Kinect-Like Depth Denoising

[†]Jingjing Fu
[†]Media Computing Group
Microsoft Research Asia
Beijing, China
{jifu, yanlu, spli}@microsoft.com

[‡]Shiqi Wang*, [†]Yan Lu, [†]Shipeng Li
[‡]The Institute of Digital Media
Peking University
Beijing, China
sqwang@jdl.ac.cn

[†‡]Wenjun Zeng
[†‡] Department of Computer Science
Univ. of Missouri
Columbia, MO, USA
zengw@missouri.edu

***Abstract*— Accuracy and stability of Kinect-like depth data is limited by its generating principle. In order to serve further applications with high quality depth, the preprocessing on depth data is essential. In this paper, we analyze the characteristics of the Kinect-like depth data by examing its generation principle and propose a spatial-temporal denoising algorithm taking into account its special properties. Both the intra-frame spatial correlation and the inter-frame temporal correlation are exploited to fill the depth hole and suppress the depth noise. Moreover, a divisive normalization approach is proposed to assist the noise filtering process. The 3D rendering results of the processed depth demonstrates that the lost depth is recovered in some hole regions and the noise is suppressed with depth features preserved.**

I. INTRODUCTION

In recent decades, various devices have been developed as an attempt to access the 3D information of the physical word, such as time-of-light (TOF) camera, stereo camera, laser scanner, and the structured light camera. However, the depth camera is not as popular as the RGB camera due to their high cost and enormous computing requirement. The launch of Kinect [1] provides a convenient way to access the depth information in real time, which has facilitated a series of applications, such as immersive gaming (Xbox), dynamic 3D reconstruction, object recognition, and so on.

The depth cameras aim to measure the distance from the camera to the target object by utilizing the light wave properties, but their working principles vary. For example, the TOF [2] measures the depth by detecting the light wave phase shift after reflection, while stereo camera [3] generates disparity map by stereo matching. Obviously, the depth generated by different devices exhibits different data characteristics. As a structured light camera, the Kinect depth is derived from the disparity between the projected infrared light pattern and the received one. The granularity and the stability of the received light speckles directly determine the resolution and the quality of the depth data.

The captured depth sequence is characterized by its large variation range and instability, which makes the Kinect depth pre-processing essential before performing future tasks such as compression and rendering. Similar to the depth derived from the stereo video, the Kinect depth suffers from the problems of depth holes and boundary mismatching due to the deficiency of the received light speckles. Moreover, even if the light speckles have been received by the sensor, the generated depth sequence is unstable in temporal domain due to the variation of the received light Patten. The depth data is likely to change from time to time, though when the Kinect is fixed to capture a static scene.

In order to enhance the quality of the stereo depth, tremendous works have been proposed to refine the depth by improving the accuracy of stereo matching [4,5]. However, few works discuss the stereo depth denoising without referencing the stereo RGB image, which is exactly the case happened to the Kinect depth. Although Kinect depth looks like a high dynamic gray image, the traditional image denoising algorithms [6, 7] cannot be directly employed due to its special disparity-like properties as will be discussed in Section II. On the other hand, the Kinect depth is a kind of noisy range data in terms of its physical meaning. A series of feature preserving denoising algorithms [8, 9] have been developed to process the range data captured by 3D scanner sensor. Also, as another representation of 3D data, the mesh generated from the range data can be denoised by bilateral filtering [10].

Considering all the special properties of Kinect depth, in this paper, we propose a Kinect denoising approach as an attempt to get a reliable and noise free depth map. In our approach, a block-based hole filling scheme is employed to predict the invalid depth values among the frames. Then the compensated depth sequence is filtered by a proposed divisive normalized bilateral filter in spatial and temporal domain simultaneously. With this approach, the properties of the depth video are greatly exploited to achieve a superior denoising performance.

The rest of the paper is organized as follows. In Section II, we give a brief analysis of the Kinect-like depth data. Section III describes the core techniques of our algorithm. The experiment results are given in Section IV. Finally, Section V concludes the paper.

