Journal of Visual Communication and Image Representation 55 (2018) 540–547

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Dynamic 3D reconstruction improvement via intensity video guided 4D fusion☆

Jie Zhang[a,b,*], Christos Maniatis[c], Luis Horna[b], Robert B. Fisher[b]

[a] The School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China
[b] The School of Informatics, The University of Edinburgh, EH8 9AB, UK
[c] The School of Mathematics, The University of Edinburgh, EH8 9AB, UK

ARTICLE INFO

Keywords:
High-speed 3D video sensor
Multi-frame 4D fusion
Intensity tracking
Dynamic object
Noise reduction

ABSTRACT

The availability of high-speed 3D video sensors has greatly facilitated 3D shape acquisition of dynamic and deformable objects, but high frame rate 3D reconstruction is always degraded by spatial noise and temporal fluctuations. This paper presents a simple yet powerful dynamic 3D reconstruction improvement algorithm based on intensity video guided multi-frame 4D fusion. Temporal tracking of intensity image points (of moving and deforming objects) allows registration of the corresponding 3D model points, whose 3D noise and fluctuations are then reduced by spatio-temporal multi-frame 4D fusion. We conducted simulated noise tests and real experiments on four 3D objects using a 1000 fps 3D video sensor. The results demonstrate that the proposed algorithm is effective at reducing 3D noise and is robust against intensity noise. It outperforms existing algorithms with good scalability on both stationary and dynamic objects.

1. Introduction

Three dimensional shape acquisition of highly dynamic and deformable objects is an increasingly active research topic in computer vision, with the development of high-speed 3D video sensors [1,2]. It is a fundamental and critical prerequisite of numerous applications, such as dynamic face recognition [3], action and behavior perception [4,5], object deformation analysis, etc. However, the 3D sequences from high-speed 3D video sensors usually suffer from serious spatial noise and temporal fluctuations, which degrades the performance of 3D reconstruction. The inaccuracy of the high frame rate 3D sequence is caused by multiple factors, including calibration error, non-uniform illumination, surface properties, motion of scenes or objects, sensor variations, etc. In passive 3D reconstruction systems (e.g. stereo vision sensors), uneven illumination or texture reflectance can cause stereo matching errors and thus poor reconstruction performance, as shown in Fig. 1. Additionally, resulting from the sensor technology, there are a small number of out-of-sync pixels that produce spatial noise and temporal fluctuations in the 3D sequence, as shown in Fig. 2. Therefore, denoising high frame rate 3D/depth sequences and thus improving the performance of 3D dynamic and deformable shape acquisition is of significant value.

In this paper, we present a method to improve the dynamic 3D reconstruction from high-speed 3D stereo video sensors, where the 3D sequence improvement framework is based on 2D intensity tracking that guides a 4D spatio-temporal fusion. The core idea is that the 2D intensity data of consecutive images can be aligned by a temporal "stereo" matching algorithm, and then the corresponding 3D point data can be fused in the spatio-temporal domain to reduce the 3D spatial noise and temporal fluctuations.

The contributions of the paper are: (1) a simple yet powerful noise reduction pipeline for boosting the 3D reconstruction of dynamic and deformable objects. (Section 4); (2) a generic 2D intensity tracking guided multi-frame 4D fusion model that integrates spatial intra-frame filtering and temporal inter-frame fusion. (Section 3). In Section 5, we demonstrate the proposed method by denoising 3D sequences of stationary, dynamic and deformable objects from a 1000 fps 3D stereo video sensor.

2. Related works

For 3D/depth noise reduction, 3D/depth noise characterization and models [6–10] provide an important basis for boosting the performance of 3D reconstruction. Noise in a 3D/depth image can be generally characterized into three types including spatial, temporal and interference noise. Each type of noise corresponds to specific theoretical or

☆ This paper has been recommended for acceptance by Adrian Barbu.

* Corresponding author at: The School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China.
E-mail addresses: zhangjie09@buaa.edu.cn (J. Zhang), rbf@inf.ed.ac.uk (R.B. Fisher).

https://doi.org/10.1016/j.jvcir.2018.07.007
Received 18 November 2017; Received in revised form 18 May 2018; Accepted 7 July 2018
1047-3203/ © 2018 Elsevier Inc. All rights reserved.

