https://doi.org/10.2352/ISSN.2470-1173.2019.16.3DMP-006
© 2019, Society for Imaging Science and Technology

Real-Time 3D volumetric human body reconstruction from a single view RGB-D capture device

Rafael Diniz [+] ***and Mylène C.Q. Farias*** [*+] ***;***
[*] ***Department of Computer Science,*** [+] ***Department of Electrical Engineering;***
University of Brasília, Brasília, Brazil

Abstract

Recently, volumetric video based communications have gained a lot of attention, especially due to the emergence of devices that can capture scenes with 3D spatial information and display mixed reality environments. Nevertheless, capturing the world in 3D is not an easy task, with capture systems being usually composed by arrays of image sensors, which sometimes are paired with depth sensors. Unfortunately, these arrays are not easy to assembly and calibrate by non-specialists, making their use in volumetric video applications a challenge. Additionally, the cost of these systems is still high, which limits their popularity in mainstream communication applications. This work proposes a system that provides a way to reconstruct the head of a human speaker from single view frames captured using a single RGB-D camera (e.g. Microsoft's Kinect 2 device). The proposed system generates volumetric video frames with a minimum number of occluded and missing areas. To achieve a good quality, the system prioritizes the data corresponding to the participants' face, therefore preserving important information from speakers facial expressions. Our ultimate goal is to design an inexpensive system that can be used in volumetric video telepresence applications and even on volumetric video talk-shows broadcasting applications.

Introduction

Real-time 3D volumetric video acquisition and display systems use either a mesh or a point-cloud format to represent the 3D objects, which allow for a full 3D representation of the objects in the scene. In particular, voxelized point-cloud is a type of point-cloud that is used to represent solid objects, where each element is a small cube, called a voxel (analog to a pixel for 2D images) [7]. Unlike in regular 2D video applications, 3D objects captured by these systems need to be reconstructed after the the acquisition. Unfortunately, given the order and topology of the 3D-frames, the computational complexity and the amount of memory required by these systems is high [3]. As a consequence, the acquisition of a complete and accurate 3D representation of scene objects is not an easy task. In fact, common 3D acquisition systems are usually composed by an array of calibrated sensors. Examples of currently available volumetric video systems include the 8i [1] and the Microsoft's Holoportation [2].

In recent years, simpler 3D volumetric systems have been proposed. Rock *et al.* proposed a volumetric scene completion system with a single view RGB-D capture [4], which uses large categorized 3D models for object shape retrieval. Song *et al.* [5] proposed a semantic scene completion method, which uses a 3D convolutional neural network (CNN) trained with large 3D scene datasets (e.g. 45,622 houses with 775,574 rooms). Yang *et al.* [6] proposed a system that uses a generative adversarial neural network to reconstruct objects from a single view capture, which also takes advantage of large databases to train the network. The method proposed by Yang *et al.* has the highest resolution among the deep learning based volumetric reconstruction methods (256^3 voxel space). It is worth pointing out that the reference volumetric video test material sent to ISO/IEC/MPEG Point-Cloud Compression group by 8i [1] has frames of up to $1,024^3$ voxels resolution and more than one million occupied voxels.

In this paper, we propose a framework that captures 3D volumetric forms using just one RGB-D capture device. The system has the goals of capturing and representing 3D representations of human figures (speaker) for applications like mixed-reality video teleconferences. The big advantage of the proposed system is its simplicity. Instead of using a deep learning approach, like previous approaches, the proposed method relies only on straightforward geometric transformations. Therefore, it is fast and does not require powerful GPUs, supporting a $1,024^3$ voxel space and more than one million voxels. The proposed solution can be used not only in live two-way communication systems, but also in volumetric video broadcasting and Internet volumetric video services. For example, the method can be used in applications where a person is giving a speech or a class, presenting the news, or playing an online game.

This paper is organized as follows. Section details the framework of the proposed system, including the acquisition and reconstruction stages. Section discusses our results. Finally, Section presents our conclusions and future works.

Proposed system

This section describes the proposed method framework, which is composed by the following stages: 3D model capture, 3D reconstruction, and experimental setup.

