Diminished Reality based on Kinect Fusion

Li, Qiaozhi / Li, Qiaozhi

(出版者 / Publisher)
法政大学大学院情報科学研究科

(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編

(巻 / Volume)
11

(開始ページ / Start Page)
1

(終了ページ / End Page)
6

(発行年 / Year)
2016-03-24

(URL)
https://doi.org/10.15002/00012918

Diminished Reality based on Kinect Fusion

Li Qiaozhi
Graduate School of Computer and Information Sciences
Hosei University
Tokyo 184-8584, Japan
qiaozhi.li.kk@stu.hosei.ac.jp

***Abstract*—While Augmented Reality allows adding virtual objects over a real scene, Diminished Reality can remove real objects from a scene. Currently existing multi-view based approaches and image completion based approaches to DR with 2D sensors. In this paper, we propose a practical approach to diminished reality for indoor environment with 3D model creation based on Kinect 3D sensors. Kinect can capture RGB information and per-pixel depth information of scenes at same time, and it can use the depth information to create 3D model. Our approach is based on the 3D model to search the region of the objects which will be removed. First we get the RGB and depth information of a scene from Kinect and get 3D model based on Kinect Fusion. After that an object is added into the scene. We compare the scene with original scene and process surface images acquired from Kinect fusion to find the accurate region of the object. At last we replace the data and diminish the object in real time. The solution uses Kinect fusion technique to create models and OpenCV for image processing. Our approach achieves a good result that we can diminish the objects clearly and restore the scene. Our approach provides real-time Diminished Reality for indoor environment.**

***Keywords*— Diminished Reality, Mediated Reality, Kinect.**

I. INTRODUCTION

In recent years, Augmented Reality (AR) develops at a high speed. AR is a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data [1]. AR enhances our perception of the real world by adding virtual objects to real scenes, not like virtual reality that replaces the real world. Some AR devices such as Google glass has been published, with these devices we can get more information.

While Diminished Reality (DR) is the opposite and the complement of Augmented Reality. DR is the name of area concerning about removing real objects from scenes in real time [2]. It also alter our perception by removing real object from the real scenes. Combine with AR, it can give people better visual experience. There are many fields that DR can be used such as removing undesirable people in live. Although several approaches have been presented recent years, our approach is to provide a way to realize DR with the Kinect sensor. Whatever is behind the object should be rendered when the object is removed. With Kinect we can get 3D information of a scene, it help us find the object that will be removed and render this region.

Kinect is a line of motion sensing input devices by Microsoft for Xbox and Windows. Kinect Fusion is a system for accurate real-time mapping of complex and arbitrary indoor scenes in variable lighting conditions [3]. User can paint a scene with the Kinect camera and simultaneously see, and interact with, a detailed 3D model of the scene [4]. Adding objects leads apparently change in depth map, so we can easily get the region of the object. And restore the scene clearly with the help of 3D model.

This paper is to present our approach to Diminished Reality and the remainder of this paper is organized as follows. Section II covers related work about Diminished Reality and Kinect. Section III describes the approach we supposed. Section IV shows the way we find the rough region to diminish based on Kinect. Section V explains image processing with OpenCV to get more accurate region. In section VI will discuss the replacement of the data from the original Kinect fusion image. Then we show some testing results for evaluation of the system performance in Section VII. In the last section, we will conclude about our approach, discuss limitations and looking into future work.

II. RELATED WORK

Existing approaches can be divided into two categories: multi-view based approaches and patch or fragment based approaches [5].

Multi-view based approaches use several cameras to get information of a certain area from different viewpoints. Through this way, the necessary information of the area behind the object to be diminished can be obtained. Zokai et al. [6] presented an approach based on multi-view. They applied a paraperspective projection model to generate the correct background from multiple calibrated view of the scene. The key point is using some calibrated images to reconstruct of a scene. And Jarusirisawad et al. [7] applied a plane-sweep algorithm to Diminished Reality. Their approach allowed for only weakly calibrated cameras, however, the number of cameras they used is up to 6. This category is use a set of cameras to get the images behind the objects. In contrast to our approach, some of approaches in this category cannot meet the requirement of real time.

