Fusion of Depth Maps Based on Confidence

Xiaoyang LIU, Xi YANG, Haobo ZHANG
International School
Beijing University of Posts and Telecommunications
Beijing China 102209
ee08b213@bupt.edu.cn ee08b256@bupt.edu.cn ee08b250@bupt.edu.cn

***Abstract*—In this paper, we propose a confidence-based depth maps fusion approach for multi-view 3D reconstruction. We present a four-stage method to refine the fused depth map to get a higher accuracy. At first, we utilize the off-the-shelf Kinect device to acquire the raw multi-view depth maps quickly. To improve the accuracy, we compute the confidence values for the estimated depth maps in the second stage to find a theoretically optimal value. Next, we discard some outliers within the confidence region to further optimize the depth maps. Finally, we attempt to fill the holes to refine the fused depth map. Similarly, we will point out some limitations of our confidence-based algorithm in obtaining the accurate depth map when putting into practice.**

Keywords- Depth Maps, Confidence, Kinect, Fusion

I. INTRODUCTION

Apparently, 3D reconstruction is quite a hot topic in computer vision. Past decades has witnessed the significant improvement and achievement in applications of accurate 3D reconstruction techniques. Appealing applications like Google Earth mushroomed recently and, undoubtedly, drew considerable attention from both the researchers and the end consumers. In addition, the state-of-the-art Kinect developed by Microsoft which can obtain the depth information simultaneously while capturing the real scenes received renewed attention in this research area. The more accurate the depth maps we can obtain, the better reconstruction results can be produced. Lying at the heart of the reconstruction procedure is to acquire such precise depth information.

In this paper, we investigate the depth maps fusion approach to achieve better accuracy of the initial depth information obtained by the off-the-shelf Kinect devices. We present a four-stage method to realize this intended goal. Firstly, we utilize the Kinect devices to obtain the multi-view depth maps. Due to variant circumstances, such as under intense light conditions, some unexpected errors and outliers can occur when using Kinect. Consequently, the accuracy of corresponding depth maps produced can be considerably reduced. In order to refine this result, we try to obtain the optimal fused depth map by computing the confidence values of all estimated depths from data captured by multiple views. Nonetheless, some supports for the estimated depth can lead to occlusions and violations which contradict the visibility constraints. Thus we perform the conflict detection in the third stage to discard some outliers within the support region to refine the fused depth map. Then, we will fill the holes in the fused depth map produced in the previous stage to optimize and smooth out the inliers. Ultimately, we present some limitations of our proposed algorithm.

II. THE PROCEDURE OF CONFIDENCE-BASED FUSION

A. *capture using Kinectcamera array*

Depth information is vital in 3D reconstruction procedure since the objects or scenes may not have consistent color and texture but must occupy an integrated region in space. Furthermore, pixels in a depth image denote calibrated depth in the scene rather than a measure of intensity or color. In addition, depth cameras possess significant advantages over traditional intensity sensor, such as working in low light levels, giving a calibrated scale estimate, being color and texture invariant, and resolving silhouette ambiguities in pose. [1] Our approach employs a set of Kinect devices which are at acceptable price and hold significant convenience in capturing relatively close scenes especially for human environment. We were inspired by the particular structure of Kinect devices which have three cameras enabling them to capture chromatic images and 3D depth images simultaneously. Actually, the middle one is the common RGB camera while those two accompanying cameras alongside it are the 3D depth sensors. The depth sensor consists of an infrared laser projector combined with a monochrome CMOS senor, which captures video data in 3D under any ambient light conditions. [2, 3]

It is this excellent feature that enables us to obtain the depth information of the objects efficiently and effectively.

Apparently, depth maps obtained by only employing such devices are relatively coarse compared to the one computed by implementing the plane-sweeping stereo. [4, 5, 6] Those obtained multi-view raw depth maps are refined in the consequent stage which applies the confidence-based fusion method to acquire the optimal depth maps.

B. *Terminology and Formula introduction in this paper*

Confidence denotes the certainty for some specific estimates. Particularly, in statistics, the confidence of one sampling estimate is defined as the probability when the error between the sampling indicator and the overall indicator is within a designated region.