II. KINECT-LIKE DEPTH CHARACTERISTICS

The Kinect consists of an infrared (IR) laser projector, an infrared sensor and an RGB sensor. The infrared projector emits the pseudo random pattern light by a diffractive mask, so that each speckle in the pattern can be recognized. Depth is measured by triangulation of each speckle between the observed light pattern and the reference light pattern which is obtained by capturing a plane at a known distance and stored in the memory of the sensor [11]. To be more specific, if a speckle is projected on an object whose distance to the sensor is different from that of the reference plane, the position of the speckle in the received image will be shifted in the direction of the baseline between the projector P and the perspective center of the infrared camera S. These shifts are measured for all speckles by a simple image correlation procedure, which yields a disparity image.

* This work was done while the author was with Microsoft Research Asia as a research intern.

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

Fig. 1 illustrates the relation between the distance D of an object point to the IR sensor and the distance D_r of a reference plane. To simplify the model, we assume that the origin of

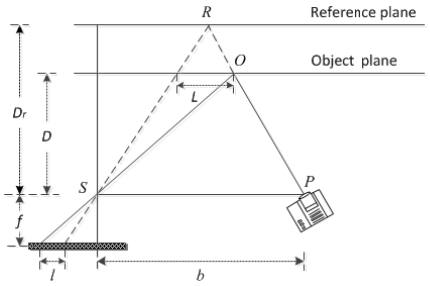

Fig. 1 Schematic representation of depth-disparity relation

depth coordinate system locates at the perspective center of the IR sensor. According to similarity of the triangles, we have

$$L/b = (D_r - D)/D_r \tag{1}$$

$$l/f = L/D \tag{2}$$

Where f is the focal length of the IR sensor; l is the relative shift length (disparity); b is the base length. After combining Eq. (1) and (2), the depth of the object is calculated as follows

$$D = \frac{D_r}{1 + D_r \cdot l/(f \cdot b)} \tag{3}$$

A. Spatial characteristcs

In [11], the author analyzes the formation of the depth errors in details. As the Kinect depth is derived from the infrared light disparity map, the depth noise exists due to the disparity accuracy limit. Moreover, the Kinect depth also suffers from the problems of depth holes and boundary mismatching caused by the light condition influence and deficiency of the imaging pattern. An example is shown in Fig.2. The RGB images and the depth images are captured simultaneously from a fixed Kinect camera. The irregular depth holes exist along the intersection between the roof and the wall, and the step-shaped depth fluctuation is noticeable after the texture is removed from the mesh (see Fig.2 (d)).

In order to analyze the depth noise, we formulate the noise as follows,

$$\widetilde{D} = M \cdot (D + e) \tag{4}$$

$\widetilde{D}$ is the output depth of Kinect, and M denotes the depth mask, indicating whether the depth value is valid at that position. The mask value of the invalid depth region is assigned to zero; otherwise the mask value is set to one. e is the depth noise introduced by the inaccurate disparity measurement.

The raw disparity length l is normalized during the depth measurement, l can be substituted by $ml^* + n$, with l^* the normalized disparity and m, n the parameters of normalized disparity. In the valid depth region ($M = 1$), the difference between the true depth and the output depth is

$$e = \widetilde{D} - D = \frac{n}{f \cdot b} D\widetilde{D} \approx C_0 n \widetilde{D}^2 \tag{5}$$

where $n \in \left[-\frac{l^*}{2}, \frac{l^*}{2}\right]$. Considering $\widetilde{D} - D \ll \widetilde{D}$, the true depth D can be replaced by $\widetilde{D}$ in Eq.(5) for approximation. Since the focal length and base length are both constant for Kinect, C_0 is used to represent the constant factor $1/fb$. Therefore, the level

