problem" that arises in the

ISBN 978-0-9928626-7-1 © EURASIP 2017

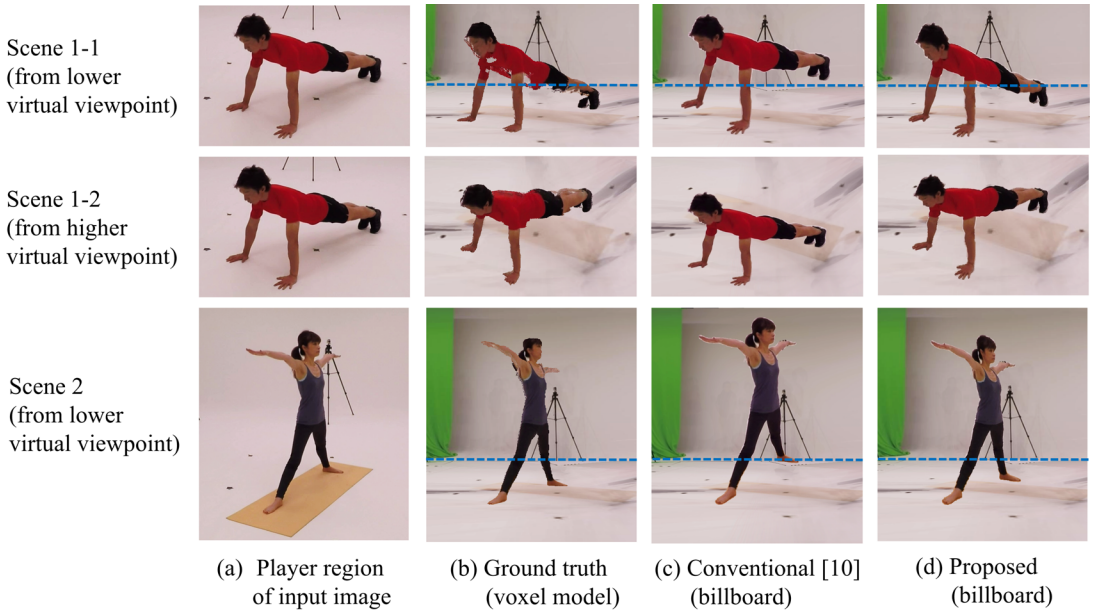

Fig. 4. Comparisons with conventional billboard method (The overlapping blue line represents horizontal line).

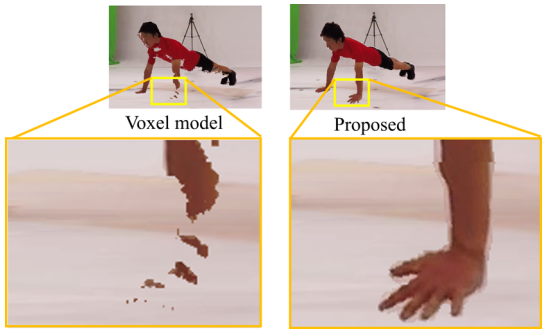

Fig. 5. Example of robustness of proposed method.

conventional billboard method. The key point of the proposed method is that a rough silhouette correspondence between the input and the optimal modification can be calculated by block matching of the patches on the edge of the silhouette. Additionally, based on the correspondence and some a priori information, we have formulated the constrained nonlinear optimization problem and solved it to generate a suitably deformed texture. Our experiments show that the proposed method can generate a silhouette similar to that from the voxel model, but is more useful as it solves the touch points problem in a way that preserves the structure while retaining the advantages of billboard.

We plan to investigate techniques for adaptive deformation that preserve specific features of the player, for instance by preserving faces, the goal being to improve the subjective quality of the deformed texture and to accelerate solving the nonlinear problem so that the method can find wide commercial use.

REFERENCES

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint tv," *IEEE Signal Processing Magazine*, vol. 28, no. 1, 2011.
[2] T. Kanade, P. Rander, and P.J. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE MultiMedia*, vol. 4, no. 1, pp. 34–47, 1997.
[3] C. Buehler, W. Matusik, L. McMillan, and S. Gortler, "Creating and rendering image-based visual hulls," *Tech. Rep.*, 1999.
[4] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *Proc. of the 27th Annu. Conf. on Computer Graphics and Interactive Techniques*, 2000, pp. 369–374.
[5] A. Ishikawa, M. P. Tehrani, S. Naito, S. Sakazawa, and A. Koike, "Free viewpoint video generation for walk-through experience using image-based rendering," in *Proc. of the 16th ACM Int. Conf. on Multimedia*, 2008, pp. 1007–1008.
[6] V. Chari, A. Agrawal, Y. Taguchi, and S. Ramalingam, "Convex bricks: A new primitive for visual hull modeling and reconstruction," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, 2012, pp. 770–777.
[7] K. Hayashi and H. Saito, "Synthesizing free-viewpoing images from multiple view videos in soccer stadiumadium," in *Proc. of Int. Conf. on Computer Graphics, Imaging and Visualisation*, 2006, pp. 220–225.
[8] T. Shin, N. Kasuya, I. Kitahara, Y. Kameda, and Y. Ohta, "A comparison between two 3d free-viewpoint generation methods: Player-billboard and 3d reconstruction," in *2010 3DTV-Conference*, 2010, pp. 1–4.
[9] Y. Yoshida and T. Kawamoto, "[demo] displaying free-viewpoint video with user controlable head mounted display demo," in *Proc. of IEEE Int. Symp. on Mixed and Augmented Reality*, 2014, pp. 389–390.
[10] S. Yamazaki, R. Sagawa, H. Kawasaki, K. Ikeuchi, and M. Sakauchi, "Microfacet billboarding," in *Proc. of the 13th Eurographics Workshop on Rendering*, 2002, pp. 169–180.
[11] N. Orman, H. Kim, R. Sakamoto, T. Toriyama, K. Kogure, and R. Lindeman, "Gpu-based optimization of a free-viewpoint video system," in *Proc. of the 2008 Symp. on Interactive 3D Graphics and Games*, 2008, p. 15:1.
[12] J. Jorge P. H. Calamai and More, "Projected gradient methods for linearly constrained problems," *Math. Program.*, vol. 39, no. 1, pp. 93–116, 1987.
[13] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

ISBN 978-0-9928626-7-1 © EURASIP 2017