EasyChair Preprint

№ 6674

Lively Free-Viewpoint Video Generation Using Video Textures

Fumiya Kimura, Hidehiko Shishido and Itaru Kitahara

September 23, 2021

LIVELY FREE-VIEWPOINT VIDEO GENERATION USING VIDEO TEXTURES

Fumiya Kimura[†] Hidehiko Shishido[†] Itaru Kitahara[†]
[†]University of Tsukuba [†]University of Tsukuba [†]University of Tsukuba

ABSTRACT

This paper introduces a method for generating free-viewpoint video that can express a sense of liveliness. One problem with conventional free-viewpoint video is that expressing such a sense of liveliness is difficult because the subject is depicted statically. We focus on "motion" as a factor for expressing this sense. Our proposed method adds a sense of liveliness to a static 3D model by expressing small movements of the body surface, such as facial expressions. The 3D model is generated from multi-viewpoint images, followed by projecting video textures to the model with randomness in the motion.

Keywords: Free-Viewpoint Video, Structure from Motion, Video Textures, Facial Expressions, Sense of Liveliness

1. INTRODUCTION

Free-viewpoint video [1][2][3], which reproduces views from arbitrary viewpoints by integrating multi-viewpoint images, is attracting attention for achieving advanced visual media. One problem with conventional free-viewpoint video is that it struggles to express a sense of liveliness because the subject is observed as a static object. We focus on "motion" as a factor for expressing liveliness. This paper focuses on portrait images, where the expression of liveliness is important.

The motion of a subject is classified into large movements caused by skeletal changes and such small movements as facial expressions. For expressing large movements, methods have been developed based on 3D motion capture, where the model's posture is modified based on the measurement of the subject's motion. For example, Shiratori [4] proposed a method for acquiring motion using sensors attached to a subject's body. Elhayek [5] proposed a method to acquire a subject's motions without markers from video captured by two cameras.

In this paper, on the other hand, we focus on the small movements of the body surface expressed by changes in textures to add a sense of liveliness to a subject by displaying these subtle movements. Even living things that remain still for a long time exhibit such subtle movements as breathing and blinking that remind us that they are indeed alive. We presume that presenting such subtle movements can impart a sense of liveliness in free view-point video. Kawabe [6] proposed Deformation Lamps to give motion to photographs and pictures by making subtle changes in textures and successfully provided the impression that the presented object is alive through subtle variations.

Structure-from-motion (SfM) [7][8] is a widely used method for 3D shape reconstruction and camera parameter estimation based on the correspondence of image feature points between multiple images of a subject. We acquire frames captured at the same time from multi-view videos of a subject's whole body and apply SfM to reconstruct the 3D shape. Textures to be mapped on the 3D model are also generated from the captured multi-view videos. By projecting video textures onto a static 3D model, subtle changes can be reproduced in body surface movements. Simply repeatedly playing the same movements causes habituation in long-term observations, thus impairing a sense of liveliness. In this paper, we introduce random elements into motions using Video Textures [9] to improve the sense of liveliness.

2. RELATED WORK

2.1. Dynamic free-viewpoint video

Carranza [2] proposed a method for generating dynamic free-viewpoint video by estimating the motion of a person's 3D model using silhouette information and mapping the corresponding texture. Collet [3] generated high-quality, streamable free-viewpoint video using a multi-view video recording system with more than 100 cameras. We aim to improve the expressive ability of free-viewpoint video by adding a sense of liveliness without such large movements as skeletal changes.

2.2. Video textures generation

The Video Textures [9] method synthesizes long time period videos that are not monotonously repetitive while maintaining video continuity by detecting similar frames in a short time period video and jumping to random frames. We apply Video Textures to free-viewpoint video. Ohta [10] confirmed that viewers are less likely to notice uncomfortable changes in textures when both the perspective and the video's subject are moving. We believe that the discomfort of flickering appearances caused by frame jumps can be further reduced by viewpoint switching.

2.3. Texture mapping methods

The following is a procedure for typical texture mapping of 3D human models: (1) generate a 3D mesh model from multi-view images; (2) expand the 3D mesh model into

2D coordinates (a *UV*-coordinate system); (3) generate texture from the multi-view images based on the *UV* coordinates (that represent the vertical and horizontal axes of the texture plane); (4) attach texture to the 3D mesh model [11]. Due to the high computational cost of generating multiple textures using the *UV*-coordinate expansion method, it is unsuitable for mapping video textures.

