Benchmark on outdoor scenes

Hairong Zhang[a*], Cheng Wang[a], Yiping Chen[a], Fukai Jia[a], Jonathan Li[b]
[a]Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Information Science and Engineering, Xiamen University, 422 Siming Road South, Xiamen, Fujian 361005, China;
[b]Dept. of Geography and Environmental Management, University of Waterloo,
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

ABSTRACT

Depth super-resolution is becoming popular in computer vision, and most of test data is based on indoor data sets with ground-truth measurements such as Middlebury. However, indoor data sets mainly are acquired from structured light techniques under ideal conditions, which cannot represent the objective world with nature light. Unlike indoor scenes, the uncontrolled outdoor environment is much more complicated and is rich both in visual and depth texture. For that reason, we develop a more challenging and meaningful outdoor benchmark for depth super-resolution using the state-of-the-art active laser scanning system.

Keywords: Depth, super-resolution, outdoor, benchmark, laser scanning

1. INTRODUCTION

Research on range image is becoming more and more popular. A variety of range image analysis and processing methods have been demonstrated in the computer graphics and computer vision literature. These include range image recognition and segmentation, range image reconstruction, range image super-resolution, etc.

A number of depth super-resolution techniques have been proposed in the literature. For example, a depth super resolution framework for range images and subsequent evaluation on Middlebury benchmark is introduced in [6]. Nevertheless, most of the available depth super-resolution techniques are tested using only indoor data sets with ground truth measurements obtained from structured light techniques. On the other hand, it has been demonstrated that many vision algorithms performing well on indoor data sets rank below average when tested under uncontrolled real world environment.

Unlike indoor scenes, the uncontrolled outdoor environment is much more challenging, and the real scenes are often rich in both visual and depth texture (e.g., bushes and trees). Additionally, as structured light techniques are vulnerable to outdoor environment where strong lighting is presented, a state-of-the-art laser scanner is used for data acquisition. We propose to develop a benchmark for quantitative evaluation of depth super-resolution algorithms using data sets captured under outdoor environment. The rest of this paper is organized as follows. Section 2 describes our system, and sections 3 and 4 introduce the registration between scanner and camera. Then we show experiments in section 5 and conclusions in section 6.

2. OUR SYSTEM

A photo of the 3D scanning system is presented in Fig. 1. It has two parts: VZ 1000 and Nikon D300s. VZ 1000 is a state-of-the-art laser scanning system, of which accuracy can reach 8mm at a distance of 100m. It is designed for dense and accurate 3D scanning and thus is suitable for generating high resolution depth images. This scanning system comes with a digital camera mounted on it manually. Nikon D300s is a high resolution camera, which can provide images of up to 4288×2848 pixels. It is used for getting color (RGB).

The resolution of the laser scanner can be even higher than the RGB camera but scanning time is too much. Sparse 3D point clouds can be captured by increasing angular step-width of the laser scanner.

[*] zhr7751662@163.com

2nd ISPRS International Conference on Computer Vision in Remote Sensing (CVRS 2015), edited by Cheng Wang, Rongrong Ji, Chenglu Wen, Proc. of SPIE Vol. 9901, 99010U · © 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2234860

For the scanner system we use, three coordinate systems should be mentioned. SOCS represents an abbreviation of scanner's own coordinate system, while CMCS an abbreviation of camera's coordinate system. At the same time, IMCS is Image's coordinate system. In the following part, we will use these abbreviated forms.

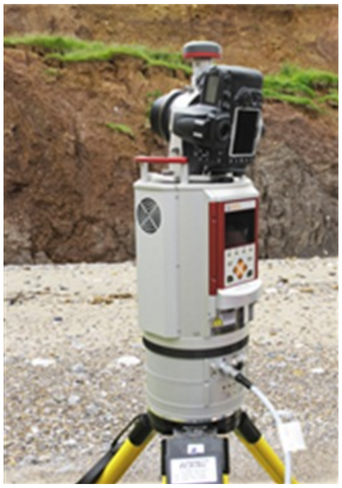

Figure 1. Our 3D Scanning System. The top part is camera and the bottom part is laser scanner.

3. COORDINATE TRANSFORMATION

This section mainly introduces coordinate transformation from scanner to camera. This includes two parts, conversion from SOCS to CMCS and from CMCS to IMCS. Through transforming, we can convert a 3D point in SOCS into a corresponding pixel in IMCS, and further convert point cloud into depth image corresponding to RGB image.

