3D Image Reconstruction and Human Body Tracking using Stereo Vision and Kinect Technology

Weidi Jia, Won-Jae Yi, Jafar Saniie and Erdal Oruklu

Department of Electrical and Computer Engineering,
Illinois Institute of Technology, Chicago IL, 60616

Abstract— **Kinect is a recent technology used for motion detection and human body tracking designed for a video game console. In this study, we explore two different types of 3D image reconstruction methods to achieve a new method for faster and higher quality 3D images. Generating depth perception information using high quality stereo image textures is computationally heavy and inefficient. On the other hand, depth information can be obtained very fast using Kinect but the overall 3D image quality is not refined and it is low resolution. Thus, in this study we explore the combination of higher quality images on a webcam and faster computation of depth information on Kinect in order to create an efficient and enhanced 3D image reconstruction system. This high resolution system has a broad range of applications including 3D motion sensing of human body, hands tracking and finger gestures.**

I. INTRODUCTION

Kinect is a motion sensing input peripheral device for the Microsoft Xbox 360 video game console, which is also known as *Project Natal* by PrimeSense. This device enables advanced user interaction in gaming experience, replacing ordinary joysticks with Natural User Interface (NUI) using gestures and voice commands.

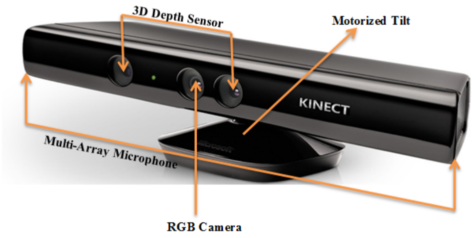


Figure 1. Microsoft Kinect

As shown in Figure 1, Kinect sensors are connected on a horizontal bar to a small base with a motorized pivoting capability designed for repositioning its view to detect players. The cameras from left to right are an infrared (IR) projector, an RGB camera and an IR monochrome camera which are combined together to obtain the 3-Dimensional (3D) image including depth map calculation. Embedded multi-array microphones enable this device to communicate with other players on Internet, and also recognize simple voice commands. Kinect is capable of providing full-body motion capture in 3D with facial and voice recognition using a Kinect Software Development Kit (SDK) [1]. Overall, the most valuable asset of Kinect is its efficient and effective depth map data retrieval mechanism. The hardware architecture of Kinect is shown in Figure 2.

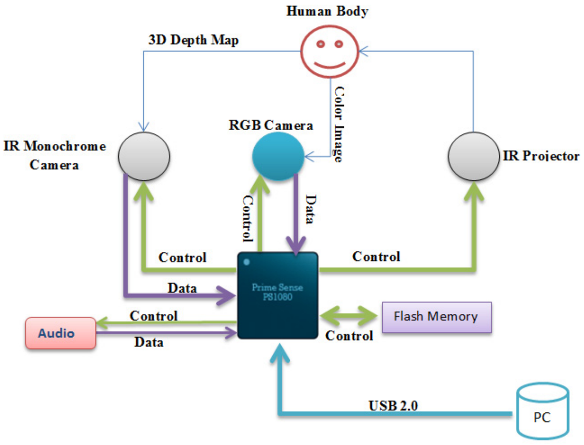


Figure 2. Kinect Hardware Architecture

A 3D depth map is generated by the reflected infrared signals retrieved on the IR monochrome camera where the signals are emitted from the IR projector. The range of the depth sensor is adjustable and Kinect SDK provides automatic calibration based on the physical environment. Chromatic and 3D depth images can be obtained simultaneously with texture from the RGB camera.

3D technology has been researched for a significant amount of time and has been focused on the study of stereoscopy in human visual system. Similar to the human stereo vision, depth perception can be obtained by projecting two images from two cameras. In this paper, we introduce a novel 3D image reconstruction method using the 2D image from a high resolution webcam combined with the Kinect 3D depth map. The proposed system provides a 3D reconstructed live image without glasses or any special display panel.

