21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Dynamic Scene Reconstruction using Asynchronous Multiple Kinects

Mitsuru Nakazawa, Ikuhisa Mitsugami, Yasushi Makihara, Hozuma Nakajima,
Hitoshi Habe, Hirotake Yamazoe and Yasushi Yagi
ISIR, Osaka University, Japan
{*nakazawa, mitsugami, makihara, nakajima, habe, yamazoe, yagi*}*@am.sanken.osaka-u.ac.jp*

Abstract

This paper proposes a novel method to reconstruct dynamic scenes by integrating depth data obtained by multiple Kinects, which cannot be synchronized to one another. In this method, the multiple Kinects located so as to cover the whole surface are firstly calibrated so that their depth data are mapped into the world coordinate system. The synchronous depth data for each Kinect is then generated by interpolation of temporally neighboring captured data. Experimental results of marching person reconstruction show the effectiveness of our method.

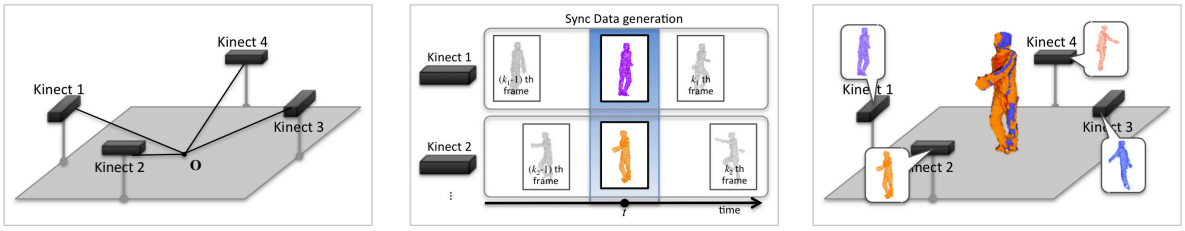
1 Introduction

The reconstruction of dynamic scenes containing moving or deformable objects is essential in various applications including mechanical analysis, virtual reality, computer graphics and robotics. One method is the shape-from-silhouette [8] technique, which uses multiple cameras to obtain the shape of an object as an intersection of visual cones corresponding to silhouettes of the object in captured images. However, this technique requires many cameras to recover detailed shapes. In addition, even with so many cameras, the method is seriously limited that it cannot recover concavity. Multi-view stereo (MVS) [3, 4, 13] is another way to reconstruct such scenes, but it works only for objects with dense textures. As a way to obtain accurate and dense shapes robustly, active 3D scanning systems are becoming popular. The minimum configuration of the system is a projector and a camera. A structured pattern is projected onto the object's surface, which is captured by the camera to obtain the shape according to triangulation. Furukawa et al. [2] extended the method to using multiple projectors and cameras to measure the whole shape. However, all the above-mentioned methods using cameras require that all cameras capture the scene synchronously. To fulfill this requirement, we must use special cameras that can operate synchronously. Since readily available cameras do not meet this requirement, we have to prepare special cameras, which are usually much more expensive.

On the other hand, consumer depth sensors such as the Microsoft Kinects are attracting the attention because of their low cost and ability to make 3D measurements. The Kinect is originally designed as a natural user interface so that it is mainly used to estimate the human pose [1, 14]. Considering their potential performance and reasonable cost, however, Kinects should not be limited to such the applications and they are expected to be applicable to our purpose, namely the reconstruction of the whole shape of an object, where the object is surrounded by multiple Kinects. Nevertheless, to reconstruct dynamic scenes using multiple Kinects, it is necessary to consider their asynchronous behavior as well as calibrate the multiple Kinects. While the cameras of multiple Kinects can be calibrated, asynchronization prevents us from obtaining the shape of dynamic objects. Since the Kinect is designed not for our purpose but as a natural interface, there is no way to achieve hardware synchronization. We are not aware of studies that solve this problem of asynchronization. Tong et al. [15] used three Kinect to scan the full human body in 3D, but the method employed can cope only with a human standing while not moving. Likewise, KinectFusion [6], which is a well-known 3D reconstruction technique, cannot treat dynamic scenes.