3.1 Extrinsic parameters

The transformation from SOCS to CMCS, a rigid transform, can be described with a rotation matrix R and a translation matrix T.

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = R_{3\times3} \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} + T_{3\times1} \tag{1}$$

For convenience, we use the following homogeneous coordinate expression:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_s \\ y_s \\ z_s \\ 1 \end{bmatrix} = E \begin{bmatrix} x_s \\ y_s \\ z_s \\ 1 \end{bmatrix} \tag{2}$$

In this step, we finish the conversion from a 3D point (x_s, y_s, z_s) in SOCS to a 3D point (x_c, y_c, z_c) in CMCS. Matrix E is extrinsic parameter matrix with 6 parameters (3 rotation variables and 3 translation variables).

3.2 Intrinsic parameters

The transformation procedure from CMCS to IMCS actually includes perspective projection, scaling and translation operations:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z_c}\begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \frac{1}{z_c} * K * \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{3}$$

Where f_x and f_y denote equivalent focal length in x and y directions respectively, and u_0, v_0 are pixel coordinates of principle point. In this step, we finish the conversion from a 3D point (x_c, y_c, z_c) in CMCS to a pixel (u, v) in IMCS. Matrix K is the well-known intrinsic parameter matrix with 4 parameters.

Combine the above steps, we have

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z_c} * K * E * \begin{bmatrix} x_s \\ y_s \\ z_s \\ 1 \end{bmatrix} \tag{4}$$

With K and E, we can finish the transition between 3D point cloud and 2D image.

3.3 Camera distortion

Actually camera model is not the ideal linear model because of imperfect camera manufacturing technology. Therefore, camera distortion should not be ignored especially when accurate camera calibration is required. Generally speaking, camera distortion mainly consists of two directions' distortion: radial and tangential distortions.

With the auxiliary terms $x = \frac{(u-u_0)}{f_x}$, $y = \frac{(v-v_0)}{f_y}$ and $r^2 = \tan^{-1}\left(\mathrm{sqrt}\left(x^2+y^2\right)\right)$, radial distortion ($\delta_{ur}$, δ_{ur}) can be computed like this:

$$\begin{cases} \delta_{ur} = x * f_x\left(k_1 * r^2 + k_2 * r^4 + k_3 * r^6 + k_4 * r^8\right) \\ \delta_{vr} = y * f_y\left(k_1 * r^2 + k_2 * r^4 + k_3 * r^6 + k_4 * r^8\right) \end{cases} \tag{5}$$

where k_1, k_2, k_3 and k_4 are radial distortion parameters. Meanwhile, tangential distortion (δ_{ut}, δ_{ut}) is like this:

$$\begin{cases} \delta_{ut} = 2 * p_1 * f_x * x * y + p_2 * f_x * \left(r^2 + 2 * x^2\right) \\ \delta_{vt} = 2 * p_2 * f_y * x * y + p_1 * f_y * \left(r^2 + 2 * y^2\right) \end{cases} \tag{6}$$

where p1 and p2 are tangential distortion parameters. Then the relation between distorted (u_d, v_d) and undistorted (u, v) image coordinates is defined by the following polynomial (closely according to the OpenCV style):

$$\begin{cases} u_d = u + \delta_{ur} + \delta_{ut} \\ v_d = v + \delta_{vr} + \delta_{vt} \end{cases} \tag{7}$$

Therefore, before we carry on linear coordinate transformation, digital image distortion must be removed. What's more, camera distortion only depends on optical equipment's interior structure. It keeps unchangeable for one existing device as well as intrinsic parameters. For our system, camera Nikon D300s itself has a fixed focal length and has been already calibrated with known distortion parameters. This makes us convenient to get undistorted images by eliminating distortion.

4. REGISTRATION

The use of our system simplifies the data acquisition problem. The major issue left is the alignment of the 3D measures and the 2D image captured by the installed RGB camera. The purpose of alignment is to solve extrinsic, intrinsic parameters, namely registration. At the beginning, we introduce the popular checkboard as calibration pattern so as to acquire adequate tie points (2D pixel dots and corresponding 3D points).

4.1 Checkerboard

Checkboard is widely used in different kinds of calibration, such as Binocular Stereo Vision. For 2D vision, through extracting corners of checkboard using Matlab Toolbox, we can get precise enough corner pixel dots as tie points for 2D image directly. However, for 3D point cloud, it becomes complicated unexpectedly.