II. DEPTH MAPPING AND 3D GEOMETRICAL MODELS

A. *Kinect for 3D Depth Mapping*

3D vision is very important since it describes not only the shape, texture, and color, but also the depth and distance from the object. Hence, it is commonly used in diverse set of applications. Reliable depth estimation is one of the basic

978-1-4673-0818-2/12/$31.00 ©2012 IEEE

techniques in a robotic control system. OpenCV (*Open Source Computer Vision Library)* stereo vision is a widely used method to reconstruct the 3D image including the depth map. Recently, Microsoft announced Kinect, a new low cost and flexible game controller peripheral. Kinect uses IR projector and receiver to construct a 3D depth map very fast in real-time. Multi-purpose design of Kinect offers a variety of visual applications. In [2], 3D measurements were established using Kinect. Similar and improved 3D measurements are obtained in [3] using OpenCV calibration procedure, combined with the inverse disparity measurement model [4,5]. In [6], Kinect was used for calibrating the embedded sensors and real-time robotic control applications in a dynamic environment. Kinect and various sensors were combined in [7] for human motion sensing and rehabilitation purposes. For human motion sensing applications, Kinect enables simpler initialization procedures, better visualization of the estimated angles, and the capability to calibrate the inertial sensors in real-time [7]. Other researchers have implemented robotic hand operations that emulate human hand movements (such as finger and hand gestures) with 3D depth information acquired by Kinect [8-13]. Thus, Kinect is a powerful device for 3D image reconstruction and researchers have demonstrated its ability of capturing and analyzing human body dynamics. However, the computational efficiency and the image quality continue to be critical factors for real-time applications. In this study, low resolution of the Kinect 3D reconstruction mechanism is improved by combining Kinect with a high definition webcam for higher 3D image quality.

B. 3D Geometrical Models

3D geometrical theory is used for construct for representing real world objects in 3D image models from 2D images captured by cameras. Since different cameras have different intrinsic parameters (focal lengths expressed in pixel-related units and principal point at the image center) and extrinsic parameters (rotation matrix and translation vector), the calibration of the camera becomes the critical objective for 3D image reconstruction.

As shown in Figure 3, single camera calibration defines a *real-world coordinate system*, a *camera coordinate system* and an *image coordinate system*. The real-world 3D point, $P(X_W, Y_W, Z_W)$, in the *real-world coordinate system* projected to the *camera coordinate system*, $P_1(x, y)$, has *real-world coordinates* of $P_1(X_c, Y_c, Z_c)$ on the image plane. Furthermore, $P_1(u, v)$ has an image pixel point (u, v) in the *image coordinate system* [15]. While calibrating the camera, point P_1 is moving from one position to another position with respect to the images taken by the camera. Thus, the camera coordinate systems rotate and translate P from one location to another location. To estimate the extrinsic**,** one needs to calculate the rotation matrix R and the translation vector t. In contrast, camera intrinsic parameters remain unchanged because camera focal length is fixed.

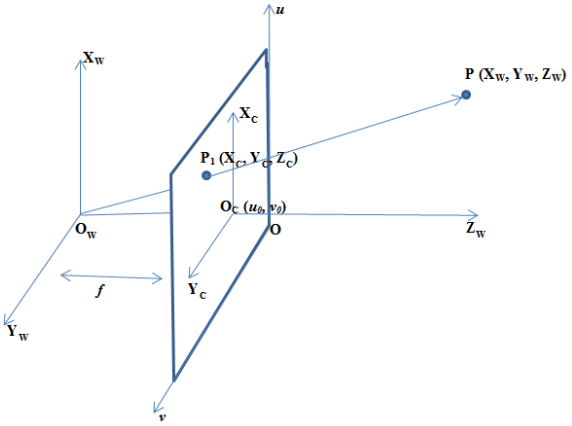


Figure 3. Single Camera Calibration Geometry

Hence, the calibration equation as given in [4] can be represented as:

$$s\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_x&0&u_0\\0&f_y&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}R&t\end{bmatrix}\begin{bmatrix}X_W\\Y_W\\Z_W\end{bmatrix}=M_1M_2P_W$$

The above equation represents the single camera calibration model, where (u_0, v_0) is the pixel location to reveal the displacement of the origin of the *camera coordinate system*, O_C, onto the origin of the *image coordinate system*, O; (f_x, f_y) is pixel-related focal lengths representing the actual focal length f on the *image coordinate system*; s is the image scale factor; matrix M_1 represents the intrinsic parameters and matrix M_2 is the extrinsic parameters; and P_W is the real-world 3D coordinates.