We solve the above mentioned problem by mapping video textures onto 3D human models using projective texture mapping [12]. Projective texture mapping is a technique that projects images onto 3D models, similar to projectors in the real world. In this mapping, captured images can be mapped onto the model without any sophisticated transformations by referring to the external parameters of the camera and positioning the projectors for projection. Therefore, projective texture mapping can be easily achieved if the video texture is generated from images taken from a fixed camera whose position and orientation do not change.

3. LIVELY FREE-VIEWPOINT VIDEO GENERATION USING VIDEO TEXTURES

Our proposed method consists of the following four elements (Fig. 1): I. multi-view video capturing, which gathers multiple views of the subject; II. 3D shape reconstruction, which reforms the 3D shape of the subject from the captured videos; III. video textures generation, which makes video textures with subtle changes in movement from the captured videos; IV. projective texture mapping, which maps the video textures onto the reconstructed 3D model.

With our proposed method, subtle changes in the movements of 3D models are expressed by projecting video textures to impart a sense of liveliness to free-viewpoint video.

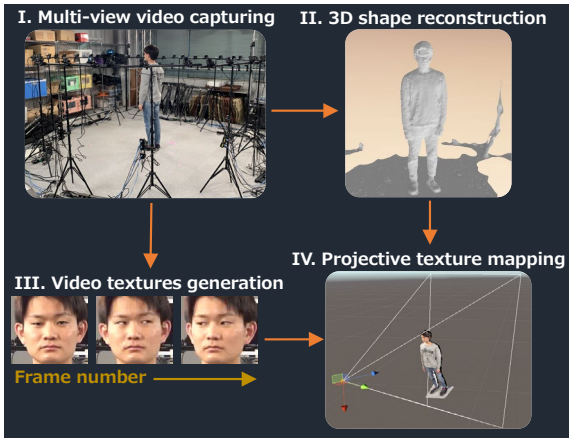

Fig. 1. Overview of lively free-viewpoint video generation using video textures

4. 3D SHAPE RECONSTRUCTION AND VIDEO TEXTURES GENERATION FROM MULTI-VIEW VIDEO

4.1. Multi-view video capturing

Figure 2 shows an overview of our developed multi-viewpoint video capturing system. Multi-view cameras are arranged in a circle around the target object to synchronously capture video for 3D shape reconstruction and video textures generation. Each camera is connected to a camera control box that synchronizes the cameras. The height of each camera is set at the same level as the subject's eyes. Similarly, the optical axis of each camera is adjusted such that the whole body of a subject falls inside the field of view.

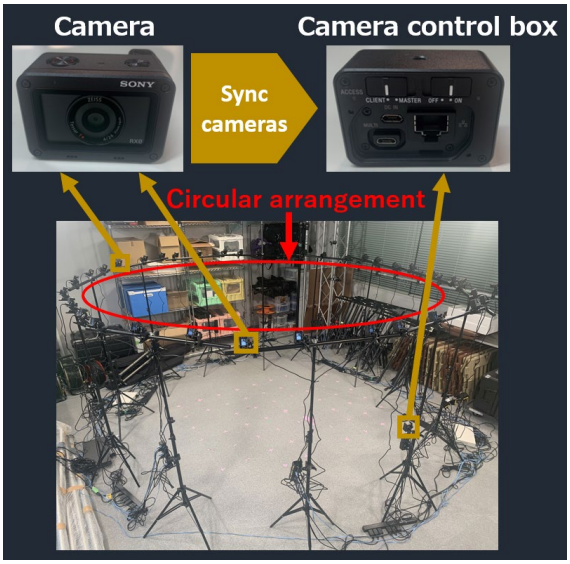

Fig. 2 Video capturing system with multiple synchronized cameras arranged in circle shape

4.2. 3D shape reconstruction

Figure 1-II shows a reconstructed 3D shape of a subject, recovered from the correspondence of the image feature points between the input images using SfM. The camera parameters for each shooting camera are estimated simultaneously. The estimated external parameters (position and orientation) of each camera are used as input for the position and orientation of the virtual projector for implementing projective texture mapping.