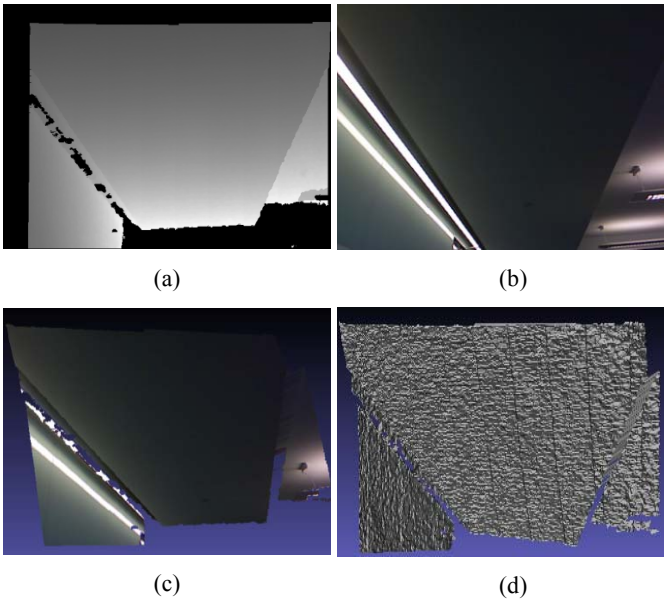

Fig. 2 Example of Kinect depth image and its corresponding RGB image (a) depth map after calibration (b) RGB image (c) 3D mesh rendered with texture (d) 3D mesh rendered without texture.

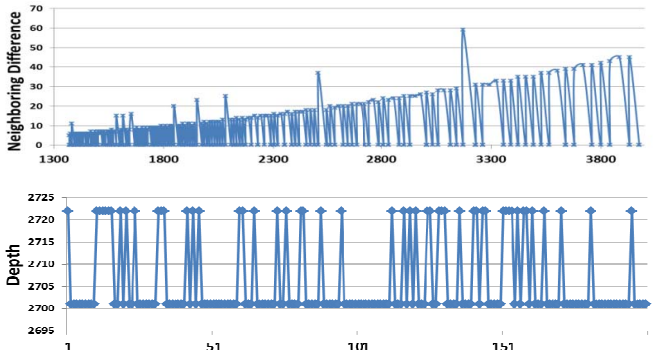

Fig. 3 Illustration of spatial and temporal depth characteristics. Top: Plot of the neighboring depth difference; the horizontal axis represents depth value. Bottom: depth variation with the time; the horizontal axis represents frame number.

of the noise e is proportional to the square of the corresponding depth value, which implies that the uniform filtering kernel is not applicable to the depth. As shown in Fig. 3(a), the neighboring depth value difference in the smooth region (roof in Fig.2) increases with the depth values, which verifies our analysis.

B. Temporal characteristics

The Kinect depth is captured at 30fps, and there exist strong temporal correlations between the true depth frames, which are of great importance for data compression and processing. However, due to the influence of the time-variant light condition, Kinect sensor may fail to capture the reflected light or fail to recognize and measure the light pattern variation during the depth generation. The depth data is likely to change from time to time, even if the Kinect is fixed to capture a static scene. Fig.3 (b) shows the depth value variation of a fixed position. For a static scene and fixed sensor, in the valid region we have,

$$\widetilde{D}^{t+\Delta t} - \widetilde{D}^{t} \approx C_0\left(\widetilde{D}^{t}\right)^2 (n^{t+\Delta t} - n^{t}) \quad (6)$$

Δt is the time intervals between frames. Assume $\{n^t\}$ an i.i.d Gaussian process with zero mean. Then the variation of the temporal difference is also proportional to the square of the corresponding depth value. For the depth hole locating in the smooth surface, it is possible to compensate it with surrounding depth information to rebuild the temporal correlation of depth.