This paper proposes a novel method to reconstruct dynamic scenes by integrating depth data obtained by multiple Kinects. We believe it to be the first study to overcome the problem of asynchronous. In the proposed method, the multiple Kinects located so as to cover the whole surface of an object are firstly calibrated so that their depth data can be mapped into the world coordinate system. The synchronous depth data for each Kinect are then generated by interpolating temporally neighboring captured data. For this interpola-

978-4-9906441-1-6 ©2012 IAPR



(a) Extrinsic calibration (b) Temporal interpolation (c) Integration

Figure 1. Outline of the proposed method

tion, we use an extended method of Earth Mover's morphing (EMM) [9], which was originally designed for 2D silhouettes.

2 Proposed method

Figure 1 outlines the proposed methods. Multiple Kinects are located to cover the whole surface of a target object for reconstructing its whole shape. Firstly, these Kinects are calibrated to estimate their extrinsic parameters. Then, in order to overcome the asynchronization problem of the Kinects, we generate depth data at a certain time for each Kinect by temporal interpolation of its temporally neighboring data. Finally, the synchronous data of all Kinects are integrated into the world coordinate system for the whole surface reconstruction.

2.1 Calibration of multiple Kinects

The Kinect consists of a color camera and range scanner, whose captured images are aligned to each other by factory default setting. Thus, extrinsic calibration of the multiple Kinects can be achieved by practical multiple camera calibration such as the bundle adjustment [16]. For our 3D reconstruction tasks, however, we have to calibrate them more precisely considering the both images of the camera and range scanner.

First, before the extrinsic calibration, intrinsic calibration for each Kinect is needed. Intrinsic parameters including those about lens distortion are simply estimated by the Zhang's method [17]. Since, as above mentioned, the range image have already been aligned to the color image by default, these images are undistorted by the parameters. In addition, it has been confirmed that raw range data usually contains error, which increases quadratically as the distance between the Kinect [7]. We capture a planar regions at various distances and obtain the quadratic coefficient to compensate the error.

After the intrinsic calibration, we apply the bundle adjustment [16] to the cameras of the Kinects to estimate their extrinsic parameters. As it is weak calibration, we then give some metric information to make them metric parameters $\boldsymbol{R}_j^0, \boldsymbol{t}_j^0$ $(j = 1, \cdots, M)$ in the global coordinate, where M is the number of the Kinects. Moreover, to find the best parameters considering the range scanner as well as the cameras, we repeatedly update the parameters so that range data of the Kinects are well aligned in the global coordinate. In this process, degree of the alignment is evaluated using planer regions in the scene. By this optimization, we finally acquire the best parameters $\boldsymbol{R}_j$, $\boldsymbol{t}_j$.

2.2 Temporal interpolation

To generate range data at a given time t, we employ a new morphing technique that is originally proposed by Nakajima et al [12], which is a variant of EMM [10]. Refer to the original paper for its details. This section describes the technique just briefly.

$\boldsymbol{X} = \{\boldsymbol{x}_i\}$ $(i = 1, \cdots, N_P)$ denotes the range image, where N_P is the number of pixels. In this section, however, it is also regarded as a cloud of N_P 3D points each of which corresponds with a pixel of the range image; the temporal interpolation method described as follows is designed as a method for such a point cloud.

First, temporally neighboring two point clouds of a certain t are selected as the source and the destination ($\boldsymbol{X}^s$ and $\boldsymbol{X}^d$). For each of $\boldsymbol{X}^s$ and $\boldsymbol{X}^d$, N_C clusters are obtained by fuzzy means clustering [5]. We define the p-th cluster's mean $\bar{\boldsymbol{x}}_p$ and weight w_p as

$$\bar{\boldsymbol{x}}_p = \frac{\sum_{i=1}^{N} m_{ip}\boldsymbol{x}_i}{\sum_{i=1}^{N} m_{ip}}, \quad w_p = \sum_{i=1}^{N} m_{ip}, \tag{1}$$

where m_{ip} is a membership of the i-th 3D point to the p-th cluster, which satisfies $\sum_{p=1}^{N_C} m_{ip} = 1 \ \forall i$.