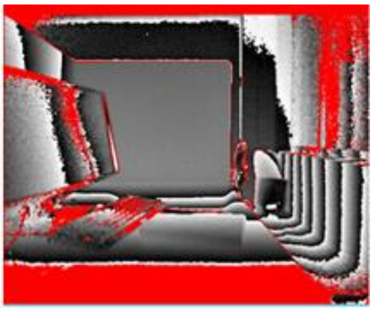


Figure 1 depth image

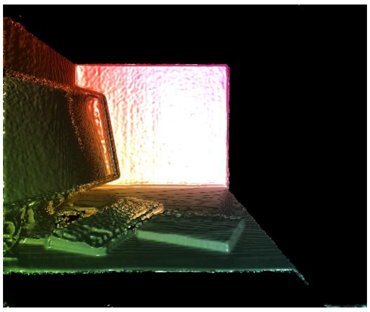


Figure 2 surface image

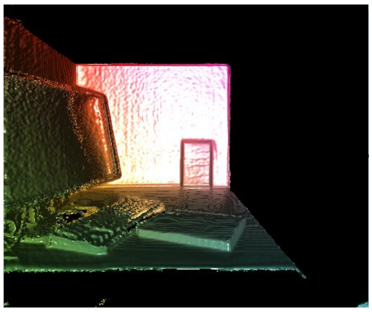


Figure 3 surface image after changing

Patch or fragment based approaches have become an attractive topic recently. This category does not need any background information. It is mainly based on algorithms of computer vision and image processing to remove objects and do reconstruction. In the area of image processing, inpainting is the technique that reconstructs small regions of images or removes small defects. At first, it was used to deal with deteriorated images to make them clearly. Recently it is used to diminish objects. Image inpainting also known as context-aware fill/repair, image completion, or image synthesis allows for filling masked or selected areas of an image by synthesized content, ideally undistinguishable from its environment, and recently becomes available as a standard feature in popular image manipulation tools [8]. In inpainting there non-exemplar based method and exemplar based method. Herling and Broll presented a real-time capable approach for object removal from video streams, allowing the manipulation of live videos [5]. It is based on an image completion and synthesis algorithm to detect the object. But it is not very good when applied to a big area of image. Then they presented a new approach for high-quality image manipulation. It gets better results for planar backgrounds. In contrast to those approaches, our approach wants to provide a real-time diminished reality based on 3D model which acquired from Kinect.

Recently 3D sensors have become available, we can expect that most robots in the future will be able to “see” the world in 3D [9]. Kinect is a good choice of 3D sensors, it is a commodity RGB−D sensor. Kinect can get color and depth information fast and not easily be interfered. Kinect uses time of flight (ToF) to get depth information, the resolution and precision is good. Kinect Fusion is a system for accurate real-time mapping of complex and arbitrary indoor scenes in variable lighting conditions [10]. The requirement for hardware is not high. Combined with augmented reality, diminished reality will lead to a broader prospect.

III. Approach

In this section we will present our approach to real-time Diminished Reality. As we use Kinect for windows v2, the latest version, there are some hardware requirements. Our environment includes a GTX760 graphics card, USB 3.0, windows 8.1 and i7-4790 CPU. Our approach needs a little pre-processing to get information about the background. While some approach find the information in pervious frames, in our approach we store one frame as a key frame for finding information. Our approach has three main steps:

1. **Object detection**

   First use Kinect fusion to get an original 3D model of the scene and store it. Then put an object into this scene and a new 3D model will be got. Through comparing this two models, we can analyze and get the rough region of the object.

2. **Image processing**

   In this step, our task is to find precise region of the object to be diminished. The rough region image has a lot of noise. The quality must be improved for following processing. Some image processing algorithm are applied to get the result.