For two cameras used for stereoscopy, epipolar theory [14] is adopted for calibrating both cameras by estimating intrinsic and extrinsic parameters with respect to the corresponding points between the two camera coordinate systems. Each camera gives a different 3D back projection line from each focal point and these projection lines, as shown in in Figure 4, coincide at the real world point, P. The P_l is the projection line from left camera focal point O_l and P_r is the projection line from right camera focal point O_r; p_l and p_r are the corresponding points in the coordinate system associated with the two cameras; and e_l and e_r are the epipolar points for left and right images. Given by this stereo setup, it is possible to calculate the 3D position of a point by observing its corresponding positions in two different cameras. Epipolar lines are shown in bold lines and are marked in Figure 4.

The property of epipolar lines is defined in [14]**:**

“*If a feature projects to a point p_l in one camera view, the corresponding image point p_r in the other camera view must lie somewhere on an epipolar line in the camera image. An image point in camera 1 corresponds to an epipolar line in camera 2 and vice versa”.*

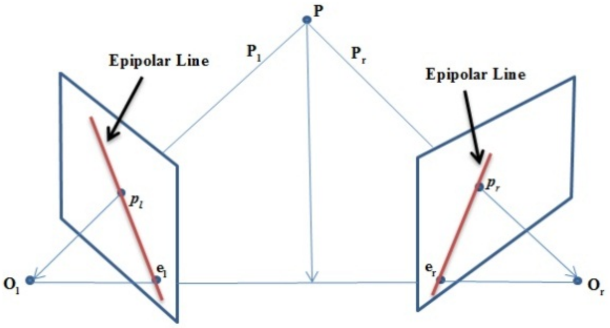


Figure 4. Two-Camera Image Planes and the Epipolar Geometry

In order to compute the 3D points, the epipolar theory is used to find the rotation matrix R and translation vector t by calculating the corresponding points on each image plane. In this study, OpenCV (*Open Source Computer Vision Library*) is utilized to calculate Matrix R and vector t by calling *StereoCalibration* function [4]. Additional details about 3D reconstruction algorithm can be found in Hartley and Zisserman's Multiple View Geometry [16].

III. 3D Reconstruction and Human Tracking

A. OpenCV Stereo Vision Calibration

The principle of the 3D geometrical model demands camera calibration and alignment for depth perception mapping. OpenCV stereo vision uses two cameras to construct stereoscopic image. Intrinsic and extrinsic parameters of the two cameras are analyzed by capturing 20 samples of stereo chessboard pictures (see Figure 5) and calculating positions and number of corners on the chessboard. In order to achieve accurate parameter estimation, a single camera calibration is performed for each camera. Then, stereo calibration is implemented by transforming those estimated parameters from each camera to their joint coordinate systems. After the stereo calibration, real-word objects can be represented in a depth map for 3D image reconstruction. The approach for the stereo calibration is shown in Figure 5 where the marked corners of the chessboard image calibrate the *real-world coordinate system* with the 3D geometrical model.

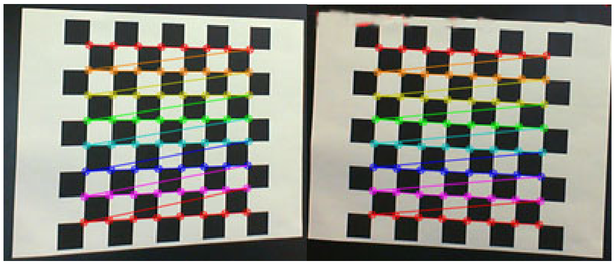
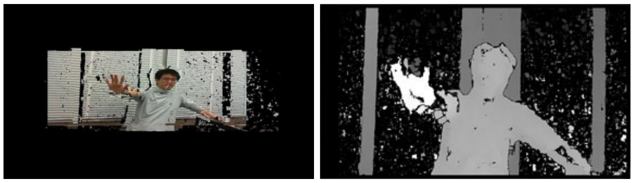