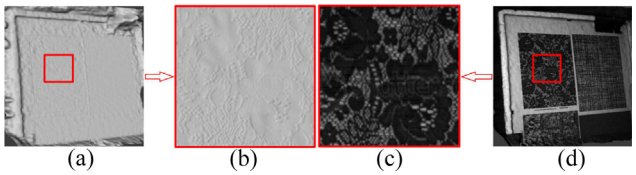

Fig. 1. Texture-related 3D noise on a static plane: (a) a 3D frame; (b) the region of interest of the 3D frame; (c) region of interest of the 3D frame with intensity texture; (d) the whole 3D frame with texture. The 3D noise in the 3D frame is closely related to the textures in the intensity image.

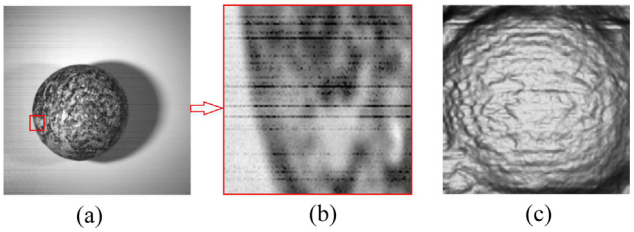

Fig. 2. Noise example: (a) an intensity frame of a falling sphere captured by a high-speed stereo video sensor; (b) invalid pixels in the intensity frames; (c) structural noise in a reconstructed 3D frame of the falling sphere.

empirical noise models. Most of 3D/depth image improvement methods mainly focus on reducing spatial axial and lateral noise, smoothing temporal fluctuations and filling non-measured pixels [11].

Existing algorithms are performed either using a single image (such as adaptive Gaussian filter (Ad-GF) [9], adaptive bilateral filter (Ad-BF) [12]) or multiple registered images (such as KinectFusion [13], imaging burst [14]). Recently, Guo et al. [15] also proposed to fuse multi-scale depth images using a hierarchical signed distance field for improved 3D reconstruction. The multi-view 3D registration based methods are helpful in smoothing 3D data and thus improving the 3D reconstruction quality, while the performance of the methods on dynamic or deformable objects is still limited.

To address this, there are existing algorithms using motion/temporal information for point-based fusion or filtering. For example, DynamicFusion [16] estimates dense non-rigid warp fields that fuse live frames of a dynamic scene to get a gradually denoised and complete 3D reconstruction. The dense SLAM system performs better on dynamic scenes compared with the KinectFusion algorithm. There are also some temporal filtering based algorithms, such as the velocity-based adaptive threshold filter (Ad-TF) [17], the spatial-temporal divisive normalized bilateral filter (DNBF) [18], and the constrained temporal averaging filter (TA) [19]). However, some are only based on the depth information of individual frames. On the other hand, depth-intensity based 3D/depth noise reduction methods including the adaptive joint bilateral filter (Ad-JBF) [20], the guided filter [21], the non-causal spatio-temporal median filter (ST-MF) [22], and the multi-sensor system [23] have been used for boosting the quality of 3D reconstruction. However, due to the limited reconstruction quality of high-speed 3D video sensors, denoising high frame rate sequences is still an open issue.

3. Proposed pipeline

The proposed system framework (Fig. 3) has 2 main stages: (1) 2D intensity tracking guided 3D motion field estimation; (2) spatio-temporal multi-frame 4D fusion. The input to the pipeline is a 3D sequence $S^t = \{\mathbf{p}_i^t \in \mathscr{R}^3\}$ with pixel-wise registered intensity $I^t = \{a_i^t \in \mathscr{R}\}$ and depth images $D^t = \{d_i^t \in \mathscr{R}\}$, where i is the pixel. In the first stage, dense tracking is performed on the intensity sequence I^t using a belief propagation based patch matching algorithm [24]. Thus, we obtain dense optical flow of I^t, which is also the continuous intensity motion field. Based on the projective camera model, the 3D motion fields of the pixel-wise registered 3D sequence P^t can be estimated by leveraging the intensity motion fields.

In the second stage, using the continuous 3D motion fields, piecewise spatio-temporal multi-frame 4D fusion is performed on the 3D sequence by fusing the registered 3D points. Rejected outliers in the 3D motion fields result in holes in the fused 3D sequence, so we perform gradient-directed hole filling to repair them. Finally, we can obtain a higher quality 3D sequence with smoother 3D spatial surface and less temporal fluctuations. More details on each stage are given in Section 4.