3. **Repalcement**

   At last, we modify the data of video stream, using the image of the original scene to replace the region we get from step 2. Finally, we realize Diminished Realty.

Step 2 is the key point of our approach. If the region of the object to be diminished is precise, the reconstruction in step 3 will perform well

IV. Object Detection

This part shows how to detect the object to be diminished. Kinect is good tool that help us to detect it. For RGB images, detection of objects is based on frequency, color or some feature vectors. One common way is get the feature of the object, than search the same feature in the image to find that object. Superpixel is also a good way to divide image by the edge of each parts in the image, but it costs a lot time. It does not quite meet our requirements for real time. Considering depth image, once object is moved, the distance between sensor and that region will change. To search the region with depth sensor is more efficient. So we make use of Kinect v2 that can get RGB image at 1920 x 1080 resolution and depth image at 512 x 424 resolution. And Microsoft also provides SDK, Kinect for Windows SDK v2.0, to help developers to create applications using Kinect sensor technology on computers.

Figure 1 shows the depth image. Kinect’s depth image ranges from 500mm to about 8000mm. The depth value maps a distance in the view of each pixel. The grayscale represents the distance. The darker the area is, the farther the distance is. The red region represents out of range.

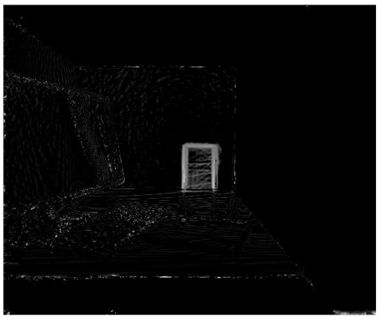


Figure 4 subtraction

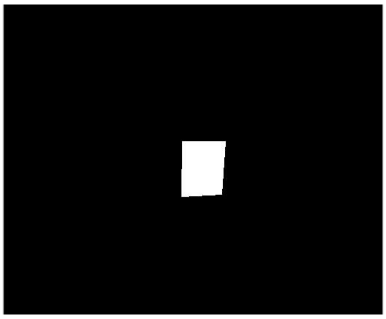


Figure 5 mask

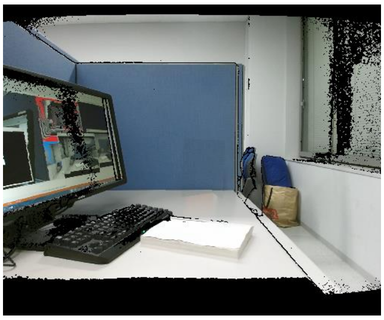


Figure 6 result of DR

Kinect fusion deals with depth image frame and get 3D model of the scene. Compare with other methods of 3D modeling, Kinect fusion is more convenient and rapid. For each depth frame, it needs to be smoothed, then process the frames in order to calculate point cloud which is a set of data points in some coordinate system, at last do 3D rendering. Figure 2 shows the 3D model of the scene we get through Kinect Fusion. It is called surface image.

From the surface image, contours of the scene can be seen clearly. If some objects are moved, distance will change. So the surface image will present the change of contours apparently. But it is hard to find contours with RGB image if the color of background is similar to the object. For example, a book is moved into the scene, then a new surface image is obtained as figure 3. Comparing figure 2 and figure 3, it is easy to find the differences and obtain the rough contour of the book.

In conclusion, in order to detect the object, first select one frame as a key frame before changing the scene. Both surface image and RGB image of this frame are stored. After changing, a new surface image is got. Then subtract the new surface image from the key surface image, result shows the rough region of the object. Figure 4 presents subtraction of figure 3 from Figure 2.

V. Image Processing

After first step, Rough region is found as shown in figure 4. But there are too much useless information in this image, it will lead to incorrect replacement in next step. It is necessary to filtrate this useless information and get precise region of the object. In this step, the useless information will be removed based on OpenCV. OpenCV (Open Source Computer Vision) is a library of programming functions mainly aimed at real-time computer vision[11] . OpenCV is a powerful tool, it has C++, C, Python and Java interfaces and supports Windows, Linux and Mac OS. OpenCV's application areas includes 2D and 3D feature toolkits, Structure from motion, Augmented Reality and so on.