Next we acquire a correspondence so as to minimize transportation cost from the source to the destination point cloud. Let sets of means and weights for

the source clusters be $\{\bar{\boldsymbol{x}}_p^s\}$, $\{w_p^s\}$ $(p = 1, \cdots, N_C^s)$, and those for the destination clusters be $\{\bar{\boldsymbol{x}}_q^d\}$, $\{w_q^d\}$ $(q = 1, \cdots, N_C^d)$. The earth mover's distance flows are optimized so as to minimize the following transportation cost as

$$\{f_{pq}^*\} = \underset{\{f_{pq}\}}{\operatorname{argmin}} \sum_{p=1}^{N_C^s} \sum_{q=1}^{N_C^d} f_{pq} \|\bar{\boldsymbol{x}}_p^s - \bar{\boldsymbol{x}}_q^d\|^2, \quad (2)$$

where f_{pq} is flow from the the p-th source cluster to the q-th destination cluster, which is subject to $\sum_{p=1}^{N_C^s} f_{pq} = w_q^d \ \forall q$, $\sum_{q=1}^{N_C^d} f_{pq} = w_p^s \ \forall p$, and $f_{pq} \geq 0 \ \forall p, q$.

Finally, the source and the destination point cloud are transported. The i-th 3D point $\boldsymbol{x}_i^s$ in the source point cloud is transported to $\boldsymbol{x}_{ipq}^s$ by the flow f_{pq}^* from the p-th source cluster to the q-th destination cluster by the weights w_{ipq}^s, which are respectively defined as

$$\boldsymbol{x}_{ipq}^s = \boldsymbol{x}_i^s + \alpha(\bar{\boldsymbol{x}}_q^d - \bar{\boldsymbol{x}}_p^s), \ w_{ipq}^s = (1-\alpha) m_{ip}^s f_{pq}^*, \quad (3)$$

$$\alpha = (t - t_s) \,/\, (t_d - t_s), \quad (4)$$

where α is the time ratio, and t_s and t_d are time of the source and the destination point cloud, respectively. The destination point cloud is also transported in the same way.

Once the point cloud at t is obtained, points with smaller weights than a certain threshold are ignored. Then we resample the points so as to generate the range image $\dot{\boldsymbol{X}} = \{\dot{\boldsymbol{x}}_i\}$. On this resampling, each pixel value of the range image is determined by averaging depths of 3D points that are along the ray corresponding to the pixel.

2.3 Integration

By the process of Subsection 2.2, we can obtain synchronous range data of the multiple Kinects at an arbitrary time. For reconstructing the whole shape of the object, we transform the synchronous range data to the world coordinate system using the extrinsic parameters obtained in Subsection 2.1. Given range data of i-th Kinect $\dot{\boldsymbol{X}}^{c_j} = \{\dot{\boldsymbol{x}}_i^{c_j}\}$, it can be transformed to that in the global coordinate as follows:

$$\begin{bmatrix} \dot{\boldsymbol{x}}_i^{w_j} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_j & \boldsymbol{t}_j \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{x}}_i^{c_j} \\ 1 \end{bmatrix}, \quad (5)$$

where $\dot{\boldsymbol{x}}_i^{w_j}$ is the i-th 3D point transformed to the world coordinate system.

These range data are integrated into a whole surface using a mesh integration method by Pietroni et al [11].

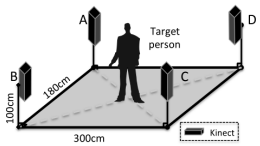
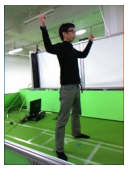


Figure 2. Layout of the environment



(a) Still pose

(b) Marching

Figure 3. Color images captured from a Kinect

3 Experiments

3.1 Experimental setup

For experimental evaluation, we located four Kinects to encircle a person who marched in place at the center of the environment (Figure 2) and direct to him/her. The Kinects were connected to different computers from one another, whose clocks were synchronized by a NTP server. The time stamp of each Kinect is then calibrated to the reference clock.

3.2 Shape reconstruction results

Before evaluating the effectiveness of the multiple Kinect synchronization, we first confirmed the accuracy of the calibration by reconstructing a static scene; here we use a still person as shown in Figure 4. The range data of each Kinect is depicted by its own color. This figure shows that the range data is well registered so as to reconstruct the person accurately, which means that the all Kinects were calibrated with accuracy enough to reconstruct the shape as a human body.