There is no available and immediate way to extract tie points required from 3D point cloud, which is also impractical on account of complexity and enormous data volume of 3D scene. As we mentioned, we can extract accurate tie points in 2D vision, which can give us a clue. If we can project 3D point cloud to 2D image, then everything is done. Indeed, we just do that.

In addition to 3D measures, the laser scanner also provides a reflectivity measure for each scanned 3D point, which greatly contributes to projection. In other words, the projected 2D image is mapped by reflectivity.

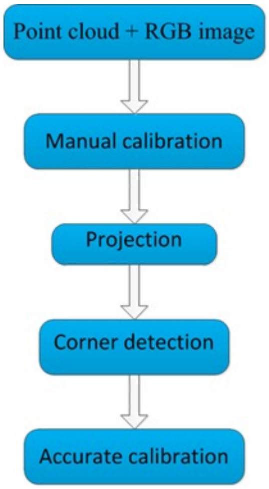

Figure 2. The calibration flow chart we propose.

Our processing scheme is shown in Fig. 2. Firstly, manual calibration or registration is carried on. In detail, we manually select some tie points, with the help of software, from point cloud and undistorted camera image for initial or inaccurate calibration. Subsequently, we project 3D point cloud into 2D reflectivity image. After that, corners of checkerboard from RGB image and reflectivity image can be extracted. With corners from reflectivity image, we can calculate their corresponding 3D points. Lastly, we use pixel dots from RGB image and points from point cloud to calibrate the camera and the scanner accurately.

4.2 Improvement

Generally, checkerboard is feasible for most calibration occasions. However, for 3D laser scanner, the situation becomes different. Unlike passive camera, the laser scanner (VZ1000) projects laser beams to measure depth, and the size of laser beams is not ignorable. Calibration using a standard checkboard cannot align point cloud and image at pixel-level accuracy as the corners extracted from reflectivity map are not real corners duo to laser beam size.

We adopt another pattern-circle as a substitute. Although we cannot directly extract a real point by projection mentioned just now, we still could calculate center of circle detected via aggregation to make up the error as a precise point. The procedure is similar to checkerboard except that we utilize center of circle detection instead of corner detection. The calibration scene can be seen in Fig. 3.

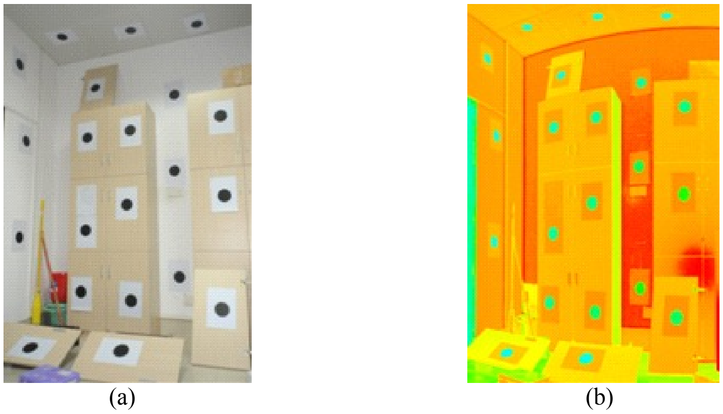

Figure 3. calibration using circle. (a) is undistorted camera image. (b) is reflectivity image projected by point cloud.