There are many algorithms in image processing about filtrating useless information. The effect of filtration will directly affect the subsequent processing of image and reliability of analysis of the image. Filtering is a common way to preserve feature and suppress noise of target image. It performs well when dealing with noise such as Gaussian noise and Salt-and-pepper noise and so on. In various filtering methods, median filtering which is a nonlinear digital filtering technique is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise [12]. After median filtering, most of the useless information has been removed. But there still a little remains. Only filtering is not enough in our situation.

Morphological processing is another way to deal with image noise. Basic operations of morphology include erosion, dilation, opening and closing. Erosion allows thicker lines to get skinny and detects the hole. Dilation is the opposite of erosion. Opening operation essentially removes the outer tiny part. While closing removes small holes in the image. Different structuring elements in morphological processing lead to different results. It is great of significance to choose a structuring appropriate element. These techniques can also be used to find specific shapes in an image. In order to get a mask which indicates the region of the object as shown in figure 5, opening is used to remove the tiny useless information. A clearly image with almost no interference is obtained. Following is dilation operation which fills the entire contour to get the mask. And threshold is a necessary step because mask is a binary image that white area represents the object.

Sometimes, Even if the procedures above were be done, there may still some tiny disturbing parts and the shape of mask is not good. To remove disturbing parts and minimize the errors, more processing procedures are needed. As there are many white parts besides the correct one, the area of each part is calculated. The useless parts usually are very small, so a minimum value is set to ensure that the correct contours are chosen. Only the chosen parts remain white. At last, we applied a polygonal approximation of contours to fix the shape of mask to make it better.

Finally, the precise region is obtained.

VI. Replacement

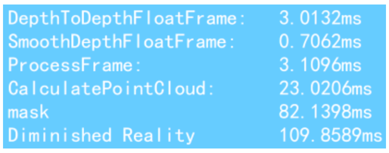
The last step is to replace the data of video stream. Every color frame will be modified. From Kinect, color frame and depth frame are obtained. Typically for ordinary users, color frame contains more information and more likely to be used. So our approach is deal with color frame.

Before dealing with color frame, each frame is copied from Kinect's buffer. As we mentioned before, the ratio of RGB image and depth image is different. If directly using the mask to deal with is color frame, it will cause some errors. Kinect has a function called MapDepthFrameToColorSpace can help us calibrate the color image with depth image. Each pixel in depth image has a corresponding color. This function can color the depth image based on the color image. After calibration, we make use of the key color image in first step and mask in second step to modify the frame. The mask provides the region and the key image provides color. With this information, the specific region is copied from the key image to current color frame. At last, the modified frame is sent to Direct2D to draw on the screen and we realize Diminished Reality. Figure 6 shows the result.

VII. Discussion

In this subsection we provide a detailed performance discussion of our approach. All tests are applied using a desktop Intel Core i7-4790 with 3.6GHz running Windows8.1, NVIDIA GTX760 graphics card, 32GB RAM, and USB 3.0. The implementation is realized in C++ with Visual Studio 2013.

Figure 7 shows an example of the time cost in each step of Kinect fusion, mask and the total time to realize diminished reality. Kinect fusion takes about 30ms to acquire a surface frame. It is the calculation of point cloud that spends most time, about 90 percent in Kinect fusion. Then It takes about 80ms to get the mask image. It spends a lot of time because the algorithms about image processing we used is slow. And many steps are executed, such as contour detection, to ensure the accuracy of mask image. Once the mask is obtained, the only necessary step is replacement. The replacement step costs only a few millisecond. In conclusion, our approach can meet the requirement of real time.



```
DepthToDepthFloatFrame:    3.0132ms
SmoothDepthFloatFrame:     0.7062ms
ProcessFrame:              3.1096ms
CalculatePointCloud:       23.0206ms
mask                       82.1398ms
Diminished Reality         109.8589ms
```

Figure 7 time of each step

Figure 8 shows the performance of other tests. It deals well with the object to be diminished. Because of the ratio of RGB image and depth image are different, the resolution of the image is limited to 512 x 424. The calibration colors the depth map, but some problems occur. As shown in figure 1, there are red areas on the display. The reason is reflection of display. Kinect uses IR projector and IR sensor to get depth image. Reflection of the surface may cause some problem. When color the depth image, these areas are not be colored.