978-1-4577-0321-8/11/$26.00 ©2011 IEEE

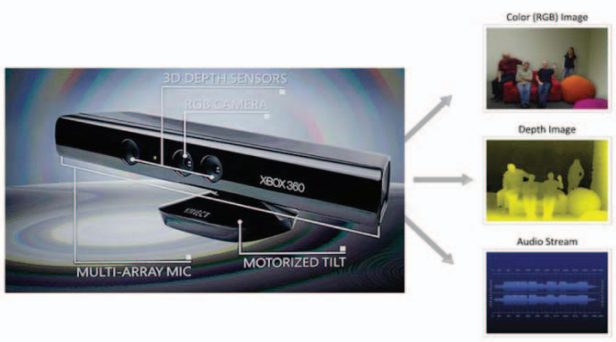

Figure 1. introduction of Kinect

Figure 2. The ideal capture result using Kinect [7]

Specifically, the concept of confidence refers to a function measuring the matching degree between the estimate value and the ground truth in both pattern recognition and image processing. Confidence has a rigorous definition in statistics. The mean value of one sampling statistics is labeled as m, its variance is σ, and the confidence region is denoted as $[m-\Delta, m+\Delta]$. Thus, the confidence for normal distribution $N(m,\sigma)$ is:

$$c = \int_{m-\Delta}^{m+\Delta} N(m,\sigma)\,dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{m-\Delta}^{m+\Delta} e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{y^2}{2}}dy \qquad (1)$$

Where $t = \Delta/\sigma$. Correspondingly, the confidence depends on t, the bigger t is, the larger this confidence value is.

However, the confidence is larger when the estimate is closer to the ground truth in the process of detecting of image feature points or some other similar random detecting results. The ground truth value is usually unknown during the detecting procedure, only the statistic algorithm can, therefore, be used.

A set of N depth maps of one designated pixel x is defined as $d_i(x)$ with the corresponding confidence labeled as $C_i(x)$. In our proposed method, the reference view can be chosen arbitrarily rather than solely selecting the one from the center camera. We obtain the depth map of each pixel in the reference view by operating the projection matrix. Consequently, we seek the optimal depth value via applying our confidence-based depth maps fusion method.

First, we define the term f(x) as the depth value of a 3D point x in the reference view. Relatively, $v_i(x)$ is for the distance between the centre of viewpoint i and this 3D point x with $d_{ref}(x)$ represents the distance of the current estimate $d_i(x)$ for the reference view. Accordingly, the depth map which is obtained by rendering the depth $d_i(x)$ into the reference view is identified as $d_i^{ref}(x)$ with C_i^{ref}representing the related confidence. Then, we further consider the possible relationships between the theoretical depths in the reference view and computed depths in other views. We define three visibility possibilities under this circumstance:

- We obtain that there exists a conflict between the estimated depth and the hypothesized one when considering the visibility relationship between the point O' observed in view i and point O in the reference view. This conflict lies in the difference between these two values $v_i(O) < d_i(O)$. The point O' could not be observed in the view i if there exactly was a surface at O, This situation is denoted as the free space violation.
- we define another possibility when the points observed in those two views (the reference view and view i) are exactly in the same location in free space. Nonetheless, we usually take two points as in agreement with each other if they satisfy the formula below in practice:

$$\frac{\left[d_i^{ref}(p) - d_i^{ref}(p')\right]}{d_i^{ref}(p)} < \varepsilon \qquad (2)$$

- The last visibility relationship is right contrary to the first situation. The point q'observed in the view i lies in front of the point q viewed by the reference camera, i.e.$d_i(q) < v_i(q)$. In this scenario, the rendered depth occludes q in the reference view. Therefore, we define this relationship as occlusion when $d_i^{ref}(q') < d_i^{ref}(q)$.

Having been informed of these fundamental ideas, we can apply our proposed confidence-based depth maps fusion method in order to acquire the optimal candidate to complete the refinement of the estimated depth maps.