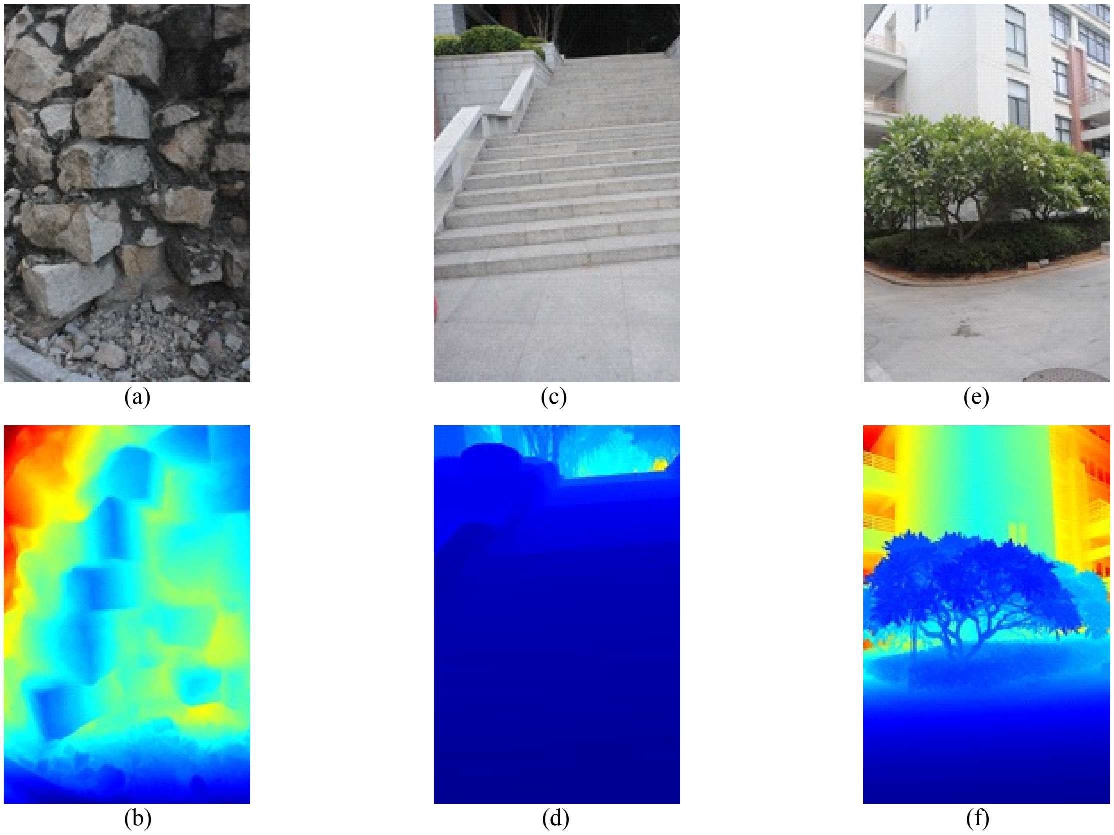

Figure 4. Our benchmark experiment results. (a) and (b) are rock and dense depth image. (c) and (d) are step and dense depth image. (e) and (f) are tree and dense depth image.

5. EXPERIMENT

We have captured a total of 40 pairs of outdoor laser scans and RGB images. 20 static scenes are captured as the same scene is captured twice from different locations so that half of data sets can be used for training. Each RGB image has five corresponding depth images consisting of 1 dense and 4 diverse sparse depth images captured by the scanner but with five different density levels.

The dense depth image, which can provide accurate depth measurement for almost every pixel except for occlusion and misalignment, is used as ground truth measurement for quantitative evaluation. Some data sets of our benchmark are shown in Fig. 4.

6. CONCLUSION

A new challenging depth super-resolution benchmark is proposed in this paper. Unlike established indoor data sets, we focus on uncontrolled and complicated outdoor real-world scenes. In the future, we plan to provide a depth super-resolution framework of our benchmark, evaluate present super-resolution algorithms and even propose an effective algorithm. Meanwhile, we also hope that it can attract more research on real-world super-resolution problem.

In addition, our benchmark and 3D measures can be used for other vision problems, including but not limited to super-pixel segmentation, 3D reconstruction, 3D object detection and 3D orientation estimation.

REFERENCES

[1] Scharstein, D. and Szeliski, R., “Middlebury stereo evaluation,” http://vision.middlebury.edu/stereo/eval/ (2002)

[2] Kopf, J., Cohen, M. F., Lischinski, D. and Uyttendaele, M., "Joint bilateral upsampling," ACM Transactions on Graphics, 96–102 (2007).

[3] He, K., Sun, J. and Tang, X., "Guided image filtering," PAMI 35(6), 1397–1409 (2013).

[4] Liu, M.Y., Tuzel, O. and Taguchi, Y., "Joint geodesic upsampling of depth images," Proc. CVPR, 169-176 (2013).

[5] Wang, L., Jin, H., Yang, R. and Gong, M., "Stereoscopic inpainting: Joint color and depth completion from stereo images," Proc. CVPR, 1-8 (2008).

[6] Yang, Q., Yang, R., Davis, J. and Nist´er, D., "Spatial-depth super resolution for range images," Proc. CVPR, 1-8 (2007).