III. SPATIAL-TEMPORAL DEPHT DENOISING

As shown in Fig.2, there exist holes in the Kinect captured depth map, which not only results in bad experience in final rendering but also disrupt the depth continuity that is crucial for denoising. Therefore, we introduce inter-frame hole values prediction to fill the hole region before depth filtering. The whole framework of the scheme is illustrated in Fig. 4. Firstly, to assist inter-frame depth prediction, the hole value is roughly padded by the spatially neighboring depth. After inter-frame hole filling, spatial-temporal divisive normalized bilateral filtering (DNBL) is applied for efficient denoising. In this approach, the filter coefficients are divisive normalized to adapt the depth map variation properties. Moreover, the temporal correlations among the depth sequences are also employed to reconstruct a noise-free depth sequence.

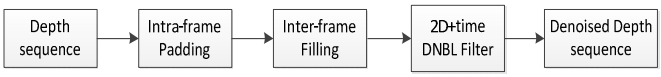

Fig. 4 Architecture of the proposed scheme.

A. Hole value predication

Similar to the image sequence, the depth value of the adjacent frames can be employed to predict the hole values of the current frame. Therefore, we adopt a non-local template matching algorithm for hole value prediction. Assume the current frame to be I_t and the adjacent frames to be I_{t+i} with $i \in [-m, m]$. For the invalid depth $I_t(x,y)$ in the current block at position (x,y), we search the corresponding blocks in the current and adjacent frames. The matching criterion is to find the minimum mean square error (MSE) between the available pixels and the corresponding pixels in the searched block. Note that the holes in the searched block are padded by the average of the neighboring pixels in order to avoid invalid reference caused by unexpected holes. The predicted value is calculated as follows

$$I_t(x,y) = \frac{1}{2m+1}\sum_{i=-m}^{m} \omega_i \cdot I_{t+i}(x + \Delta x_i, y + \Delta y_i) \quad (7)$$

where $(\Delta x_i, \Delta y_i)$ represents the motion vector for frame R_i. ω_i is defined as the weight for each reference pixel and it is defined as follows

$$\omega_i = G_\sigma(\mathrm{MSE(i)}) \quad (8)$$

where $G_\sigma(\mathrm{x})$ denotes the 2D Gaussian kernel:

$$G_\sigma(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (9)$$

Note the hole caused by out-of-range depth will not be processed, since little reference information can be extracted for depth prediction.

B. Spatial-temporal DNBL

The Bilateral filter has been successfully applied in many areas [12] such as denoising, data fusion and depth reconstruction. Assume the position of the to be filtered pixel is p and the pixel value is $I(\mathrm{p})$. The local neighborhood set of p which may have influence on $I(\mathrm{p})$ is denoted as $S(\mathrm{p})$. Then the bilateral filter BF can be formulated as follows:

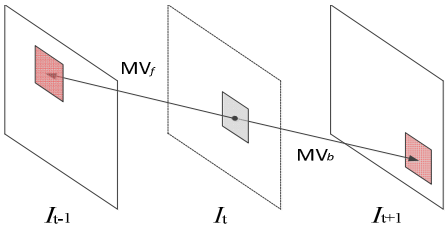

Fig. 5 Motion trajectories of the depth map.

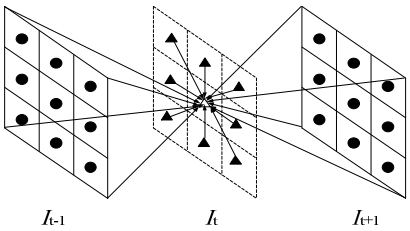

Fig. 6 Spatial-Temporal Bilateral Filtering Model.

$$\mathrm{BF}[I]_p = \frac{1}{Wp}\sum_{q\in S(\mathrm{p})} G_{\sigma_\mathrm{s}}(||p-q||)G_{\sigma_\mathrm{r}}(||I(p)-\mathrm{I}(q)||)I(q) \quad (10)$$

Where W_p is the normalization factor.

$$W_p = \sum_{\mathrm{q}\in \mathrm{S(p)}} G_{\sigma_\mathrm{s}}(||p-q||)G_{\sigma_\mathrm{r}}(||I(p)-I(q)||) \quad (11)$$