C. *Confidence-based depth map fusion*

Due to the coarse quality of depth maps obtained by the Kinect camera set, we still need some refinements to get an optimal candidate. Since these depth maps contain a host of errors and hold a relatively low matching degree, some error correction and conflicts detection tasks are performed in the fusion step.

We have already obtained a set of N depth maps of one point with its corresponding set of confidence values. Then we select one camera as the reference view and render other different views into the reference one to compute the depth estimate of this point in the reference view. Unlike other precedent algorithms, our confidence-based depth maps fusion method test only one estimate by combining several close depth maps into this single one. Consequently, it decreases the computation complexity significantly where there are only $O(N)$ renderings to compute.

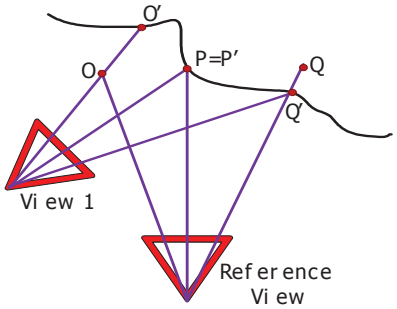

Figure 3. three visibility possibilities

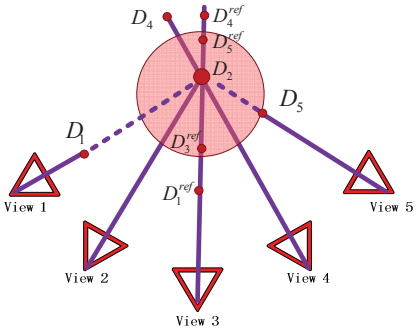

Figure 4.depth maps fusion

As what we have mentioned above, our method first renders all the obtained depth maps into the designated reference view. Then we take the depth estimate of each point with the highest confidence as the initial estimate. Specifically, the obtained information at each point x contains two types of quantities: the current depth estimate and the related level of support. Next, we denotes $f'(0)$ and $k'(0)$ as the initial depth estimate and the initial confidence separately. Lying at the center of this method is computing these two quantities iteratively to get the k_{th} iteration values. Therefore, we use $f_k(x)$ and $s_k(x)$ to represent the depth estimate and support level value correspondingly at iteration k .

We make further decision based on the following principle: we deem these two different views could reconstruct the same surface accurately when the depth value $d_i^{ref}(x)$ which has been rendered into the reference view lies within a small enough region around the initial depth estimate $f'(0)$.

Then we compute the weighted average value of the depth estimate in each view by the corresponding confidence according to the formula:

$$f_{k+1}(x) = \frac{f_k(x)s_k(x)+d_i^{ref}(x)C_i(x)}{s_k(x)+C_i(x)} \quad (3)$$

$$s_{k+1}(x) = s_k(x) + C_i(x) \quad (4)$$

Finally, we obtain the depth estimate for each pixel with its support level representing the matching degree of the depth maps with the depth estimate. In the next stage, we find the depth values that contradict the obtained optimal estimate $f_k(x)$ to verify whether it is correct or not.

Conflict Detection and Elimination

In this section, we will eliminate the conflicts which occurred out of confidence area. In C section which calculated confidence and depth maps weights, we consider the depth maps which is out of support region. Due to the conflict of visibility, it will influence on the accuracy of depth maps fusion. Consequently, we need wipe out these conflicts of visibility.

In B section, we have introduced the visible conflicts which consist of two parts. Next, we will eliminate the conflicts from these two situations. From figure 4, there are two views occurring occludes, and now we need to wipe out this situation which is out the supported region. In addition, the occlusion happened when the view has projected into reference view, thus the current confidence is equal to the confidence which worked out in the section C subtracts the confidence which occurred occlusion out of supported region:

$$s_{k+1}(x) = s_k(x) - C_i^{ref}(x) \quad (5)$$

Another situation is violation, showing in figure 3, which is resemble occlusion. In figure 3, the depth maps in view 4 occurs violation out of the confidence region. Thus, we need to eliminate the confidence of depth maps in the view 4. It is different from the situation from occlusion that happened in the reference view into which the view projected.

$$s_{k+1}(x) = s_k(x) - C_i(x) \quad (6)$$

After the conflicts detection and elimination, we have achieved the relatively accurate fusion of depth maps which is captured by Kinect. The principle which we followed is that if the depth information in one view has positive influence, we retain it and if the depth maps have negative influence, we need to eliminate them. Consequently, the depth maps after fusing will be more accurate.