3D Model Capture

The first stage of the framework consists of capturing a complete volumetric representation of the heads of the persons who will join the volumetric video session, i.e. the 3D models. The proposed framework creates a complete volumetric representation of the upper body of a person. We capture the data to create this volumetric model by moving the capture device (the Kinect 2) around the head of the person. We assume that:

1. The back of the head of the person is non-deformable;
2. The speaker is looking ahead during most of the time, allowing the RGB-D camera to capture the mouth and eyes of

Figure 1: 3D captured model.

the speaker;

3. Self-occlusions do not occur often.

Next, the volumetric model of the person's head is stored in a point-cloud format. Figure 1 shows an example of the captured model. With the volumetric model of the person's head stored, we can reconstruct the whole head of the speaker. We use a methodology based on the Truncated Signed Distance Function [8] and the Kinect Fusion algorithm [9] to assemble the volumetric 3D object representation model. Our methodology assures that higher dynamics of the object (mouth, nose, eyes) are fully represented in the reconstructed 3D volumetric stream. Since the method is not specific for the human body (eg. [10]), the propsoed framework can be easily extended to represent other types of objects.

Finally, we segment the captured 3D representation in two parts and represent them in cloud-point format. The first segmented part of the 3D representation corresponds to the person's nose and neighboring face regions, including the area of the eyes. The second segmented part of the 3D representation is the back of the head. The segmentation process uses maximum (or minimum) depth heuristics (depending on coordinate system) to identify each region of the face. Figure 2 shows the segmented point-clouds extracted from the 3D captured model. In the reconstruction step, this 3D model is registered and merged with the point-cloud frame captured live from the Kineptic 2 device.

Figure 2: Nose and adjacency (bottom) and back of the head (top) point-clouds extracted from the model.

3D reconstruction

After the model is captured, we can start the volumetric object reconstruction from a live capture system. First, the following pre-processing procedures are applied to each captured RGB and Depth frame pair:

Figure 3: Point-cloud from live single view capture.

- For each RGB and Depth frame pair captured, since the timestamps of the color and depth frames differ from 10 ms to 20 ms, we perform an alignment. It is worth pointing out that depth frames are especially important to convey high speed movements, with spatial details and temporal precision. Therefore, although the time differences are less than a frame period (∼33ms at 30fps), they have a big effect on the representation of movements.
- Using Kinect's intrinsic parameters, the RGB and Depth frames are converted to a point-cloud format, with camera coordinates converted to world coordinates, as shown in Figure 3;

Next, the volumetric object reconstruction is performed as follows:

- The point-cloud obtained from the live feed has its nose and adjacent areas segmented;
- Using a fast global registration method [11], we compute the transformation matrix between the segmented "face" from the model and the segmented input point-cloud;
- The segmented "back of the head" from the model is transformed using the computed transformation matrix, computed in the previous step;
- The transformed 3D model and live captured point-cloud "nose areas" are merged and the live reconstructed volumetric video frame is created.

Figure 4 shows several views of the reconstructed point-cloud. It is worth pointing out that the heart of our approach is the fast global registration algorithm [11], which aligns the pre-processed 3D model to the pre-processed point-cloud obtained directly from Kinect 2 (or any other capture device). Given its high efficiency, the model aligned with the live single view 3D input can be nicely merged to re-create the head of the telepresence participant.

Figure 4: Several views of a point-cloud volumetric form, reconstructed using the proposed system.

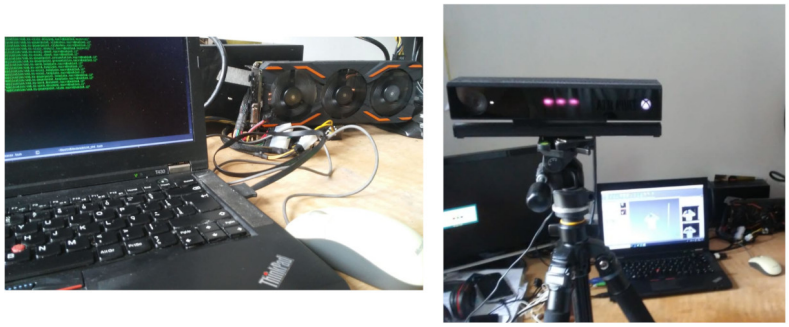


Figure 5: Left: Partial view of the notebook with an external GPU attached; Right: Kinect 2 connected to the notebook computer used for capture outside the laboratory (right).