σ_s and σ_r are the filtering parameters, G_{σ_s} is a spatial Gaussian that decreases the influence of distant pixels, and G_{σ_r} is a range Gaussian that decreases the influence of pixels q with a large intensity difference. Since the weight of each pixel is determined by its position and value difference from the current pixel, the filtering could efficiently smooth the image while preserving the edge. However, as analyzed in Section II, the level of depth noise is proportional to the square of depth value. If the depth is directly processed by bilateral filtering with uniform Gaussian convolution kernel, the depth in a small value range will be over smoothed, and noises in a large depth region will be preserved as depth edges. In order to avoid this problem and preserve the edge information as much as possible, we apply an adaptive scaling technique to the range Gaussian kernel as follows:

$$G'_{\sigma_\mathrm{r}}(||I(p)-\mathrm{I}(q)||) = \frac{1}{2\pi{\sigma_\mathrm{r}}^2}\exp\left(-\frac{||I(p)-I(q)||^2}{2{\sigma_\mathrm{r}}^2\Theta(I(p))}\right) \quad (12)$$

Where $\Theta(I(p))$ is the depth variance of depth error e caused by inaccurate disparity measurement, which can be estimated by depth values statistics analysis. The difference between two pixels is normalized in terms of their error level to evaluate the importance of the depth discontinuity.

As we analyze in Section II, the temporal depth difference possesses similar property as the depth noise. Therefore, the high degree of correlation between adjacent frames can be used to achieve better estimate of the original signal since additional information is available from nearby frames. However, this process is complicated by the motion between video frames as the depth map is corrupted by the noise. Fortunately, the depth map is smoother than natural images. Moreover, in physics, an object can always be treated as moving in a straight line, which

motivated us to employ the block matching to obtain the motion trajectories.

For each block, we can identify a corresponding block in the previous and next frames, as shown in Fig. 5. Following the adaptive scaling bilateral filtering approach, for the to-be-filtered pixel $I(p)$, we can set the prediction windows in the previous frames and next frames. Denote the pixel set of the prediction windows as $S_{t-j}(p)$ and $S_{t+j}(p)$, with $j \in [-r, r]$. r is the temporal radius of the spatial-temporal bilateral filter, which can be different from the spatial radius. Then the pixel $I(p)$ will be filtered by $(2r+1)$ prediction windows along the time axis,

$$\hat{I}(p) = \frac{1}{W_p} \sum_{j=-r}^{r} \sum_{q_i \in S_{t+j}(p)} G_{\sigma_s}(||p - q_j||) G'_{\sigma_r}(||I(p) - I(q_j)||) I(q_j) \tag{14}$$

If r is set to one, the pixel $I(p)$ is filtered using the previous, current and next frame (see Fig. 6).

IV. Experimental Results

When the high dynamic range depth is normalized to grey image, most of details in the depth map are eliminated by the large step quantization, e.g. the smooth region in Fig. 2(c). Although the depth of the roof region seems to be smooth, noticeable ripples extend on the roof surface in the rendered mesh. Therefore, the performance of our proposed denoising algorithm is evaluated by rendering the depth map into a 3D mesh. As no ground truth references are available for the real scene, we subjectively compare the mesh results of processed depth with that of the original kinect depth.

Firstly, we compare the 3D rendered meshes without texture. Fig. 7 shows two groups of mesh which are generated by the third frame of the depth sequences ("roof" and "cubicle") given that the temporal radius is set to one in our experiment. As shown in Fig. 7 (a) and (c), the depth information is lost in some regions, such as the intersection between the roof and the wall, boxes under the desk and so on. Also, the depth values vary in a step-sharped manner within the smooth region. After hole filling and filtering by our algorithm, the processed depth is rendered as mesh, as shown in Fig. 7 (b) and (d). Our rendered mesh looks more continuous than the original one and the noise is well suppressed. More mesh results with texture are given in Fig. 8. It can be clearly observed that our approach can generate a mesh with better quality which results in more enjoyable watching experience. The lost depth values are properly predicted and the local depth features are preserved during the filtering process.