D. Holes filling and smooth.

In the section D, we eliminate some depth information which occurred conflicts. It improves the accuracy of the whole depth maps, but leaves some holes in the picture and not smooth. Thus, we need to fill the holes and make them smooth.

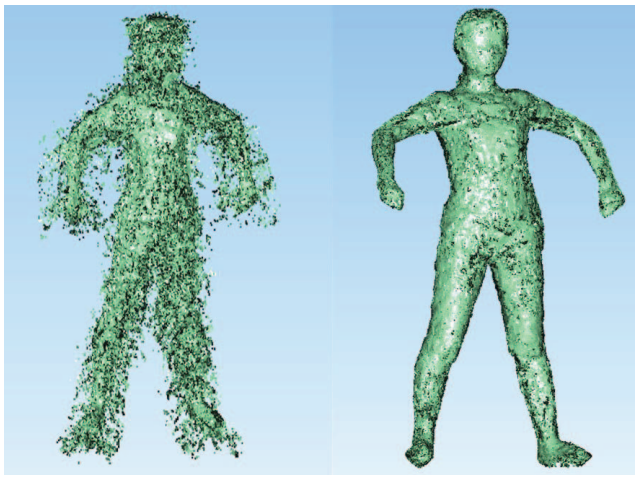

Figure 5. the left one is the result of fusion based on Confidence and the right one is the result of eliminating conflicted depth information .

There are many ways which can achieve the holes filling, we will choose one simple way. Firstly, we find the hole which is eliminated and neighbors which are in very small region. Then we use the neighbors' information to calculate the eliminated hole information by weighted average. The last step is to smooth out the filling information.

III. CONCULSION

We have presented a fast, convenient and accurate method to achieve the 3D reconstruction in four-stage process. The algorithm which is based on Confidence makes the process faster than many other methods about 3D reconstruction. We introduce the four steps to achieve the whole project in detail. They are separately collecting data using Kinect, fusion depth maps based on Confidence, detecting and eliminating the conflicts and filling holes. The most important steps in the process are fusing depth information and eliminating conflicts. The former makes the algorithm very fast but not accurate enough because it does not need to use traverse method for whole data, and the later improves the accuracy of the whole project.

Our method tested by multi-view 3D reconstruction. Firstly, we collect data from 12 views which are distributed averagely in one circle by Kinect. Then, we used the algorithm which is introduced before in detail to get the result showing in Figure 5. Our method suits the situation which aims to improve operation rate especially the 3D real-time reconstruction, and also can improve from the accuracy of the reconstruction.

IV. ACKNOWLEDGMENT

Supported by International School Beijing University of Posts and Telecommunications and Queen Mary, University of London.

V. REFERENCES

[1] Lu Xia, Chia-Chih Chen and J. K. Aggarwal. Human Detection Using Depth Information by Kinect.

[2] ^ a b c d e f ""Project Natal" 101". Microsoft. June 1, 2009. Archived from the original on June 1, 2009. Retrieved June 2, 2009.

[3] ^ a b Totilo, Stephen (June 5, 2009). "Microsoft: Project Natal Can Support Multiple Players, See Fingers". Kotaku. Gawker Media. Retrieved June 6, 2009.

[4] R.Collins. A space-sweep approach to true multi-image matching. In CVPR, pages 358-363, 1996.

[5] D.Gallup, J.-M.Frahm, P.Mordohai, Q.Yang, and M.pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In CVPR, 2007

[6] R. Yang and M.Pollefeys. A versatile stereo implementation on commodity graphics hardware. Journal of Real-Time Imaging, 11(1):7-18,2005

[7] Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio, Richard Moore Alex Kipman Andrew Blake, Real-Time Human Pose Recognition in Parts from Single Depth Images In CVPR 2011