Figure 5. Image Calibration using Chessboard

B. 3D Image Reconstruction

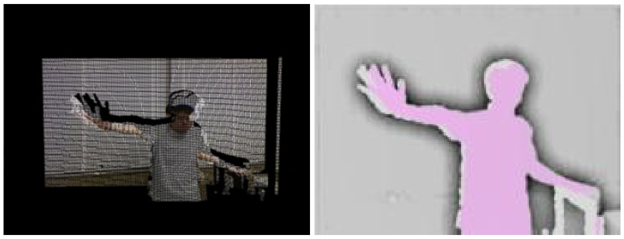
3D reconstruction is an important topic in computer vision. Past decades had witnessed significant achievements in applications of accurate 3D reconstruction techniques. Computation of depth perception image is essential in 3D image reconstruction. It is used in autonomous navigation, map building and obstacle avoidance. The more accurate the depth maps we can obtain, the better reconstruction results can be produced. In this research, we used OpenCV stereo vision combined with the OpenCV calibration function to estimate the intrinsic and extrinsic matrices. Then, we calculated the 3D depth map and derived 3D reconstructed image. Figure 6(a) is the high resolution 3D reconstructed image and Figure 6(b) is the corresponding 3D depth map. Noise is quite noticeable on the depth map which led to a corrupted image for 3D reconstruction due to high resolution mapping and the depth map estimation errors. In addition, heavy computation for OpenCV stereo vision 3D depth map is counterproductive for real-time 3D video streaming. However, the texture of the image is adequately clear for object recognition.

Compared to the stereo vision using two HD cameras, the Kinect 3D reconstruction method achieved the opposite results: low image quality and fast depth map computation. The main drawback of OpenCV stereo vision was the computation time for 3D depth map construction. IR sensors embedded on Kinect solves this issue by delivering instant and accurate depth perception information for faster 3D reconstruction. In this study, with the help of OpenGL (Open Graphics Library), we integrated the 3D depth map, obtained from Kinect SDK, with the image from the RGB camera on Kinect. Figure 7 shows the result of our experiment. Figure 7(a) is the reconstructed 3D image and Figure 7(b) is the 3D depth map. Compared to Figure 6(b), this method generates not only accurate and clear depth information, but also faster computation results. However, due to the low quality of the Kinect RGB camera, the texture of reconstructed 3D image is not as high quality as the previous method (see Figure 6a).



(a) (b)

Figure 6. Stereo Vision 3D Reconstruction (a) 3D reconstructed image; (b) Depth map

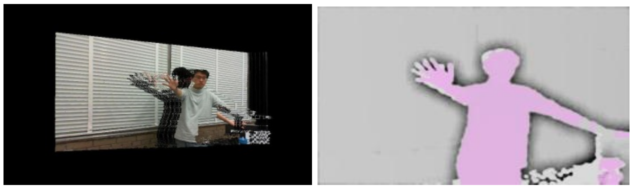


(a) (b)

Figure 7. Kinect 3D Image Reconstruction: (a) 3D image, (b) Depth map

C. 3D Image Reconstruction Using Kinect and HD Camera

Based on the experimental results discussed above, enhanced implementation should be considered to establish faster and higher quality 3D images. In this study, fast computation of depth perception information from Kinect and high quality texture from the HD webcam are combined to improve the outcome. Using Kinect SDK, a 3D depth map was retrieved and integrated with high quality image from the camera in OpenGL. Figure 8(a) shows the reconstructed 3D image in real-time using a high resolution webcam. Figure 8(b) is the generated 3D depth map by Kinect. Figure 9 describes the proposed 3D image reconstruction design flow.



(a) (b)

Figure 8. Combined method for 3D image Reconstruction (a) 3D image, (b) Depth map

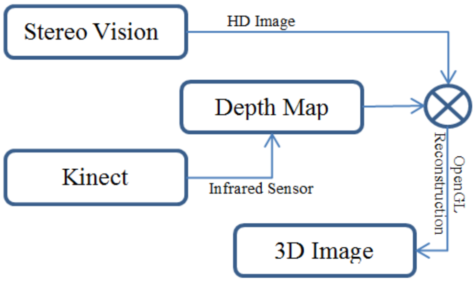


Figure 9. Proposed 3D Image Reconstruction Design Flow

This design flow based on Kinect and one HD camera offers real-time high resolution 3D image with 30 FPS (Frame per Second). On the other hand, for depth map calculation in OpenCV stereo vision, a high-end PC (Intel i7-920 CPU and 12GB 1333MHz RAM and ATI Radeon HD 5870 graphic card with 1GB memory) could not manage to deliver real-time 3D images with the same FPS.