We then evaluate the effect of the temporal interpolation. Figure 5 shows results of the person who marched in place. Since his/her shape was continuously changed and the Kinects could not be synchronized, the range data of multiple Kinects could not be well registered without the temporal interpolation. In fact as shown in Figure 5(a), the range data could not be well registered especially around his/her left arm. On the other hand, with the temporal interpolation, it is confirmed that the shape is reconstructed accurately; no misalignment of multiple range data could not be observed as shown in Figure 5(b).

4 Conclusion

We proposed a novel method that reconstructs dynamic scenes using multiple Kinects instead of ex-

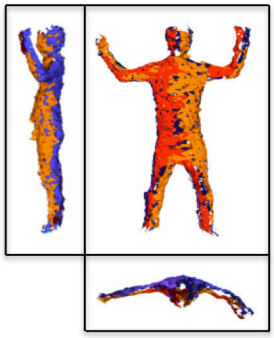
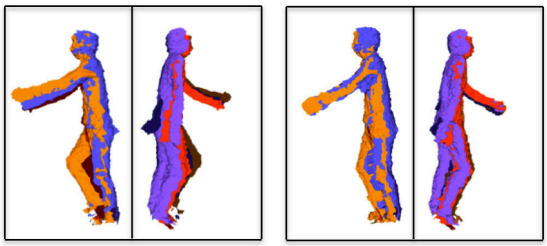


Figure 4. Reconstruction result of the still person. (Top left: side view, top right: front view, bottom: topview.)



(a) Without temporal interpolation (b) With temporal integration

Figure 5. Reconstruction result of the person who march in place. (Left: front side view, right: back side view.)

pensive synchronous cameras. In this method, multiple Kinects are accurately calibrated so that the range data can be well registered in the world coordinate. Moreover, to overcome the asynchronous problem of the Kinects, virtual synchronous range data is generated by the extended method of the EMM. The experimental results confirmed the effectiveness of the proposed method; the whole shape of a marching person, which has never been reconstructed, could be well reconstructed by the proposed method. In future work, we plan to temporally interpolate the color images as well as the range images for more realistic 3D modeling in real scenes.

Acknowledgement

This work was partly supported by the JST CREST "Behavior Understanding based on Intention-Gait Model" project.

References

[1] K. Berger, K. Ruhl, C. Brümmer, Y. Schröder, A. Scholz, and M. Magnor. Markerless motion capture using multiple color-depth sensors. In *Vision, Modeling and Visualization (VMV) 2011*, pages 317–324, 2011.

[2] R. Furukawa, R. Sagawa, A. Delaunoy, and H. Kawasaki. Multiview projectors/cameras system for 3d reconstruction of dynamic scenes. In *Proc. Workshop on Dynamic Shape Capture and Analysis*, 2011.

[3] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. In *Proc. CVPR*, 2007.

[4] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. In *Proc. CVPR*, 2008.

[5] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. John Wiley and Sons, 1999.

[6] S. Izadi, R. A. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. J. Davison, and A. Fitzgibbon. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks*, page 23. ACM, 2011.

[7] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[8] A. Laurentini. How far 3d shapes can be understood from 2d silhouettes. *IEEE Trans. on PAMI*, 1995.

[9] Y. Makihara and Y. Yagi. Earth mover's morphing: Topology-free shape morphing using cluster-based emd flows. In *Proc. Asian Conference on Computer Vision*, pages 2302–2315, 2010.

[10] Y. Makihara and Y. Yagi. Earth mover's morphing: Topology-free shape morphing using cluster-based emd flows. In *Proc. of the 10th Asian Conf. on Computer Vision*, pages 2302–2315, 2010.

[11] P. C. N. Pietroni, M. Tarini. Almost isometric mesh parameterization through abstract domains. *IEEE Transactions on Visualization and Computer Graphics*, 16(4):621–635, 2010.

[12] H. Nakajima, Y. Makihara, H. Hsu, I. Mitsugami, M. Nakazawa, H. Yamazoe, H. Habe, and Y. Yagi. Point cloud transport. In *Proc. ICPR*, 2012.

[13] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, 2006.

[14] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. In *Proc. of CVPR*, volume 2, pages 1297–1304, 2011.

[15] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–50, apr 2012.

[16] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment -a modern synthesis. *Lecture Notes in Computer Science*, 1883:298–372, 2000.

[17] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. on PAMI*, 22(11):1330–1334, 2000.