4. Intensity tracking guided 4D fusion

This section details the intensity tracking guided 3D motion field estimation and the spatio-temporal multi-frame 4D fusion model for 3D sequence improvement.

4.1. Intensity-guided 3D motion field estimation

For a dynamic 3D object, we assume that each intensity image point in n consecutive frames is trackable in the temporal domain. To achieve this, dense tracking is performed on the pixel-wise registered intensity sequence I^t using a particle belief propagation method [24]. This gives an intensity motion field $\{\mathbf{s}^{t,t+1} \in \mathscr{R}^2\}$ between each pair of consecutive 2D intensity frames I^t, I^{t+1}.

The intensity correspondence field $\mathbf{s}^{t,t+1} = \{\mathbf{s}_i^{t,t+1}\}$ is obtained by minimizing an objective function that combines a unary term evaluating point similarity and a pairwise term for piecewise smoothness as:

$$\hat{\mathbf{s}}^{t,t+1} = \underset{\mathbf{s}^{t,t+1}}{\operatorname{argmin}} \sum_i \left(\psi_1(\mathbf{s}_i^{t,t+1}) + \sum_{n \in N_I(i)} \psi_2(\mathbf{s}_i^{t,t+1}, \mathbf{s}_n^{t,t+1})\right) \tag{1}$$

In Eq. (1), $N_I(i)$ are the neighbors of the i_{th} 2D intensity pixel a_i^t in frame I^t;$\psi_1(\mathbf{s}_i^{t,t+1})$ is the unary term that represents the discrepancy of a pair of corresponding 2D intensity patches centered on the i_{th} pixel in consecutive frames I^t, I^{t+1}, as

$$\psi_1(\mathbf{s}_i^{t,t+1}) = \sum_{n \in N_I(i)} w_{1n} \left\| I^{t+1}(\mathbf{k}_i + \mathbf{k}_n + \mathbf{s}_i^{t,t+1}) - I^t(\mathbf{k}_i + \mathbf{k}_n) \right\| \tag{2}$$

where $\mathbf{k}_i$ is the 2D coordinates of the i_{th} pixel in frame I^t;$\{\mathbf{k}_n\}$ is the 2D coordinates of the intra-frame neighbors of the pixel $\mathbf{k}_i$;w_{1n} is a weight assigned to each neighbor $\mathbf{k}_n$, emphasizing closer points to the center.

$\psi_2(\mathbf{s}_i^{t,t+1}, \mathbf{s}_n^{t,t+1}) = w_{2n} \left\| \mathbf{s}_i^{t,t+1} - \mathbf{s}_n^{t,t+1} \right\|$ is a smoothness term to regularize the correspondence field, which can be optimized by minimizing the message (smoothness error) passed by the intra-frame neighboring intensity patch n to the patch i. w_{2n} is a weight assigned to each neighboring motion vector $\mathbf{s}_n^{t,t+1}$.

The resulting pixel-wise continuous intensity motion fields $\mathbf{s}^{t,t+1}$ give pixel-wise correspondences for the registered depth frames D^t. We iterate the correspondences across time t so each point has a linked position $\mathbf{p}_i^t$ in the depth frame D^t (3D frame S^t).

Using the projective camera model (assuming that the intensity pixels are distortion-free), the point $\mathbf{p}_i^t$ in the 3D frame S^t can be expressed as

$$\mathbf{p}_i^t = d_i^t \left[f_x^{-1}(x_i^t - u_0),\ \ f_y^{-1}(y_i^t - v_0),\ \ 1 \right] \tag{3}$$

where f_x, f_y, u_0, v_0 are the calibration parameters (focal length and centers) of the camera, $d_i^t$is the depth value, and x_i^t, y_i^t are intensity image pixel coordinates.

For an intensity field, the registration from frame I^t to frame I^T is

$$\mathbf{s}_i^{t,T} = [s_{ix}^{t,T}, s_{iy}^{t,T}] \tag{4}$$

where $s_{ix}^{t,T} = x_i^T - x_i^t$ and $s_{iy}^{t,T} = y_i^T - y_i^t$. The 3D correspondence vector $\mathbf{m}_i^{t,T}$ for the point i from the corresponding frame S^t to frame S^T can be estimated by:

Download English Version:

https://daneshyari.com/en/article/6938148

Download Persian Version:

https://daneshyari.com/article/6938148

Daneshyari.com