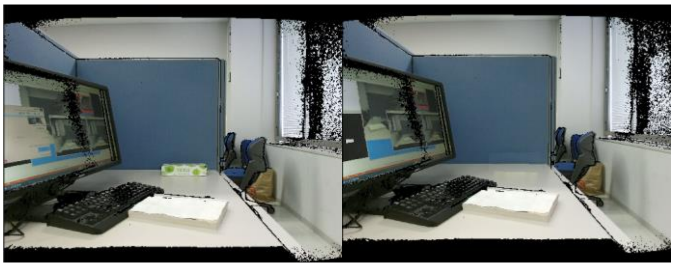


Figure 8 other test

VIII. Conclusion

In this paper we presented our Kinect based approach to real-time Diminished Reality. It concerns finding the region to be diminished based on 3D surface model with Kinect. While the procedure to obtain mask takes some times, it is acceptable for requirement of real-time. While many approaches applied inpaint algorithm to Diminished Reality which the result is not good when the background is complex, our approach provides real-time Diminished Reality with pre-processing and get a better result in some situations.

Now, in our approach, the position of Kinect must be fixed. It is a limitation. Tracking camera’s position is one of our future works. There are some function for Kinect to track camera. If we can apply it with our approach, it could be wider used. Other future work is to improve the quality of mask and reduce the time cost.

References

[1] Rolim Cledja and Veronica Teichrieb. "A viewpoint about Diminished Reality." 2012.

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

[3] Microsoft, MSDN Library “Kinect for Windows – Kinect fusion” , https://msdn.microsoft.com/en-us/library/dn188670.aspx.

[4] Herling Jan and Wolfgang Broll. "Advanced self-contained object removal for realizing real-time diminished reality in unconstrained environments." Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on. IEEE, 2010.

[5] Zokai S., Esteve J., Genc Y., and Navab N. "Multiview paraperspective projection model for diminished reality." Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on. IEEE, 2003.

[6] Jarusirisawad Songkran, Takahide Hosokawa and Hideo Saito. "Diminished reality using plane-sweep algorithm with weakly-calibrated cameras." Progress in Informatics 7 (2010): 11-20.

[7] Herling, Jan, and Wolfgang Broll. "Pixmix: A real-time approach to high-quality diminished reality." Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on. IEEE, 2012.

[8] Rusu, Radu Bogdan, and Steve Cousins. "3d is here: Point cloud library (pcl)." Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011.

[9] Izadi Shahram, Kim David, Hilliges Otmar, Newcombe Richard, Molyneaux David, Kohi Pushmeet, Shotton Jamie, Hodges Steve Fitzgibbon Andrew, Davison Andrew and Dustin Freeman. "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera." Proceedings of the 24th annual ACM symposium on User interface software and technology. ACM, 2011.

[10] Raman Maini & Dr. Himanshu Aggarwal “Study and Comparison of Various Image Edge Detection Techniques” International Journal of Image Processing (IJIP), Volume (3): Issue (1).

[11] Enomoto Akihito and Hideo Saito. "Diminished reality using multiple handheld cameras." Proc. ACCV. Vol. 7. 2007.

[12] Leao C. W. M., Lima J. P., Teichrieb V., Albuquerque E. S. and Kelner J. "Altered reality: Augmenting and diminishing reality in real time." Virtual Reality Conference (VR), 2011 IEEE. IEEE, 2011.