4.3. Video textures generation

From multiple viewpoint cameras, we select the main camera that is likely to influence the expression of the subject's sense of liveliness. In general, the camera that captures the subject from the front is most likely to be selected. Thereafter, the face region is extracted from the main camera video using a cascaded classifier [13]. The similarity between the extracted face regions is calculated. The pair of frames with highest similarity is selected as the transition frame pairs (Fig. 3).

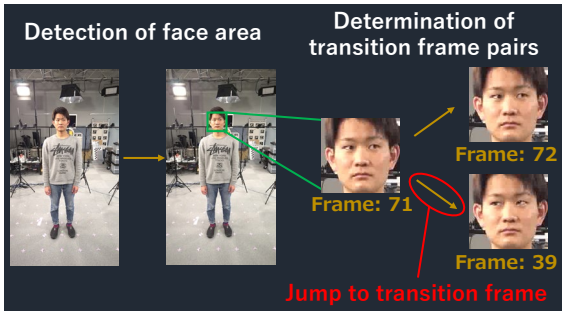

Fig. 3 Process of detecting subject's face region from main camera video using a cascaded classifier and determining transition frame pairs.

The following is the specific method for determining the transition frame pairs: steps a-d.

a. Generate distance matrix D, which is the sum of the squares of the differences in pixel values between frames:

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}, \quad (1)$$

where n is the number of frames of the captured video and d_{ij} is the sum of the squares of the difference of pixel values between the *i-th* and *j-th* frames of the detected face area.

b. Convolve obtained distance matrix D with diagonal matrix W to produce matrix D' with time-series weighting:

$$D' = D * W. \quad (2)$$

c. Find several components (i, j) of generated matrix D' with local minima, i.e., components (i, j) with high similarity in matrix D'.

d. Among the calculated components (i, j), select the one whose difference between i and j exceeds a threshold and make it a pair of transition frames. Here the threshold describes the minimum distance between both frames in a time series to avoid selecting neighboring frames and search for similar patterns in distinct time.

For each estimated pair of transition frames, we execute a backward jump with a certain probability. A backward jump is a move to a frame that goes back in a time series between transition frame pairs. By executing backward jumps, the variation in facial expressions will no longer be monotonous and repetitive, preventing viewers from becoming accustomed to the video.

5. DYNAMIC FREE-VIEWPOINT VIDEO GENERATION USING PROJECTIVE TEXTURE MAPPING

5.1. Projective texture mapping

This section describes implementation using a CG environment. The 3D model of the subject and the virtual projectors are placed in the constructed environment. For the position and the orientation of the virtual projector, we used the external parameters estimated by SfM. To present free-viewpoint video, the virtual projector must be selected based on the observation viewpoint. In this paper, we used a virtual projector with the shortest distance from the observation viewpoint. The left side in Fig. 4 shows how eight virtual projectors were arranged in a circle. In the same figure, the projection was performed using projector 1, which has the shortest distance from the camera for observation. The right side in Fig. 4 shows the result of projective texture mapping from projector 1, observed from the virtual camera.

The same process is repeated for each frame of the presented video to map dynamic textures onto the 3D model. The frame switching speed is set to match the frame rate of the captured video.

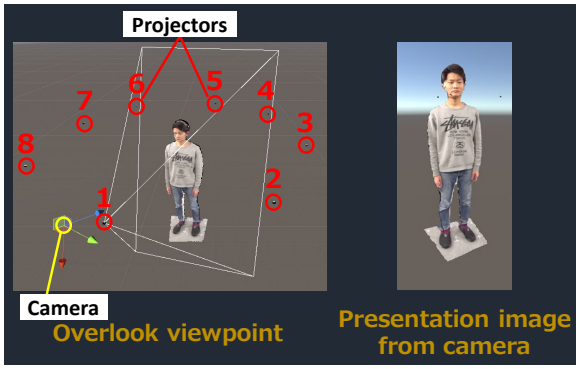

Fig. 4 3D model and projector arrangement in CG environment, and projective texture mapping from projector No.1 observed from overlook viewpoint (left)
Projective texture mapping from projector No.1 observed as images presented by camera (right).

5.2. Dynamic free-viewpoint video generation by subtle changes in textures

Figure 5 shows an example of a free-viewpoint video rendered by our proposed method. Its top row shows the projection results using projector 1 shown in Fig. 4, the bottom row of Fig. 5 shows the projection results using projector 8, and the numbers indicate the frame number of the projected video textures. These projection results in Fig. 5 confirm that the projector can be arranged in an appropriate position using the camera parameters estimated by SfM.