D. Skeleton Image and Video Tracking

Video tracking has a wide range of applications such as human-computer interaction, security and surveillance, video communication and compression, traffic control and motion sensing of human body. In this study, Kinect was explored to estimate more accurate human motion. The depth map generated by Kinect can not only be used for 3D image reconstruction but also for skeletal tracking. This skeletal tracking is done by detecting the human body within the image to track movement. This image is created by estimating joint locations and producing an arbitrary skeleton image on the screen. Finger movement can be determined by assigning individual coordinate values for each joint within the image. Hand movement sensing is implemented in this study by analyzing the skeleton image, and the motor movement of Kinect can be controlled with respect to the position of the hands.

IV. Conclusions

In this paper, OpenCV stereo vision and Kinect are introduced for 3D image reconstruction. Taking advantage of the Kinect depth map with infrared sensors, faster generation of depth map is realizable. Also, with the high definition webcam, texture of reconstructed 3D images can be improved. Thus, combing the Kinect with the HD webcam can deliver high quality 3D image reconstruction for real-time video streaming. This high quality 3D image can be used for hand tracking and finger gesture detection and recognition.

References

[1] Kinect for Windows. [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/

[2] Smisek, Jan; Jancosek, Michal; Pajdla, Tomas, "3D with Kinect," *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp.1154-1160, 6-13, Nov. 2011.

[3] Burrus, N.; Kinect calibration on cameras. [Online]. Available: http://nicolas.burrus.name/index.php/Research/KinectCalibration.

[4] (2011) OpenCV documentation: Camera Calibration and 3D reconstruction. [Online available: opencv.willowgarage.com] http://opencv.itseez.com/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html.

[5] (2010) Kinect _node. [Online]. Available: http://www.ros.org/wiki/kinect_node.

[6] Stowers, J., Hayes, M., and Bainbridge-Smith, A.; "Altitude control of a quadrotor helicopter using depth map from Microsoft Kinect sensor," *Mechatronics (ICM), 2011 IEEE International Conference on*, pp.358-362, 13-15 April 2011.

[7] Bo, A.P.L., Hayashibe, M., and Poignet, P.; "Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect," *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp.3479-3483, Aug. 30 2011-Sept. 3 2011.

[8] Chaudhary, A., Raheja, J.L., Das, K., and Raheja, S.; "A Survey on Hand Gesture Recognition in context of Soft Computing", Published as Book Chapter in *Advanced Computing CCIS*, Springer Berlin Heidelberg, Vol. 133, pp. 46-55, 2010.

[9] Garg, P., Aggarwal, N., Sofat, S., "Vision Based Hand Gesture Recognition", *World Academy of Science, Engineering and Technology*, Vol. 49, pp. 972-977, 2009.

[10] Chaudhary, A., Raheja, J.L, and Das, K.; "A Vision based Real-time System to Approximate Fingers Angles", *Proceedings of IEEE International Conference on Computer Control and Automation (ICCCA 2011)*, Jeju Island, South Korea, pp. 118-122, 1-3 May, 2011.

[11] Chaudhary, A., Raheja, J.L., Singal, K., and Raheja, S.; "An ANN based Approach to Calculate Robotic fingers positions", Published as Book Chapter in *Advances in Computing and Communications, CCIS*, Springer Berlin Heidelberg, Vol. 192, pp. 488-496, 2011.

[12] Raheja, J.L., Chaudhary, A., and Singal, K.; "Tracking of Fingertips and Centers of Palm Using KINECT," *Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third International Conference on*, pp.248-252, 20-22, Sept. 2011

[13] Liu, X., Yang, X., and Zhang, H.; "Fusion of depth maps based on confidence," *Electronics, Communications and Control (ICECC), 2011 International Conference on*, pp.2658-2661, 9-11 Sept. 2011.

[14] Hillman, P.; *White Paper : Camera Calibration and Stereo Vision," Matrix,* pp. 1-14, Square Eyes Software, www.peterhillman.org.uk, 2005.

[15] Bradski, G., and Kaehler, A.; *Learning OpenCV, Computer Vision with the OpenCV Library*, O'Reilly Media Publisher, 2008.

[16] Hartley, R. and Zisserman, A.; *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.