V. Conlusions

In this paper, we analyze the characteristics of the kinect like depth data and accordingly propose a spatial-temporal denoising algorithm, in which intra frame and cross frame information are exploited to compensate for the depth hole and suppress the depth noise. Moreover, a divisive normalized approach is employed for Bilateral filtering process. Significant mesh quality improvement is observed in the comparison between the 3D mesh rendered from the denoised depth map and that of the original depth, which verifies the efficacy of our algorithm.

References

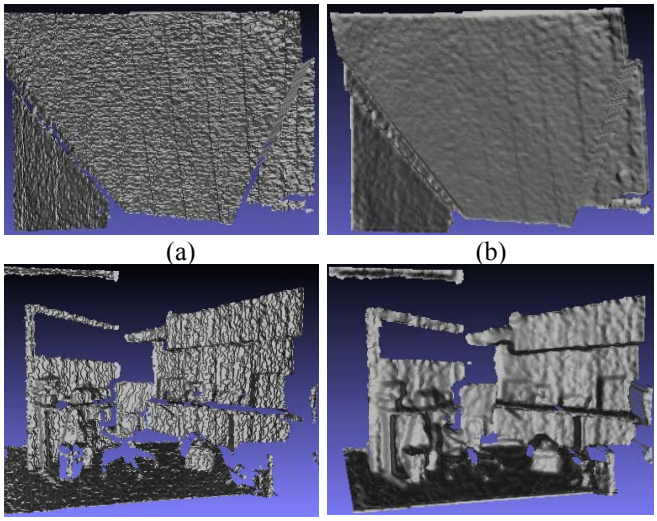

Fig. 7 Performance comparison of the rendered mesh (a) "roof" mesh rendered before denoising (b) "roof" mesh rendered after denosing. (c) "cubicle" mesh rendered before denosing (d) "cubicle" mesh rendered after denoising.

Fig. 8 Performance comparison of the mesh rendered from "cubicle" with texture. left: mesh rendered before denosing with texture. right: mesh rendered after denoising with texture

[1] Microsoft Kinect. http://www.xbox.com/de-de/kinect

[2] S.B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor-System Description, Issues and Solutions," in Proc. CVPR, 2004.

[3] T. Kanade, and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," in Proc. Pobotics and Automation, 1991, pp. 1088-1095.

[4] L. Alvarez, R. Deriche, J. Sanchez, and J. Weickert, "Dense Disparity Map Estimation Respecting Image Discontinuities: A PDE and Scale-Space Based Approach," Journal of Visual Communication and Image Representation, vol. 13, no. 1-2, pp. 3-21, Mar., 2002.

[5] R. Khoshabeh, S.H. Chan, and T.Q. Nguyen,"Spatio-temporal consistency in video disparity estimation," in Proc. ICASSP, 2011, pp. 885-888.

[6] L.I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D., vol. 60, no.1-4, pp. 259-268, 1992.

[7] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in Proc. CVPR, 2005, pp. 60-65.

[8] O. Schall, A. Belyaev, and H. P. Seidel, "Feature-preserving non-local denoising of static and time-variant range data" in Proc. ACM SPM 2007.

[9] D. Chan, "Noise vs. Feature: Probalitlistic Denoising of Time-of-Flight Range Data," [Online]

[10]S. Fleishman "Bilateral Mesh Denoising", in Proc. ACM SIGGRAPH 2003, Volume 22 Issue 3, July 2003

[11]K. Khoshelham, "Accuracy analysis of kinect depth data," ISPRS WORKSHOP 2011, Calgary, Alberta, Canada.

[12]S. Paris, P. Kornprobst, J. Tumblin, and F. Durand. "A Gentle Introduction to Bilateral Filtering and its Applications," In Proc. ACM SIGGRAPH Course, 2007.