By observing the rendered result (Fig. 5), we can observe subtle variations in the textures from the projection by visually inspecting the changes in the eye position and the face orientation between the frames of the 3D model.

We subjectively evaluated the impression changes of the sense of liveliness depending on the following three

Fig. 5 Free-viewpoint video with projective texture mapping of video textures

texture conditions: "static textures," "video textures with frame jumps," and "video textures without frame jumps." As a result, we confirmed that mapping the video textures improves the sense of the liveliness of a 3D model in free-viewpoint video compared to mapping static textures. We conducted our evaluation with 15 participants (mean age: 24.0 years, standard deviation: 2.34) who observed videos that were virtually captured from an orbital trajectory around a target object in 24 seconds in a CG environment. After observing the videos with each texture condition, they rated the sense of liveliness on a 1-5 scale: 1: absolutely no sense of liveliness, 2: almost no sense of liveliness, 3 neither, 4: some sense of liveliness, 5: strong sense of liveliness. The average of their responses was 2.7 for "static textures," 3.7 for "video textures with frame jumps," and 3.5 for "video textures without frame jumps." The t-test result shows a statistically significant difference between the sense of liveliness between "static texture mapping" and "video texture mapping" with a 1% significance level.

6. CONCLUSIONS

This paper proposed a method that imparts a sense of liveliness to free-viewpoint video by expressing subtle changes in movement using video textures and projective texture mapping on a static 3D model.

We captured synchronized multi-view videos of the subject and applied SfM to estimate the 3D shape and camera parameters. By using Video Textures, which incorporate random elements into loop playback, we generated textures with subtle motions, such as facial expressions, and mapped them using projective texture mapping. We confirmed the effectiveness of our method in experimental evaluations.

This work was partly supported by Grants-in-Aid for Scientific Research (19H00806) and (17H01772).

7. REFERENCES

[1] I. Kitahara, H. Saito, S. Akimichi, T. Onno, Y. Ohta, and T. Kanade, "Large-scale Virtualized Reality," 2002.

[2] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-Viewpoint Video of Human Actors," ACM Trans. Graph., vol. 22, no. 3, pp. 569-577, 2003, doi: 10.1145/882262.882309.

[3] A. Collet et al., "High-Quality Streamable Free-Viewpoint Video," ACM Trans. Graph., vol. 34, no. 4, 2015, doi: 10.1145/2766945.

[4] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion Capture from Body-Mounted Cameras," ACM Trans. Graph., vol. 30, no. 4, 2011, doi: 10.1145/2010324.1964926.

[5] A. Elhayek et al., "Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3810-3818, doi: 10.1109/CVPR.2015.7299005.

[6] T. Kawabe, T. Fukiage, M. Sawayama, and S. Nishida, "Deformation Lamps: A Projection Technique to Make Static Objects Perceptually Dynamic," ACM Trans. Appl. Percept., vol. 13, no. 2, 2016, doi: 10.1145/2874358.

[7] S. Agarwal et al., "Building Rome in a Day," Commun. ACM, vol. 54, no. 10, pp. 105–112, 2011, doi: 10.1145/2001269.2001293.

[8] J. L. Schönberger and J. Frahm, "Structure-from-Motion Revisited," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104-4113, doi: 10.1109/CVPR.2016.445.

[9] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa, "Video Textures," in Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, 2000, pp. 489-498, doi: 10.1145/344779.345012.

[10] Y. Ohta, I. Kitahara, Y. Kameda, H. Ishikawa, and T. Koyama, "Live 3D video in soccer stadium," Int. J. Comput. Vis., vol. 75, no. 1, pp. 173-187, 2007, doi: 10.1007/s11263-006-0030-z.

[11] P. S. Heckbert, "Survey of Texture Mapping," IEEE Comput. Graph. Appl., vol. 6, no. 11, pp. 56-67, 1986, doi: 10.1109/MCG.1986.276672.

[12] C. Everitt, "Projective texture mapping," NVIDIA SDK, White Paper, 2008.

[13] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in Proceedings. International Conference on Image Processing, 2002, vol. 1, pp. I-I, doi: 10.1109/ICIP.2002.1038171.