Implementation and Experimental Setup

The proposed system is implemented in C and C++[1]. The code contains functions that perform basic point cloud operations, like geometric transformations, point-to-point distance measurement, crop and merge of areas. For the development of this work, we used the following libraries:

- OpenKinect's project libfreenect and libfreenect2 [12];
- Open3D [13].

The proposed system was tested using a high-end computer and a regular notebook computer. The high-end computer was a dual eight-core (32 SMT[2]) Intel Xeon E5-2620, with 80GB of RAM memory and two video cards, an NVidia Quadro P6000 and a NVidia GeForce GTX 1080. The notebook computer is a Lenovo ThinkPad T430 with a dual-core (4 SMT) Intel Core i5-3320M with 8GB of RAM with an external NVidia GTX 1080 GPU (see fig. 5). The notebook setup was also used when capturing outside the laboratory.

Early developments were performed using a Kinect 1 device. Later, the authors decided to change the capture device to a Kinect 2, which uses a time-of-flight ranging technology, as opposed to the structured light technology used by Kinect 1 [14]. Kinect 2 is also the most widely used RGB-D capture device for volumetric video production. It has a 1920×1080 RGB camera and its time-of-flight range sensor outputs a 512×424 depth frame. In the proposed system, the RGB frame captured by Kinect 2 is scaled and chopped to match the 512×424 depth resolution. The depth sensor supports distances of 0.5 m to 4.5 m and provides a field-of-view of 70.6° by 60° (H×V), with millimeter accuracy. In other words, Kinect 2 has a better accuracy and provides an output that is less noisy than Kinect 1. Nevertheless, the code was developed to work on both versions of Kinetic.

[1]Implementation source code is available, upon request.

[2]SMT: Simultaneous multithreading permits current CPUs to share CPU resources among 2 independent threads, improving the overall performance.

Results

Our tests show that each captured frame is processed, on average, under 33ms (the CPU used was a 16-core Intel Xeon E5-2620 at 2.1GHz). This acquisition and processing time guarantees that a 30fps input can be processed in realtime. More specifically, our system is a computationally fast volumetric video system that reconstructs 3D representations of the speaker in real-time. Since the system is implemented using a consumer CPU, there is room for improvement in terms of computation optimization, like implementing a GPU processing offload.

The proposed method produces a better mixed reality experience, when compared to other incomplete and open volumetric representations produced using just one RGB-D capture device. Similarly to methods that recreates a complete human body [10], the proposed framework can be extended to capture different types of objects, for which the changes occur mostly in one side/face of the object.

Conclusions

In this paper, we proposed a system that captures and displays 3D volumetric forms, using just one RGB-D capture device. The system is designed for human figures (speakers), targeting applications like mixed-reality video teleconferences. One contribution of the proposed system is the design of a CPU efficient volumetric video system, with state-of-the-art registration techniques that allow an efficient 3D capture setup and a real-time 3D object reconstruction. In the implemented method, the pre-processing steps, which are performed before the registration step, have an important role in determining the final quality of the

reconstructed volumetric form. We noticed that the fast global registration technique is also important and, if not adequately implemented, may incur in an imperfect transformation matrix. Finally, undersampling specific model areas also affects the quality of the reconstructed volumetric form.

Among the external factors that affect the quality of the reconstructed volumetric video are the differences in illumination between the model and the live RGB-D video frames. These differences cause color differences between the point-clouds and, consequently, produce merging (blending) artifacts. Figure 6 shows an example of a reconstructed head with blending artifacts, which can be seen in the area where the hair starts. Besides improving the registration algorithm, to solve issues caused by lighting differences between the model and the live frame, it is also necessary to improve the merging method. For this, we need first to implement an algorithm to monitor and compensate lighting changes. Also, we must implement a smart voxelization algorithm that correctly merges the point clouds into one voxelized point cloud [7]. With this algorithm, the merging procedure can prioritize voxels from one point-cloud over another to improve the performance of the reconstruction method. Finally, one possible future work can be the implementation of the proposed method using multiple RGB-D capture devices, working in parallel.

Figure 6: Reconstructed volumetric video frame sowing blending artifacts.

Acknowledgments

This work was supported in part by the Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the University of Brasília (UnB).

References

[1] E dEon, B Harrison, T Myers, and PA Chou, "8i voxelized full bodies, version 2–a voxelized point cloud dataset," *document MPEG*, p. m74006, 2017.

[2] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al., "Holoportation: Virtual 3d teleportation in real-time," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 741–754.

[3] Kazuo Sugimoto, Robert A Cohen, Dong Tian, and Anthony Vetro, "Trends in efficient representation of 3d point clouds," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 364–369.

[4] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem, "Completing 3d object shape from one depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2484–2493.

[5] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser, "Semantic scene completion from a single depth image," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 190–198.

[6] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen, "3d object dense reconstruction from a single depth view," *arXiv preprint arXiv:1802.00411*, 2018.

[7] Tommy Hinks, Hamish Carr, Linh Truong-Hong, and Debra F Laefer, "Point cloud data conversion into solid models via point-based voxelization," *Journal of Surveying Engineering*, vol. 139, no. 2, pp. 72–83, 2012.

[8] Brian Curless and Marc Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 303–312.

[9] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[10] Dimitrios S Alexiadis, Nikolaos Zioulis, Dimitrios Zarpalas, and Petros Daras, "Fast deformable model-based human performance capture and fvv using consumer-grade rgb-d sensors," *Pattern Recognition*, vol. 79, pp. 260–278, 2018.

[11] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun, "Fast global registration," in *European Conference on Computer Vision*. Springer, 2016, pp. 766–782.

[12] J Blake, F Echtler, and C Kerl, "Openkinect: Open source drivers for the kinect for windows v2 device," 2015.

[13] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[14] Diana Pagliari and Livio Pinto, "Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors," *Sensors*, vol. 15, no. 11, pp. 27569–27589, 2015.

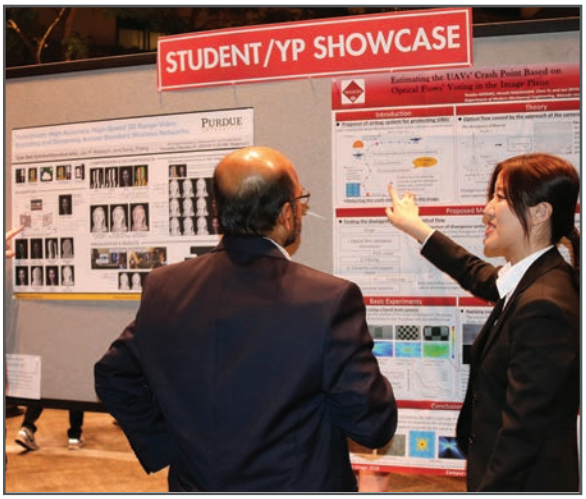
JOIN US AT THE NEXT EI!

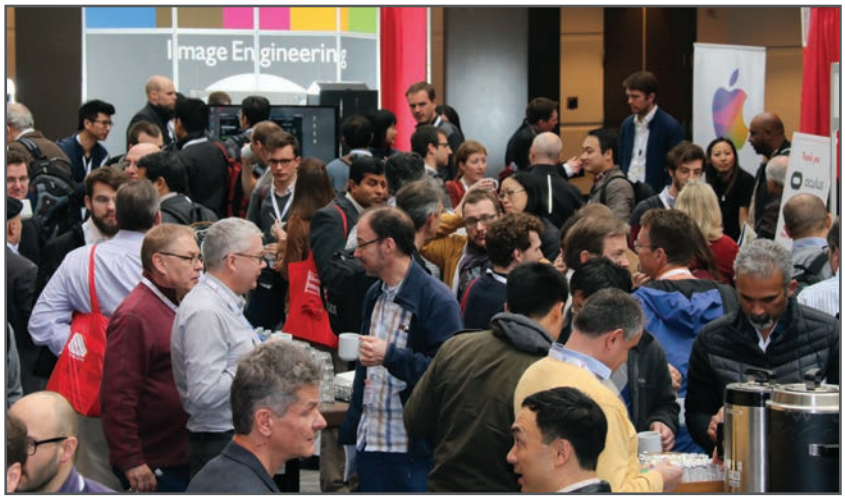
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org